Genomic Studies on Floral and Vegetative Development in the Genus *Streptocarpus* (Gesneriaceae)

Yun-Yu Chen

Doctor of Philosophy
The University of Edinburgh
Royal Botanic Garden Edinburgh
2019

Declaration

I declare that this thesis has been composed solely by myself and that it has not been submitted, in whole or in part, in any previous application for a degree. Except where stated otherwised by reference or acknowledgement, the work presented is entirely my own.

Yun-Yu Chen

Yun-Yu Chen

Date

15/07/2019

Acknowledgements

I would like to express my utmost gratitude to the wonderful people who joined and contributed to this project and helped me throughout the PhD. First are my supervisors, Michael Möller, Catherine Kidner, and Kanae Nishii – thank you for giving me the opportunity, guidance, joy, intellectual support, and sometimes physical hard work! - throughout my study. Thank you all very much for your time and effort in discussing and commenting the works we produced together, and in helping me to practice and improve myself through numerous posters, presentations, and written pieces of work. This thesis is a result of all of our dedications. I would also like to thank our collaborators, Christine Hackett from the BioSS for the course and discussion on linkage mapping and QTL mapping, and Atsushi Nagano from Ryukoku University for the RAD-Seq experiments. I feel extremely lucky and grateful to have worked with you all, and the memory will forever stay dearly in my heart. My sincerest thanks also goes to my examiners, Michelle Hart and Minsung Kim, on reviewing this brick of papers and for stimulating the questions and discussions during the viva. You have helped me improve the work and myself even more.

My gratitude also goes to the supporting teams at the University of Edinburgh and the Royal Botanic Garden Edinburgh. To my thesis committee member, Prof Andrew Hudson, who joined all my PhD milestones, providing invaluable discussions about the project. To all members in the RBGE horticultural team, particularly Sadie Barber, Nate Kelso and Andy Ensoll for taking care of the plant materials – without you the project would not have happened. Thanks to the RBGE STS team, Laura Forrest and Ruth Holland for the support in molecular lab, and to Frieda Christie for the training and discussion in microscopy and SEM. Thanks to the ICT team members for fulfilling my endless software and server requests.

A big thank you to all the friends in Edinburgh - Julieth, Roosevelt, Javier, Karina, Flávia, Kristine, Pakkapol, Andrés, Mauricio, Lucía, Surabhi, Nadia, Thibauld and so many more - I will always remember the happiness and time we shared together in the BO8 office as well as in the pubs. To people in the tropical science group: Axel, Peter, Mark, Tiina, Toby and others, for the fun times and intellectual discussions. Special thanks to my former landlady Susan and her family for their hospitality. Also to Moth, their little granny cat for the cuddling service provided.

Finally, my sincerest gratitude goes to my family - for my parents' love and endless supports, for waiting for me at the other side of the world patiently. To my wife Andrea, who appeared at the most surprising moment, and accompanied me to complete this journey. Thank you for your patience and understanding, going through all the hard work and stressful moments with me. Wherever I am, I know I can look to you and feel the calmness and supports you reserved for me. I owe you everything.

Abbreviations

BAM	Binary sequence alignment map (file format)
BLAST	Basic local alignment search tool
BLAT	BLAST-like alignment tool
BM	Basal meristem
Bp	Base pairs
BTL	Binary trait loci
BUSCO	Benchmarking universal single-copy orthologs
cM	Centimorgan
CTAB	Cetyl trimethylammonium bromide
DBG	De Bruijn graph
DNA	Deoxyribonucleic acid
EDTA	Ethylenediaminetetraacetic acid
Gbp	Giga base pairs
gDNA	Genomic DNA
GM	Groove meristem
GWAS	Genome wide association study
KEGG	Kyoto Encyclopedia of Genes and Genomes
Mbp	Mega base pairs
NGS	Next generation sequencing
ORF	Open reading frames
PCR	Polymerase chain reaction
QTL	Quantitative trait loci
RAD-Seq	Restriction-site associated DNA sequencing
RBGE	Royal Botanic Garden Edinburgh
RNA	Ribonucleic acid
RNA-Seq	RNA sequencing
Rpm	Revolution per minute
SAM	Sequence alignment map (file format)
SEM	Scanning electron microscope

Abstract

The genus *Streptocarpus* consists of around 180 species with diverse morphologies. At least three main types of vegetative growth forms can be distinguished: caulescent, rosulate (acaulescents with multiple leaves), and unifoliate (acaulescents with one leaf). Floral size, shape, and pigmentation pattern are also highly variable between species. Previous studies have suggested that some of the morphological characters are inherited as Mendelian traits. For instance, the rosulate growth form is dominant over the unifoliate, and the rosulate / unifoliate growth form was hypothesised to be determined by two genetic loci, based on the Mendelian segregation ratios recorded in backcross and F2 populations. However, the identity of the loci and the underlying molecular mechanisms remain unknown. In this study, *Streptocarpus rexii* (rosulate) and *Streptocarpus grandis* (unifoliate) were used to study the genetic basis of morphological variation in *Streptocarpus*. The aim is to use modern next generation sequencing (NGS) technologies to build draft genomes, transcriptomes, and genetic maps for the non-model *Streptocarpus* plants, and carry out quantitative trait loci (QTL) mapping to locate the causative loci.

First, suitable DNA and RNA extraction methods for obtaining NGS-quality nucleic acids from *Streptocarpus* were established. For DNA extraction this was a modified protocol of the ChargeSwitch gDNA Plant Kit, and for RNA extraction a TRIzol reagent plus phenol:chloroform:isoamyl alcohol wash protocol was devised. The nucleic acid samples extracted were subsequently used for library preparation and NGS sequencing experiments.

Whole genome shotgun sequencing was performed for *S. rexii* and *S. grandis* using Illumina HiSeq 4000 and HiSeq X. *De novo* assembly of the sequence data produced a *S. rexii* draft genome of 596,583,869 bp, with 95,845 scaffolds and an N50 value of 35,609 bp. The *S. grandis* draft genome had a total span of 843,329,708 bp, with 127,951 scaffolds and an N50 value of 31,638 bp. The genome assemblies served as references for subsequent NGS data analysis.

The RNA samples derived from various vegetative and floral tissues of *S. rexii* and *S. grandis* were sequenced on MiSeq and HiSeq 4000 platforms. The transcriptome assembly was carried out using *de novo* and reference-based methods (i.e. mapped to the obtained draft genomes), followed by putative protein-coding open reading frame identification and annotation. For *S. rexii*, 60,500 and 53,322 transcripts were constructed in the *de novo* and reference-based assemblies respectively. For *S. grandis*, 51,267 and 46,429 transcripts were constructed respectively.

A *Streptocarpus* genetic map was constructed using restriction-site associated DNA sequencing (RAD-Seq) genotyping of a backcross population ((*S. grandis* × *S. rexii*) × *S. grandis*). The RAD-Seq data were analysed using a *de novo* approach and reference-based approaches with two different aligners, and the RAD-markers recovered from the three approaches were combined to maximise the genetic map density. Different marker-filtering strategies with varying stringencies were also tested and compared. The results showed that the most stringently filtered map had 377 mapped markers in 17 linkage groups, and a total distance of 1,144.2 cM. On the other hand, the densest map consisted of 853 markers in 16 linkage groups (matching the basic haploid chromosome number of the *Streptocarpus* species used here), and a total distance of 1,389.9 cM.

The maps constructed were used for QTL mapping of growth form variation, identifying up to 5 effective loci for the rosulate / unifoliate phenotypes, with two of the loci on LG2 and LG14 consistently found in all mapping attempts. The results suggest that the variation in growth form may be regulated by two major loci, but a few additional minor loci might also be associated with the trait. Several QTLs for floral dimension, flowering time, and floral pigmentation patterns were also found, and the genetic regions associated with the floral traits of *Streptocarpus* were revealed for the first time.

During this study valuable genomic resources were generated for future research to identify the genes underlying different morphologies in the genus *Streptocarpus*. The reported QTLs narrow down the genetic region for fine-mapping studies, and the genome and transcriptome resources will aid the isolation of candidate gene sequences. Identifying the genetic loci and their crosstalk behind the variable morphologies in future work will greatly add to our knowledge on how the highly diverse genus *Streptocarpus* has evolved and on how fundamental developmental processes of plants are regulated.

Lay summary

An important question in biology is how differences in shape and form between species have evolved. To answer this question, a key step is to investigate how shape and form develop and understand the genetic mechanisms regulating these processes. The identification of the genes responsible for the developmental differences underlying the morphological differences is a milestone in understanding the evolution of diversity.

Streptocarpus is a group of plants including some which have become popular houseplants (e.g. African Violet and Cape Primroses). Some species in the genus have unconventional developments that have attracted botanical research for over 70 years. Most flowering plants form shoots with clearly defined growing tips which produce ‘conventional’ leaves and flowers but some *Streptocarpus* species produce leaves in unconventional ways – from meristems that develop at the base of leaves, or plants that produce a single leaf that grows for the whole life span of the plant. However, the genetic changes which cause this dramatic shift in form remain unknown. Here, we used modern next generation sequencing (NGS) technologies to build genetic resources for *Streptocarpus* that are required for gene identification, and to isolate the causative genes inferring morphological variation in the genome.

This study describes the generation of the first draft genome sequence for *Streptocarpus*, identifies the expressed part of the genome through RNA sequencing, and creates a genetic map (i.e. the graphical representation of a chromosome with linear arrangements of genetic markers). By studying two *Streptocarpus* species with distinctive morphologies, *S. rexii* with multiple leaves in an irregular rosette and *S. grandis* with only one leaf, we identified genomic regions where the causative genes were most likely to be located.

This is the first study reporting association mapping, identifying association between morphologies and genome sequences, of vegetative and floral traits in *Streptocarpus* to aid the isolation of candidate gene sequences. The presented work provides the basis for further studies to understand the molecular mechanism of *Streptocarpus* development, which will greatly add to our knowledge on how this morphologically highly diverse genus has evolved and on how fundamental developmental processes are regulated in plants.

Table of contents

Chapter 1 Introduction

1.1 Background

Studies in model organisms provide important insights into fundamental biological processes. On the other hand, non-model organisms may have unusual yet interesting properties that are not observed in model systems (Russell et al., 2017). For instance, non-model organisms may show unique morphologies, and can serve as valuable materials to study the evolution of morphological diversity (Mauricio, 2001; Bolger et al., 2017a).

In my thesis, I focus on genomics and genetics analyses in the non-model genus *Streptocarpus*. *Streptocarpus* belongs to the family Gesneriaceae and shows a wide range of morphological variations (Hilliard and Burtt, 1971; Möller and Cronk 2001; Nishii et al., 2015). Some of the species display an unordinary growth of their above-ground shoots, such as one-leaf plants, which retain an enlarged cotyledon as a sole above ground vegetative organ (Jong, 1970; Jong and Burtt, 1975; Möller and Cronk, 2001; Nishii et al., 2015). The aim of this study is to increase our understanding of the genetic mechanisms that regulate the morphological variations in these *Streptocarpus* species.

1.2 The genus *Streptocarpus*

1.2.1 Overview

The genus *Streptocarpus* was first reported by John Lindley in the 1828 edition of *The Botanical Register* (Figure 1.1). He described it as rivalling the famous ornamental *Gloxinia* species in looks, while *"surpasses it in the elegance of its figure, and the delicacy of its colouring"* (Lindley, 1828). The name *Streptocarpus* was derived from the Greek στρεπτός (twisted) καρπός (fruit) characterising their twisted capsule, which was a traditional taxonomical character for this genus although it was later found to be lost in some species in the light of its new delineation (Möller and Cronk, 1997; Nishii et al., 2015). *Streptocarpus* includes some popular ornamental plants with important horticultural value, such as the African Violets (section *Saintpaulia*) and the Cape Primroses in section *Streptocarpus* (Nishii et al., 2015). Both are known commercially important flowering plants, and are of particular interest to plant breeders world-wide (Buta et al., 2010; Currey and Flax, 2015). Many cultivars based on *Streptocarpus* hybrids are commonly cultivated throughout Europe, America, and Asia for ornamental purposes (Reinten et al., 2011; Maria et al., 2004).

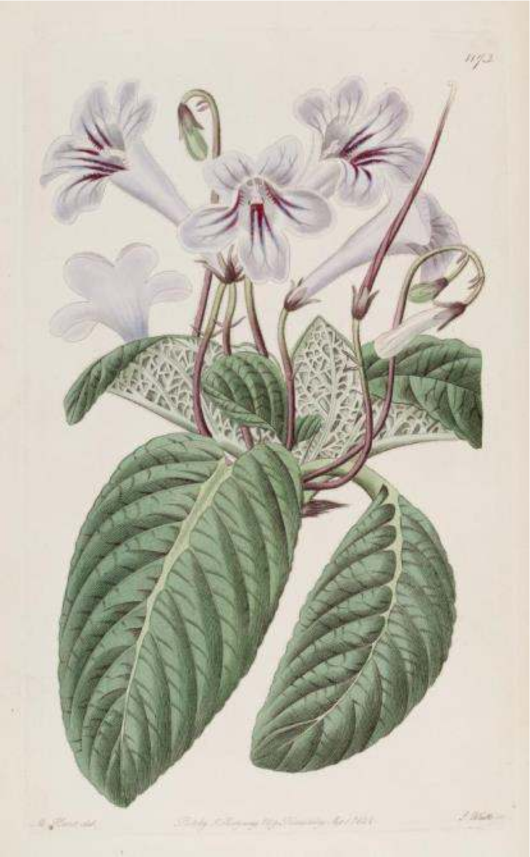


Figure 1.1 Illustration of *Streptocarpus rexii*, the type species of the genus *Streptocarpus*. (Figure from Plate 1173, Lindley 1828, Botanical Register)

In Gesneriaceae, the genus belongs to subfamily Didymocarpoideae, tribe Trichosporeae, subtribe Streptocarpinae (Weber, et al., 2013). The latest phylogenetic study using molecular markers, the Internal Transcribed Spacer region (ITS) and three chloroplast sequences, revealed the relationships among more than 130 (out of >176 described) species in the genus (Nishii et al., 2015). The genus is divided into two subgenera, *Streptocarpella* and *Streptocarpus*, consisting of seven and five sections, respectively. Geographically, the genus is distributed across Africa, Madagascar and the Comoro Islands, with the subgenus *Streptocarpus* found throughout eastern and southern Africa, and the subgenus *Streptocarpella* distributed widely in central and eastern Africa; both subgenera are found in the Madagascar and the Comoro Islands (Hilliard and Burtt, 1971). *Streptocarpus* plants are either monocarpic (i.e. die after flowering and fruiting) or perennial plants, with herbaceous or shrubby-woody habits (Hilliard and Burtt, 1971; Humbert 1971; Jong et al. 2012). In their natural habitat they are generally found growing on rocks, ravines, forest floors, less commonly under rock boulders outside forests, and rarely epiphytically in shady gorges (Lawrence, 1940; Hilliard and Burtt, 1971).

Streptocarpus species show several notable properties, and have been studied in many different disciplines and subject areas. These include studies on phylogeography (e.g. Hughes et al., 2005), population genetics (e.g. Hughes et al., 2004; 2007), morphology and evolution (e.g. Jong, 1970; Jong and Burtt, 1975; Möller and Cronk, 2001; Nishii et al., 2017), meristem development (e.g. Jong 1970; Jong and Burtt 1975; Imaichi et al., 2000; Rauh and Basile, 2003; Nishii et al., 2004; Nishii and Nagata, 2007; Mantegazza et al., 2007; Mantegazza et al., 2009; Nishii et al., 2010a; Tononi et al., 2010; Nishii et al., 2012a), floral development and evolution (e.g. Harrison et al., 1999; Hughes et al., 2006), physiology such as photoperiodism (Nitsch, 1967), hormone responses (e.g. Rosenblum and Basile, 1984; Nishii et al., 2012a; Nishii et al., 2014; Chen et al., 2017), biochemistry (Scott-Moncrieff, 1936; Stöckigt et al., 1973; Inoue et al., 1984; Sheridan et al., 2011; Inoue et al., 1982), cytology (e.g. Lawrence et al., 1939; Ratter, 1963; Jong and Möller, 2000; Möller and Pullan, 2015; Möller, 2018), transcriptomics (Chiara et al., 2013), and chloroplast genomics (Kyalo et al., 2018). In particular, the morphological variations observed in *Streptocarpus* are extraordinary for land plants, and suggests a greater flexibility of developmental programs than those revealed in model plant systems. In my PhD project, I focus on studying the morphological variations and the genetic basis of vegetative and floral characters.

1.2.2 Variation and inheritance of vegetative forms

One of the most unusual features observed in the development of *Streptocarpus* species is their diverse vegetative growth forms (Figure 1.2). The classification of their growth forms attracted early taxonomic attention (Fritsch 1893-1894), and later detailed morphological studies were carried out (Jong, 1970; Jong and Burtt, 1975). In general, *Streptocarpus* species can be roughly grouped into caulescents (with typical shoot apical meristems, SAM; Figure 1.2 a, d) and acaulescents (lacking a conventional SAM; Figure 1.2 b, c, e, f). Acaulescent species can further be distinguished into two subgroups: rosulates (with multiple leaves; Figure 1.2 b, e) and unifoliates (with a single leaf; Figure 1.2 c, f) (Jong, 1970; Hilliard and Burtt, 1971). The latest study identified over 30 caulescents, 50 rosulate and 40 unifoliate species and some intermediate forms in a total set of 167 species (Nishii et al., 2015). The evolutionary history of the growth forms is not fully resolved, and hybridisation, ecological niche adaptation, frequent transition between the growth forms may all be involved in shaping the species' vegetative habit (Möller and Cronk, 2001; Nishii et al., 2017).

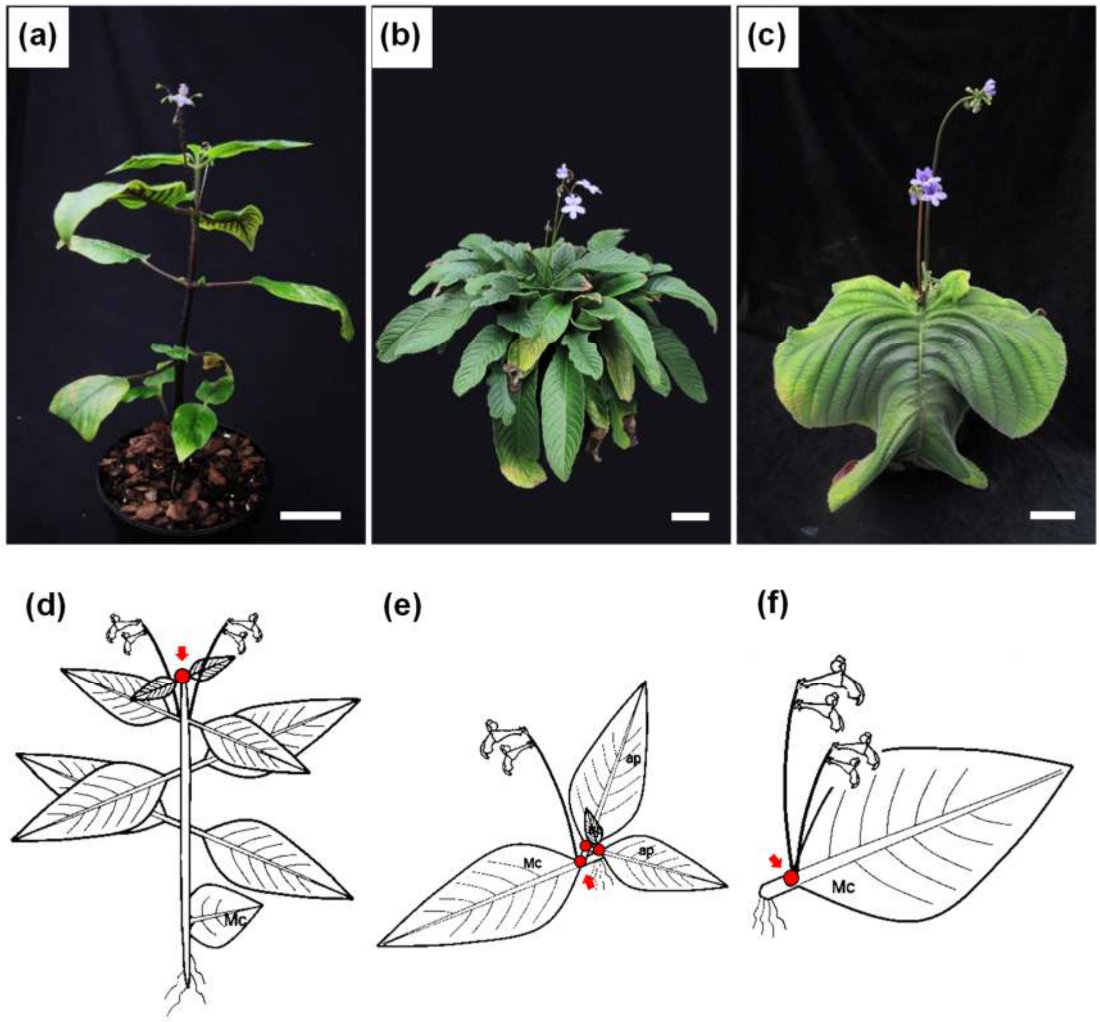


Figure 1.2 Examples of variation of growth forms observed in *Streptocarpus*. **(a)** *Streptocarpus thysanotus*, caulescent. **(b)** *Streptocarpus johannis*, excentric rosulate. **(c)** *Streptocarpus wendlandii*, unifoliate. **(d)** - **(f)** Schematic illustrations of the growth forms. **(d)** Caulescent. **(e)** Rosulate. **(f)** Unifoliate. Red circles and arrows indicate location of leaf-forming meristems, *Mc* macrocotyledon, *ap* additional phyllomorphs. Bars = 5 cm. (Illustrations modified from Nishii et al. 2016)

The classification of the genus is closely linked to growth form and has been revised several times over the years. The different morphs of *Streptocarpus* were first recognised by De Candolle (1845), who noticed the short stem of some of the species and grouped the known species into the *caule abbreviato* (abbreviated stem) and the *caulescentes* (caulescents). Fritch 1893-1894 further divided the growth forms into three taxonomical groups: the *caulescentes* (caulescents), *unifoliati* (unifoliates), and *rosulati* (rosulates). It was later recognised that there is no hard line between unifoliate and rosulate growth forms, with some rosulate species showing unifoliate morphology at early life stages, or unifoliate species that can produce additional leaves (Burtt, 1939). Nevertheless, this classification system was further extended to include the *sub-unifoliate* growth form, describing some small-leaved unifoliates that produce additional leaves occasionally (Lawrence, 1958). Later, extensive morphological studies were carried out on *Streptocarpus fanniniae* and other species, and the species categorised into four major groups: unifoliates, plurifoliates, rosulates, and caulescents (Jong, 1970; Jong and Burtt, 1975). Humbert (1971) described additional growth forms in Madagascar including the species with leaves in basal rosette with long petioles, species with leaves in basal rosette with veins ascending from the base and shrubby species with short filaments and non-coherent anthers. Hilliard and Burtt (1971) placed all the caulescents and the petioled Madagascan species into subgenus *Streptocarpella* and the remaining acaulescent species in subgenus *Streptocarpus* without further formal subdivision. In the latest study, a total of six fundamental growth patterns were defined, including the caulescent, rosulate, unifoliate, creeping rhizomatous stem, shrubby, and *Saintpaulia*-like rosette (Nishii et al., 2015). On the basis of phylogenetic results, floral and growth patterns, a classification with two subgenera and 12 sections was proposed, in which unifoliates and rosulates occurred in mixed sections, the African section *Streptocarpus* and the Madagascan sections *Colpogyne* and *Plantaginei* (Figure 1.3 Nishii et al., 2015). The subgenus division of Nishii et al. (2015) is fully supported by cytology with subgenus *Streptocarpella* possessing a basic chromosome number of $x = 15$, while those of subgenus *Streptocarpus* have $x = 16$.

(Next page) **Figure 1.3** Latest molecular systematics of Streptocarpus with the growth habits mapped on each taxon **(a)** subgenus *Streptocarpella*, **(b)** subgenus *Streptocarpus*. Growth habit: ● caulescent, ★ creeping rhizomatous stem, ■ rosulate, ▬ Saintpaulia-like rosette, ▲ unifoliate, ✦ shrubby. (Figure modified from Nishii et al., 2015)

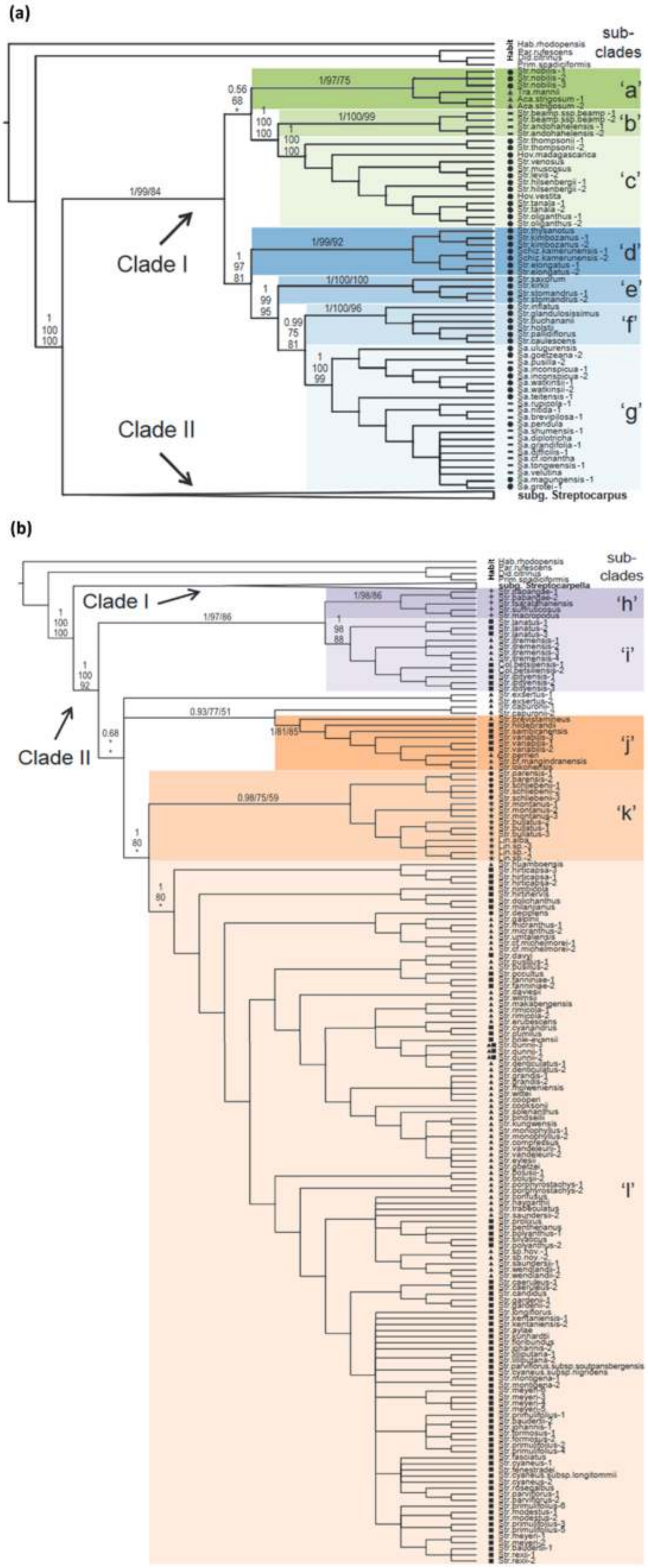


Figure 1.3 Latest molecular systematics of *Streptocarpus* with the growth habits mapped on each taxon. Full legend given on previous page.

My PhD project focuses on the developmental differences between the two acaulescent growth forms, rosulate and unifoliate. The above-ground vegetative body of acaulescent *Streptocarpus* is composed of specialised organs named phyllomorphs (Figure 1.4; Jong, 1970). A phyllomorph is a leaf/stem construct bearing several meristems. Each phyllomorph consists of a lamina and a petiolode, the latter with functions of petiole and stem. Three meristems are found on a phyllomorph: the basal meristem maintains lamina growth and is located at the proximal end of the lamina (Figure 1.4; bm). The petiolode meristem controls the petiolode and midrib extension and thickening (Figure 1.4; pm). The groove meristem located at the juxtaposition between the lamina and petiolode, gives rise to additional phyllomorphs and/or inflorescences (Figure 1.4; gm).

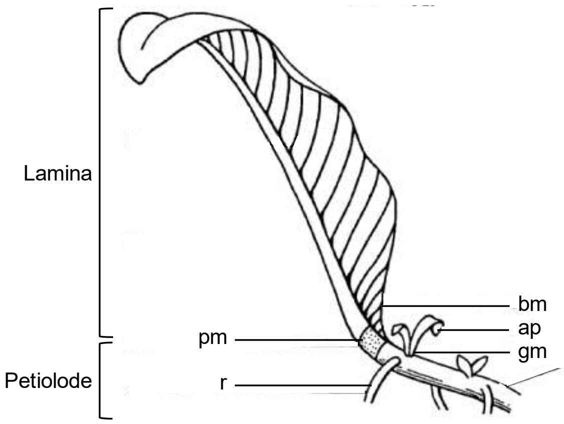


Figure 1.4 Schematic illustration of a phyllomorph, with morphology based on *Streptocarpus fanniniae* C. B. Clarke. *bm* basal meristem, *ap* additional phyllomorph, *gm* groove meristem, *pm* petiolode meristem, *r* root. (modified from Jong and Burtt, 1975)

A major distinction between the rosulate and unifoliate growth forms is the differentiation of the groove meristem (Jong, 1970). At seed germination and early seedling development stages, the morphology between the two growth forms are very similar (Jong, 1970), which both rosulate and unifoliate species showing anisocotylous development (Figure 1.5, Figure 1.6). Anisocotyly is the unequal growth in a pair of cotyledons, where one of the two cotyledons grows large and becomes a major photosynthetic organ (cotyledonary phyllomorph; Caspary, 1858; Fritsch, 1904; Jong, 1970; reviewed in Nishii et al., 2010b). But as the plant grows, the difference between rosulate and unifoliate becomes apparent; in rosulate species the groove meristem will develop additional phyllomorphs, and the successive production of further phyllomorphs from the groove meristem of the preceding phyllomorph arranges them in either a more-or-less regular (centric rosulate) or irregular (excentric rosulate) rosette (Figure 1.5; Jong, 1970; Nishii and Nagata, 2007). Later on, each phyllomorph will produce inflorescences at the base of the lamina. On the other hand, in unifoliate species the enlarged cotyledonary phyllomorph is the only above-ground vegetative organ and the groove meristem will differentiate into inflorescences (Figure 1.5; Jong, 1970; Jong and Burtt, 1975; Imaichi et al., 2000).

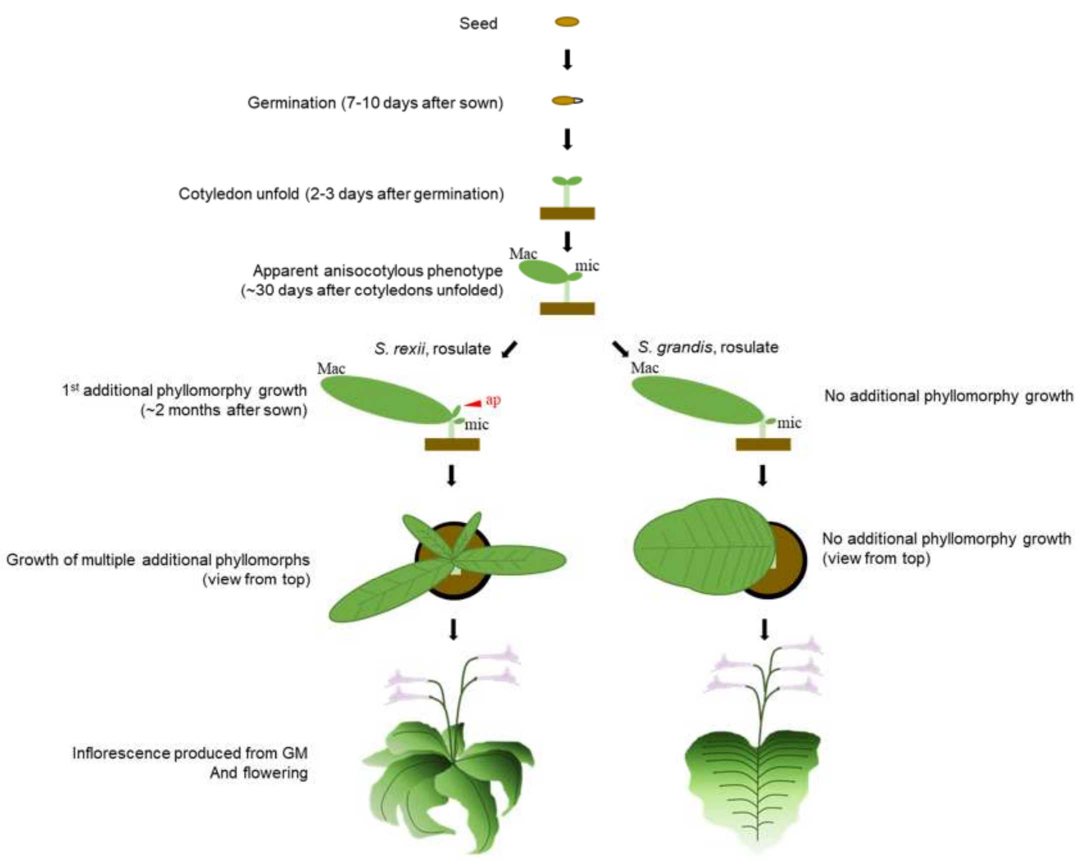

Seed
Germination (7-10 days after sown)
Cotyledon unfold (2-3 days after germination)
Mac
mic
Apparent anisocotylous phenotype
(~30 days after cotyledons unfolded)
S. rexii, rosulate
S. grandis, rosulate
Mac
1st additional phyllomorphy growth
(~2 months after sown)
ap
mic
Mac
No additional phyllomorphy growth
mic
Growth of multiple additional phyllomorphs
(view from top)
No additional phyllomorphy growth
(view from top)
Inflorescence produced from GM
And flowering

Figure 1.5 Schematic illustration of rosulate and unifoliate development, using *S. rexii* and *S. grandis* as example. The actual development time may vary depending on the growth condition.

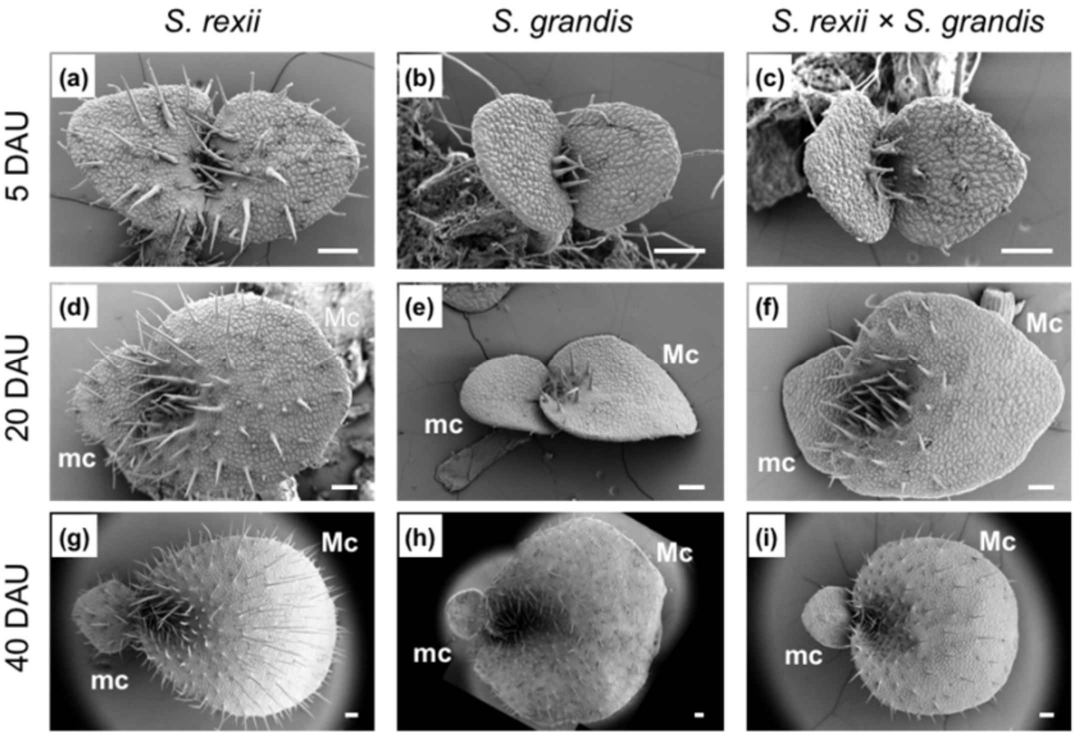


Figure 1.6 Seedling morphologies of the parental lineages. **(a)** 5 DAU isocotylous seedling of *S. rexii*. **(b)** 5 DAU isocotylous seedling of *S. grandis*. **(c)** 5 DAU isocotylous seedling of *S. grandis* × *S. rexii* F1 hybrid. **(d)** 20 DAU anisocotylous seedling of *S. rexii*. **(e)** 20 DAU anisocotylous seedling of *S. grandis*. **(f)** 20 DAU anisocotylous seedling of F1 hybrid. **(g)** 40 DAU anisocotylous seedling of *S. rexii*. **(h)** 40 DAU anisocotylous seedling of *S. grandis*. **(i)** 20 DAU anisocotylous seedling of F1 hybrid. Mc: macrocotyledon. mc: microcotyledon. Bar: 200 µm.

The differentiation in the groove meristem can be well illustrated under electron microscope (Figure 1.7). In 60 DAU seedlings of *S. rexii* and *S. rexii* × *S. grandis* F1 hybrid, a patch of small cells was seen on the adaxial side of the petiolode, adjacent to the proximal end of the macrocotyledon (Figure 1.7 a, c; arrows). The cells formed a bulging round-shaped cluster of around 150 µm to 200 µm in diameter (Figure 1.7 d, f). These cells presumably represent the groove meristem, and do not carry trichomes in contrast to the tissues surrounding them that are densely covered with short glandular and long eglandular trichomes. On the other hand, *S. grandis* seedlings showed no apparent sign of bulging in the groove meristem area (Figure 1.7 b), only a patch of cells that were not covered with trichomes, about 50 µm in diameter (Figure 1.7 b arrow). At 65 DAU seedlings of *S. rexii* and F1 hybrid showed apparent signs for the formation of a bulged GM of about 200 µm in diameter (Figure 1.7 f and g). The first primary phyllomorph emerged from the bulge, and showed adaxial-abaxial polarity with trichomes appearing from abaxial side (Figure 1.7 j, l, m; P1). On the other hand, the groove meristem area of *S. grandis* appeared flat and dormant at 65 DAU (Figure 1.7 e), 90 DAU (Figure 1.7 h), and 150 DAU (Figure 1.7 k). At the same time, the groove meristem area without trichome-growth seemed to enlarged in older materials. In 90 DAU *S. grandis* plant, the groove meristem area was about 100 µm in diameter and showed slightly bulge-shape morphology (Figure 1.7 h). In 150 DAU plant, the groove meristem area became about 200 µm in diameter though appeared to be flatten again (Figure 1.7 k).

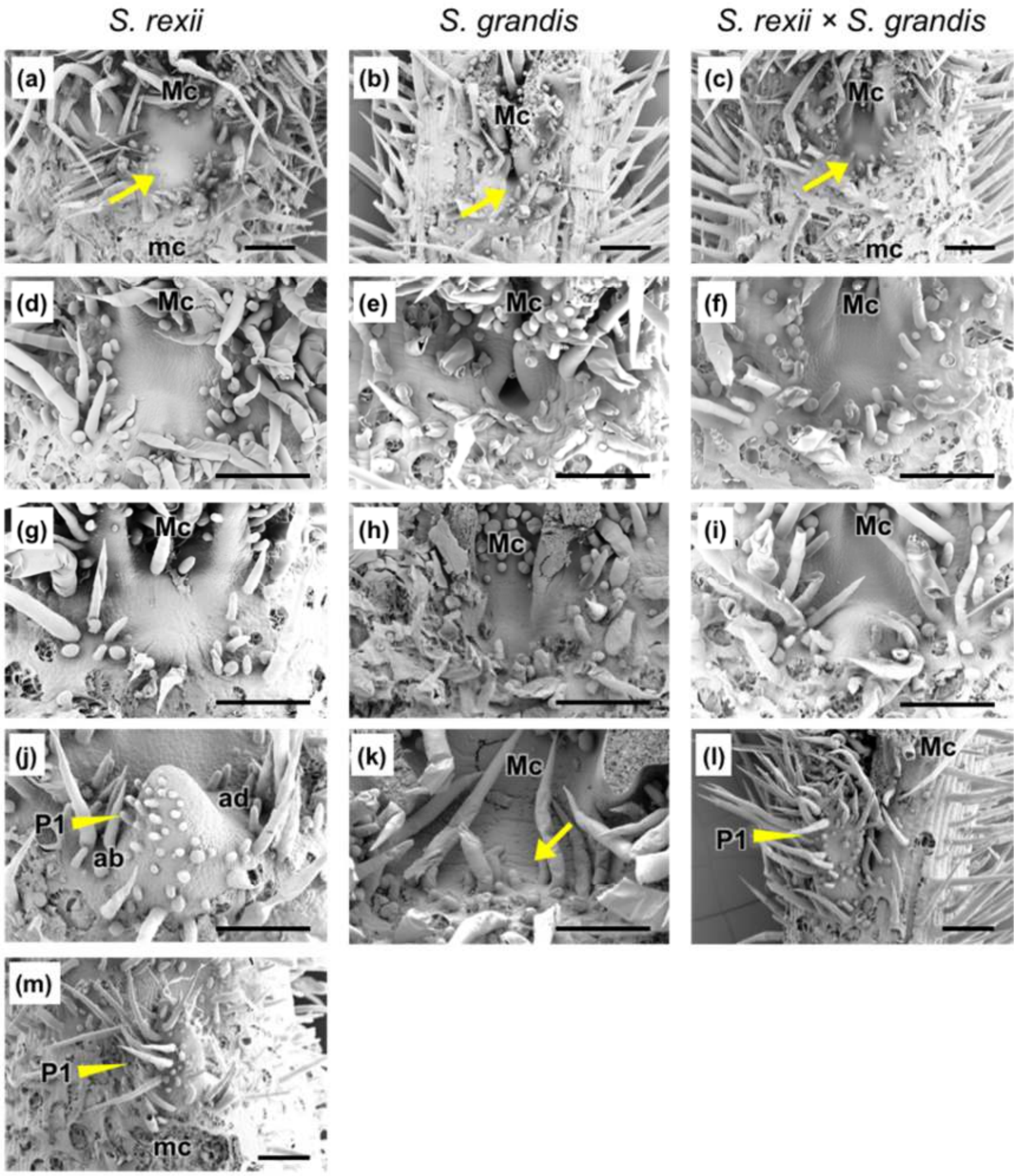


Figure 1.7 Development of the groove meristem of the three parental *Streptocarpus* parental lineages. All images are oriented to show the macrocotyledon (Mc) at the top. The microcotyledon and the trichomes surrounding the groove meristem tissue were removed. **(a)** *S. rexii* 60 DAU. **(b)** *S. grandis* 60 DAU. **(c)** F1 hybrid 60 DAU. **(d)** *S. rexii* 65 DAU, close

up view of the groove meristem. **(e)** *S. grandis* 65 DAU, close up of the groove part. **(f)** F1 hybrid 65 DAU, close up view of the groove meristem. **(g)** *S. rexii* 65 DAU, with a bulge shape groove meristem. **(h)** *S. grandis* 90 DAU, with a bulge shape groove meristem. **(i)** F1 65 DAU, with a bulge shape groove meristem. **(j)** *S. rexii* 65 DAU, with a growing phyllomorph primordium. **(k)** *S. grandis* 150 DAU, with flat groove meristem. **(l)** F1 65 DAU, with a developed primary phyllomorph. **(m)** *S. rexii* 65 DAU, with a developed primary phyllomorph. Mc: macrocotyledon. mc: microcotyledon, which was removed to reveal the groove meristem tissue. Yellow arrows: groove meristem. P1: primary phyllomorph. ad: adaxial. ab: abaxial. Bars: 200 µm.

As rosulates and unifoliate species of subgenus *Streptocarpus* both have the same chromosome count of *2n* = 32, viable off-springs can be produced (Lawrence et al., 1939; Möller and Pullan, 2015). Oehlkers (1938; 1942) carried out some of the earliest studies of the inheritance of the growth forms. In hybridisation experiments between rosulate and unifoliate *Streptocarpus*, all F1 hybrids were rosulate in form, indicating that rosulate is the dominant phenotype. Furthermore, their backcross progenies were reported to segregate in a Mendelian ratio, with the rosulate:unifoliate ratio of 3:1 in backcross progenies. And in F2 progenies, the segregation ratio was 15:1 (Table 1.1; Oehlkers, 1938; 1942; Harrison et al., 2005). Both ratios indicate that two unlinked genetic loci define the growth form, and that the growth form variation is a Mendelian trait (Oehlkers, 1938; 1942; Harrison et al., 2005). In addition, it was reported that one of the loci may act at an early stage, which resulted in rosulate individuals appearing at about 6 months after sowing, and the other locus at a later stage, at about 9 months after sowing (Oehlkers, 1942).

Table 1.1 Segregation ratios of rosulate × unifoliate experimental crosses

Population*	No. rosulate	No. unifoliate	Rosulate:unifoliate	Reference
S. wendlandii × (*wendlandii* × *rexii*)	318	120	3:1 (P = 0.246, X^2 test)	Oehlkers, 1938
(*S. wendlandii* × *rexii*) × (*rexii* × *wendlandii*)	48	3	15:1 (P = 0.9136, X^2 test)	Oehlkers, 1938
S. grandis × (*grandis* × *rexii*)	145	41	3:1 (P = 0.351, X^2 test)	Oehlkers, 1942
S. wittei × (*wittei* × *rexii*)	98	30	3:1 (P = 0.683, X^2 test)	Harrison *et al.*, 2005

* All species except for *S. rexii* (rosulate) are unifoliate

However, the identity of the two rosulate loci remains unknown to date, and plant hormones, sugar signalling and environmental factors may all have an effect on the rosulate and unifoliate morphologies. In terms of hormonal signalling, external treatment of gibberellin (GA) on unifoliate seedlings was found to induce the formation of an additional leaf (Dubuc-Lebreux, 1978; Rosenblum and Basile, 1984; Nishii et al., 2012a), and also induced the formation of an apical leaf bud in the rosulate *S. rexii* (Nishii et al., 2014). Sugar signalling may have a similar effect, with the treatment of β-glucosyl phenyglycoside (β-D-Glc)$_3$, a sugar molecule that specifically bind to the membrane Arabinogalactan-Proteins (AGPs), fascilitate the formation of additional leaf shoots in *S. prolixus*, which in natural condition produce 2 – 3 leaves (Rauh, 2001; Rauh and Basile, 2003). On the other hand, treatment of a structurally-similar β-galactosyl Yariv reagent that does not bind to AGPs failed to produce the same phenotype (Rauh, 2001; Rauh and Basile, 2003). In certain cases the growth form can be affected by environmental factors, such as in the caulescent species *Streptocarpus nobilis*, which when grown under adverse condition grows no additional leaves (Hilliard and Burtt, 1971).

A most direct attempt to identify the loci was the genetic association with the meristematic class I *KNOX* (*KNOXI)* gene. Mutation of the *KNOX* gene *SHOOT MERISTEMLESS* (*STM*) was shown to produce a phenotype lacking the SAM during embryogenesis in *A. thaliana* (Barton and Poethig, 1993). The gene homolog *SSTM1* was studied in the backcross populations *S. dunnii* × (*S. dunnii* × *S. rexii*) and *S. wittei* × (*S. wittei* × *S. rexii*) and were found expressing in the groove meristems, but was also found to be un-linked to the unifoliate growth form (Harrison, 2002; Harrison et al., 2005). Other developmental genes studied in *Streptocarpus* sp. include *WUSCEL* (as *SrWUS*; Mantegazza et al., 2009), *AS1* / *ROUGH SHEATH2* / *PHANTASTICA* (as *SrARP*; Nishii et al., 2010), *GA20-oxidase* and *GA2-oxidase* (as *SrGA20ox* and *SrGA2ox*; Nishii et al., 2014), and *ISOPENTENYLTRANSFERASE* (as *SrIPT*; Chen et al., 2017). Among these the gene transcripts of *SrWUS*, *SrARP*, *SrGA20ox*, *SrIPT5* and *SrIPT9* were found located in the groove meristem and basal meristems of rosulate *S. rexii* (while the gibberellin synthesising *SrGA2ox* expressed in the surrounding tissues outside of the meristem). In addition, the KNOX1 homolog STM was also found in the groove and basal meristems of the unifoliate *S. wendlandii* (Nishii et al., 2017).

1.2.3 Variation and inheritance of floral colour and morphology

The flowers of *Streptocarpus* species have several features in common: they are zygomorphic (with bilateral symmetry), gamopetalous, five-lobed, two-lipped, and are produced in pair-flowered cymes (Hilliard and Burtt, 1971; Haston and Ronse de Craene, 2007). On top of these features, the genus shows a wide range of variation in terms of floral dimension, corolla shape and colour, pigmentation pattern, and scent (Hilliard and Burtt, 1971; Harrison et al., 1999; Möller et al., 2019).

Hilliard and Burtt (1971) first described the floral types of *Streptocarpus* based on the shape of the corolla mouth. Three floral types were characterised as open (funnel shape flower), the key-hole-type (the opening of the tube is laterally compressed to a narrow slit), and personate (the ridges of lower lobe mask the corolla tube entrance). Later, Harrison et al. (1999) conducted measurements on the floral shape of 39 *Streptocarpus* and *Saintpaulia* species, and grouped them into six floral types based on morphometric analyses. These types included the open-tube type, the key-hole type, the personate type, the *Saintpaulia* type (reduced corolla tube, enantiostyly), the small pouch type (small size, pale colour and relatively wide tube), and the *dunnii* type (distinctive flower of *Streptocarpus dunnii* with red colour). Nishii et al. (2015) further expanded on this with two additional floral types, the Acanth-type (for the inclusion of the unusual corolla shape of *S. lilliputana* that matches certain genera in *Acanthaceae*) and the labellanthus-type (with a forward directing lip and reduced upper lip) (Figure 1.8). Since this categorisation did not account for the wide range of corolla shapes in the open tube, Möller et al. (2019) reassessed the flower classification and recognised seven main types of Nishii et al. (2015) and subdivided the open-tube type into six subtypes that included the Acanth-type and the new Acicularis-type.

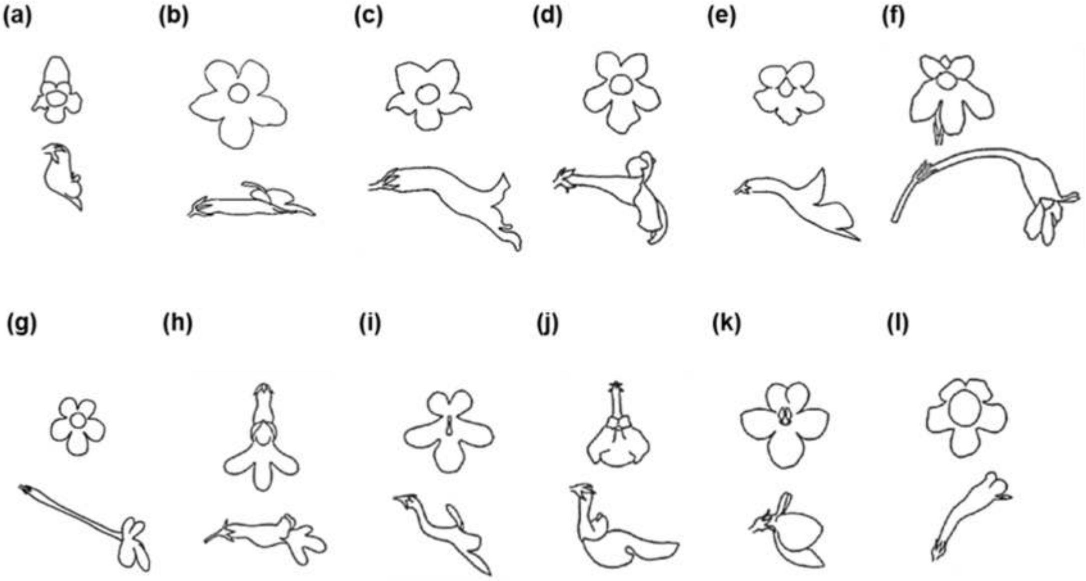


Figure 1.8 Different floral types of the genus *Streptocarpus* **(a)** Small pouch type, S. *beampingaratrensis* subsp. *beampingaratrensis* **(b)** Open cylindric tube, with narrow tube, *S. kentaniensis* **(c)** Open cylindric tube, with broad tube, *S. grandis* **(d)** Open tube with pollination chamber, *S. pumilus* **(e)** Inverted V-type, *S. wendlandii* **(f)** Acanth-type, *S. lilliputana* **(g)** Acicularis-type, *S. acicularis* **(h)** labellanthus-type, *S. thysanotus* **(i)** Key-hole type, *S. saxorum* **(j)** Personate type, *S. glandulosissimus* **(k)** Flat-faced type, *S. shumensis* **(l)**

Bird-pollination-type, *S. dunnii*. (Figure **(b)** **(c)** **(e)** were modified from Möller et al., 2019; figure **(g)** was modified from Darbyshire and Massingue, 2014; all other figures were modified from Nishii et al., 2015)

Between 1940 and 1960, a series of investigations were carried out by Lawrence to study the genetic inheritance of floral colour and pigmentation patterns of *Streptocarpus* species. These studies examined the segregation ratio of floral traits in multiple crosses, and aimed to find their inheritance patterns (Lawrence et al., 1939; Lawrence, 1947, 1957, 1958; Lawrence and Sturgess, 1957). *Streptocarpus* flowers usually have different intensities of purple to blue and pink shades with occasional exceptions, e.g. *S. lutea* and *S. bindseilii* which have ivory white flowers, and *S. dunnii* has red flowers (Hilliard and Burtt, 1971; Lawrence et al., 1939). The underlying pigmentation molecules are possibly delphinidin derivatives (a kind of anthocyanin), predominantly malvidin, followed by pelargonidin and peonidin (Scott-Moncrieff, 1936; Lawrence et al., 1939; Lawrence and Sturgess, 1957). The inheritance of floral colours was studied in *S. rexii* (blue) and *S. dunnii* (red) crosses, and several different colour classes were identified in the backcross and F2 populations, including blue, mauve, magenta, rose, pink, salmon and ivory (varying from the most intense blue to white, i.e. acyanic). Through studying the crosses and segregation ratios it was suggested that these colour classes were controlled by at least nine genes (Table 1.2; Lawrence et al., 1939; Lawrence and Sturgess, 1957). However, the identity of these loci remains unknown to date, and it has not been further examined using molecular marker or other genotyping methodology.

The genetics of the pigmentation patterns in the flower were also studied, including (1) the anthocyanin blotch in the corolla tube, (2) anthocyanin accumulation in glandular hairs of the pistil, (3) anthocyanin-coloured lines on the petals, and (4) the yellow pigment in the central stripe of the corolla tube (Lawrence, 1957). This study was carried out with multiple crosses between garden forms and acyanic forms of *Streptocarpus* (cultivars originated from hybridisation between *S. dunnii*, *S. rexii* and *S. parviflorus*). The inheritance of the anthocyanin blotch, hair colours (on the pistil), and anthocyanin lines on lower petals were all found to segregate in a 1:1 and 3:1 ratio in the backcross and F2 populations respectively (Lawrence, 1957). The yellow pigment (inside the lower petal of the corolla tube) was found to segregate in a 1:1 ratio in the backcross, but deviates from the expected 3:1 ratio in the F2 population, possibly due to a more complicated genetic basis. These characters also showed varying degrees of linkage (Lawrence, 1957), and it was concluded that the pigmentation patterns possibly follow Mendelian inheritance, and are probably controlled by a ‘supergene’ consisting of five individual genes that are closely located on a chromosome and are genetically linked (Table 1.3; Lawrence, 1957, 1958). Later, the inheritance of the yellow spot was studied in crosses using *S. rexii*, *S. parviflorus*, *S. montigena* and *S. cyaneus* (Oehlkers, 1966; 1967). The result suggests that the presence of the yellow spot is a dominant phenotype, and is likely to be monogenic with the segregation ratio of absence to presence of 1:1 and 3:1, in backcross and F2 population respectively, in contrast to Lawrence’s study.

Table 1.2 Summary of the hypothetical genes involved in *Streptocarpus* floral colouration

Hypothetic gene code	Hypothetic function	Phenotype	Reference
V	General production of anthocyanin in all tissues	*Dominant* Coloured (red) inflorescence and flower *Recessive* Green inflorescence stem and white flower	Lawrence and Sturgess 1957
F (or A)	General production of anthocyanin in flowers	*Dominant* Non-white flower *Recessive* White flower	Lawrence 1939 Lawrence and Sturgess 1957
I	Increase production of anthocyanin in flower	*Dominant* Medium to intense anthocyanin colour in corolla *Recessive* Pale anthocyanin colour in corolla	Lawrence and Sturgess 1957
C	Production of anthoxanthin co-pigment	*Dominant* Presence of anthoxanthin (white to yellowish pigments) *Recessive* Absence of anthoxanthin	Lawrence and Sturgess 1957
R	Convert pelargonidin to cyaniding	*Dominant* Presence of cyanidin *Recessive* Absence (or traces) of cyanidin	Lawrence 1939
O	Convert pelargonidin to delphinidin	*Dominant* Presence of delphinidin, resulted in mauve or blue flower *Recessive* Absence of delphinidin, resulted in lighter-colour flower	Lawrence 1939
D	Produce 3:5 dimonoside	*Dominant* Presence of solely 3:5 dimonoside pigments *Recessive* 3:5 dimonoside, 3-pentoseglycoside and 3-monoside mix	Lawrence 1939
X,Z	Complementary for 3:5 dimonoside production	*Dominant* Produce limited amount of 3:5 dimonoside pigments *Recessive* 3:5 dimonoside, 3-pentoseglycoside and 3-monoside mix	Lawrence and Sturgess 1957

Note. In the presence of both dominant R and dominant O, the pigment malvidin is produced

Table 1.3 Summary of the hypothetical genes involved in *Streptocarpus* floral pigmentation patterns

Hypothetic gene	**Hypothetic function**	**Phenotype**	**Reference**
B	Production of anthocyanin blotch at the anterior part of corolla	*Dominant* Presence of blotch or a deeper anthocyanin *Recessive* Absence of the trait	Lawrence 1957
H	Production of anthocyanin accumulation in hairs on the pistil	*Dominant* Presence of colour in the stalk of glandular hairs on pistil *Recessive* Absence of the trait	Lawrence 1957
L	Production of anthocyanin lines at the posterior part of lower petal	*Dominant* Presence of lines on the lower petal *Recessive* Absence of the trait	Lawrence 1957
Y	Production of yellow pigment down the central part of corolla	*Dominant* Yellow spot presence if both loci have at least one dominant allele *Recessive* Yellow spot absence if either of the loci are having two recessive allele	Lawrence 1957

In summary, *Streptocarpus* species show distinctive caulescent and acaulescent vegetative growth forms and diverse floral morphological characters. These morphologies are well documented, and preliminary genetic studies suggest the possible underlying genetic mechanism, i.e. two genetic loci for rosulate / unifoliate growth, nine genetic loci for floral colour, and a supergene for pigmentation pattern. However, the important question of physical identification of the actual loci remains unresolved. Gaining further knowledge of the genetic regulation of these phenotypic traits will increase our understanding of the evolution of the genus *Streptocarpus* and the Gesneriaceae family. It will also provide a broader understanding of how plant development is regulated to produce the unique morphologies observed in *Streptocarpus* and relate these to model plant systems.

1.3 Applications of next generation sequencing technologies to study interspecific genetics

1.3.1 Next generation sequencing technologies and Gesneriaceae resources

Next generation sequencing (NGS) refers to a wide range of high-throughput and in-parallel sequencing methods that emerged around 2004 (reviewed in Reuter et al., 2015; Kulski, 2016). These methods are distinct from the traditional Sanger sequencing method (Sanger et al., 1977) in their chemical reactions, and can generate giga base pairs (Gbp) of sequence data overnight at much lower cost. While the Sanger method sequences longer strands of DNA fragments based on the polymerase-chain reaction (PCR, usually around 1000 bp), NGS methods often involve shearing of DNA into much smaller fragments (from 25 bp to 500 bp), and each fragment is sequenced in parallel (reviewed in Glenn, 2011; Reuter et al., 2015). Thereby the number of base pairs (bp) sequenced per unit cost is greatly increased (Figure 1.9; Stein, 2010).

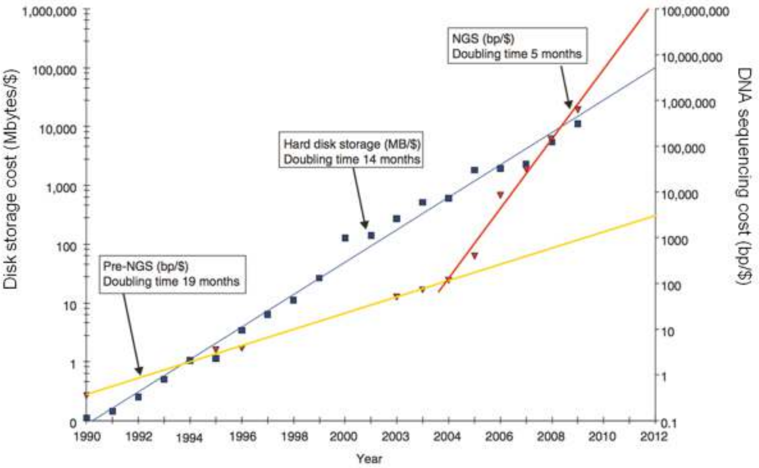


Figure 1.9 The trend of DNA sequencing cost versus the cost for hard disk storage. The cost is in US dollar (Stein, 2010)

Prior to the emergence of NGS, genome-scale sequencing and analysis have been restricted to a few selected model species, e.g. fruit fly, *Arabidopsis thaliana*, and humans (Adams et al., 2000; Arabidopsis Genome Initiative, 2000; International Human Genome Sequencing Consortium, 2001). However, with the lowered price of sequencing and the development of more user-friendly bioinformatics software, whole genome sequencing projects of non-model organisms became popular (Figure 1.10; Genome 10K Community of Scientists, 2009; Ellegren, 2014; Smith, 2016). As of April 2016, there were more than 60,000 prokaryotic genomes and over 2,700 eukaryotic genomes stored in GenBank, and between 2010 and 2015 alone, more than 2,000 mitochondrion genomes were published (Smith, 2016). Hence, the emergence of NGS technologies is an important milestone for genomic studies of non-model organisms (Sboner et al., 2011; Van Nimwegen et al., 2016).

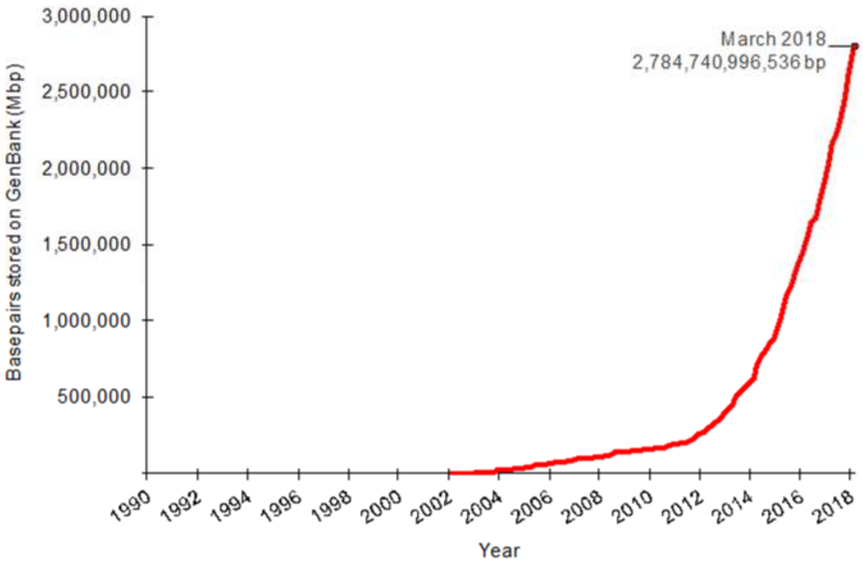


Figure 1.10 Number of base pairs (bp) stored in the GenBank database that were derived from whole genome shotgun sequencing experiments (from GenBank and WGS statistics https://www.ncbi.nlm.nih.gov/genbank/statistics/)

The usage of NGS technologies greatly accelerated research progress and resource availability in Gesneriaceae. For instance, the nuclear genome assembly of *Dorcoceras hygrometricum* (as *Boea hygrometrica*, see also Puglisi et al., 2016) is constructed (Xiao et al., 2015). The species belongs to the Loxocarpinae, closely related to subtribe Streptocarpinae where *Streptocarpus* resides (Möller et al., 2009), and has nine pairs of chromosomes (Kiehn et al., 1998). RNA sequencing-derived transcriptomes have been produced for several genera of Gesneriaceae, such as *Streptocarpus* (Chiara et al., 2013; Matasci et al., 2014), *Dorcoceras* (Xiao et al., 2015), and *Primulina* (Ai et al., 2014). A genetic map was constructed for the genera *Rhytidophyllum* and *Primulina*, using genetic markers derived from Genotyping-By-Sequencing (GBS) and transcriptome-derived single-nucleotide polymorphism (SNP) markers respectively (Alexandre et al., 2015; Feng et al., 2016).

1.3.2 Next generation sequencing genomic resources for *Streptocarpus*

Compared to other Gesneriaceae species or model plants, the available NGS derived genomic resources for the genus *Streptocarpus* are limited. A transcriptome of *S. rexii* is available at the online database ANGeLDUST (Chiara et al., 2013). The only genome resource available so far is the circular chloroplast sequence of *S. teitensis* (Kyalo et al., 2018). At the beginning of my PhD project, there was no nuclear genome reference or genetic map available for *Streptocarpus* (Chen et al., 2018).

Nevertheless, sequence resources are fundamental for genetic and genomic studies in the genus. Genome sequences can serve as backbone for reference-based SNP calling, and for the assembly of genotyping or RNA sequencing data (Davey et al., 2011, Korpelainen et al., 2014). A well annotated genome is very useful for the identification of functioning genes and gene structure (Ekblom and Wolf, 2014). With advanced NGS platforms such as Illumina HiSeq 4000 and HiSeq X, one lane of sequencing can generate up to 900 Gbp of data, that can provide ~900× depth of coverage for a 1 Gbp genome (Shen et al., 2014), which is suitable for assembly of the medium sized genome of *Streptocarpus* species which is on average ~0.8 Gbp for diploids (Möller, 2018).

Transcriptome profiles provide important information on the expressed genes in a genome. RNA sequencing (RNA-Seq) is a powerful approach for building a transcriptome database, which yields the sequence and structural information of genes. The RNA-Seq reads and the assembled transcriptome will be beneficial for annotating the nuclear genome (Hoff et al., 2016). In addition, the gene sequence information are valuable resources for gene isolation for future candidate gene studies (Wolf, 2013; Korpelainen et al., 2014). Well established bioinformatics tools are readily available for RNA-Seq data analysis, allowing the sequence to be assembled without the need of a complete reference genome (Haas et al., 2013), and annotation of the transcripts can be performed using existing pipelines (Conesa et al., 2005; Lohse et al., 2013; Kanehisa et al., 2016; Bolger et al., 2017b). Thus, RNA-Seq would be a feasible approach to generate the transcriptome database as a fundamental resource.

A genetic map is an essential resource for studying the genetic basis of phenotypic variation. It is required for mapping causative loci conferring a phenotype, such as quantitative trait loci analysis (QTL) or the mapping of simple-inherited trait loci (reviewed in Lynch and Walsh, 1998; Broman and Sen, 2009). Traditionally, most of these mapping experiments involved relatively labour-intensive genotyping methods, such as Restriction Fragment Length Polymorphism (RFLP), Amplified Fragment Length Polymorphism

(AFLP), or microsatellite markers (Kole and Abbott, 2008). The incorporation of NGS technologies allows sequence-based genotyping of a mapping population, and has been proven to be a successful approach for evolutionary genetic studies in non-model species (e.g. Chutimanitsakun et al., 2011; Kakioka et al., 2013; Palaiokostas et al., 2013; Campbell et al., 2014; Gonen et al., 2014). New methodologies, such as Reduced Representative Libraries (RRL; first described in Van Tassell et al., 2008), Restriction-site Associated DNA sequencing (RAD-Seq; first described in Baird et al., 2008), and Genotyping-By-Sequencing (GBS; first described in Elshire et al., 2011) enables the genotyping of thousands to tens of thousands of markers (Davey et al., 2011). This greatly enhances the linkage map density and the resolution of QTL mapping, with the great advantage that the data analysis can be done without the need of a complete reference genome (reviewed in Davey et al., 2011; Nielsen et al., 2011; Leggett and Maclean, 2014). Among these methods, the RAD-Seq approach has well developed and established analysis pipelines and has been successfully applied many times for constructing ultra-dense linkage maps (Davey and Blaxter, 2010; Davey et al., 2011; Catchen et al., 2011, 2013). It utilises the availability of diverse restriction enzymes, for fragmenting the genomic DNA and sequences hundreds to thousands of genetic markers (Reviewed in Lowry et al., 2016; Catchen et al., 2017; McKinney et al., 2017; Lowry et al., 2017). Thus, RAD-Seq is a promising approach for genetic mapping of traits for *Streptocarpus*.

The quality of NGS data is affected by the initial quality of DNA or RNA for library preparation (Healey et al., 2014). Some plant material may have high polysaccharide and secondary metabolite content, which affects the quality and quantity of extracted nucleic acids (Križman et al., 2006; Elshire et al., 2011). These contaminants inhibit the downstream experiments such as restriction digestion and PCR, thus reducing the efficiency of NGS library preparation (Zhang et al., 2000; Healey et al., 2014). The extraction of high molecular weight nuclear DNA is important for whole genome shotgun sequencing and RAD-Seq. Severely degraded DNA can cause the loss of genetic polymorphisms and loss of important genome information (Yang et al., 2014), and can also lead to reduced RAD-tags and sites of variance (Etter et al., 2011; Graham et al., 2015). Since DNA and RNA extraction methodologies for NGS experiments have not been established for *Streptocarpus*, different extraction methods will be tested and optimised in this thesis.

1.4 Objectives

In this study, essential genomic resources for genetic studies in the genus *Streptocarpus* will be acquired. Using NGS technologies, we will assemble reference genome sequences, transcriptome data, and build a genetic map for our target *Streptocarpus* species to carry out QTL mapping of the target traits. Two *Streptocarpus* species were chosen as the study material. One is the type species *Streptocarpus rexii*, which has an excentric rosulate growth form (Figure 1.11 a), with open-tube type flowers with pollination chambers (Figure 1.11 b). The other is *Streptocarpus grandis*, which has a unifoliate growth form (Figure 1.11 c), with open-type flowers with broad cylindrical tubes (Figure 1.11 d). Both species represent the section *Streptocarpus* in sub-genus *Streptocarpus* (Nishii et al., 2015), with 16 pairs of chromosome and viable interspecies hybrids can be produced

(Oehlkers, 1938; 1942; Möller and Pullan, 2015). The following are the specific objectives for my PhD project:

1. Determine the DNA and RNA extraction methods optimal for *Streptocarpus* NGS experiments.
2. Construct draft genomes for *S. rexii* and *S. grandis*.
3. Assemble transcriptomes of *S. rexii* and *S. grandis*, based on a range of tissue types to obtain as wide as possible gene expression profiles.
4. Calculate a genetic map for *Streptocarpus* using a mapping population generated from a backcross population (*S. grandis* × *S. rexii*) × *S. grandis.*
5. Perform QTL mapping for vegetative and floral characters, and search of candidate genes for future fine mapping approaches.

These data obtained in the study will not only be useful for the isolation of specific genetic loci in future studies, but will also serve as an important resource for future genomic studies across this morphologically challenging group.

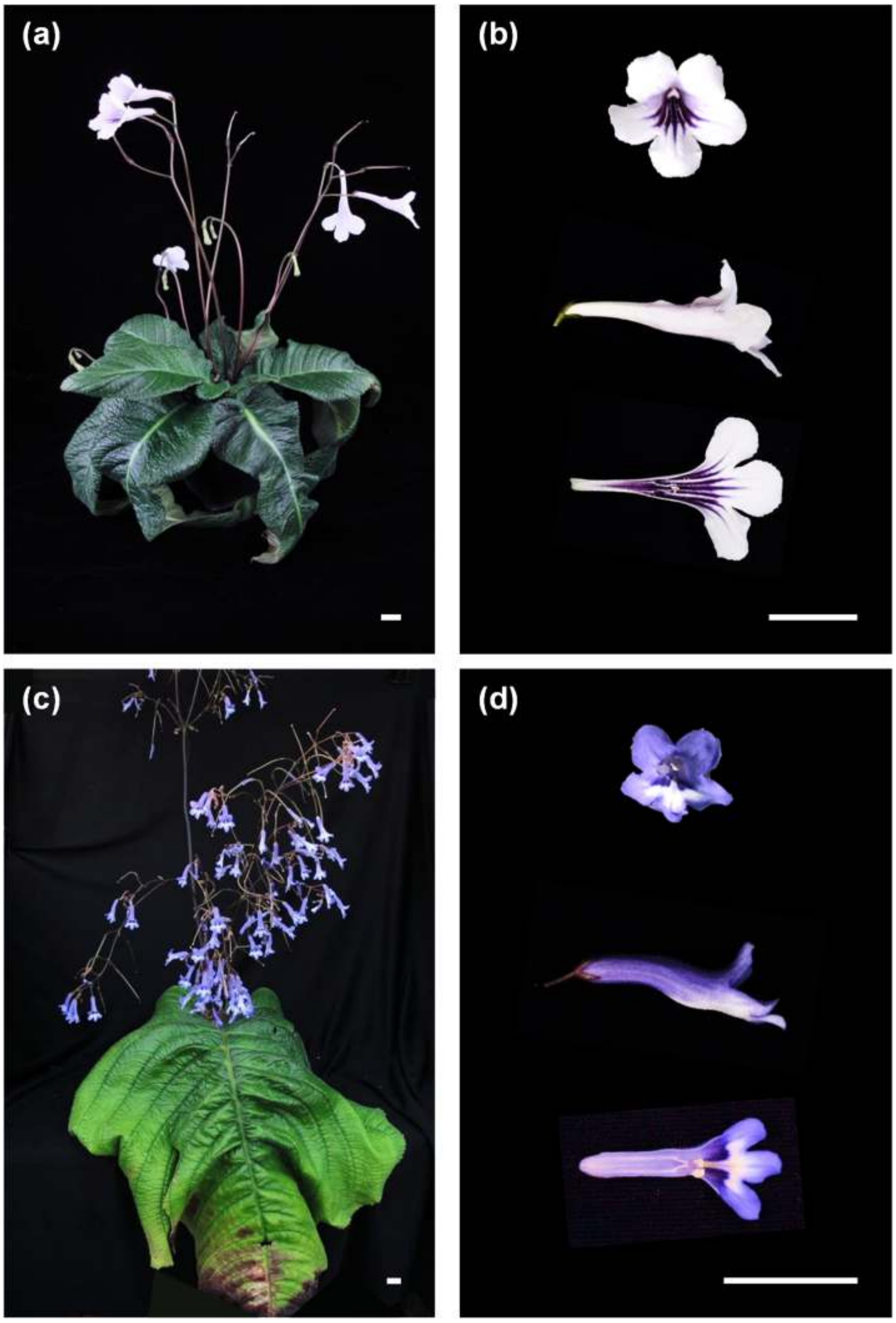


Figure 1.11 Study materials **(a)** *S. rexii*, mature flowering plant **(b)** Flower of *S. rexii*. Top: front view. Middle: Side view. Bottom: Ventral corollas of a dissected flower **(c)** *S. grandis*, mature flowering plant **(d)** Flower of *S. grandis*. Top: front view. Middle: Side view. Bottom: Ventral corollas of a dissected flower. Bars = 2 cm.

Chapter 2 Establishing DNA and RNA extraction methods for *Streptocarpus* for next generation sequencing (NGS)

2.1 Introduction

2.1.1 Impact of DNA and RNA quality on NGS experiments

High quality nucleic acids are an essential prerequisite for NGS experiments. Contamination and nucleic acid degradation during extraction can have profound negative impacts on an NGS run, such as reducing the efficiency of NGS library preparation (Zhang et al., 2000; Healey et al., 2014). DNA degradation may results in the loss of important genome regions to be sequenced, reducing the detection of genetic polymorphisms and number of single nucleotide polymorphisms (SNPs) recovered (Yang et al., 2014; Graham et al., 2015; Hart et al., 2016). Large amounts of DNA and RNA are required for library preparation and library quality check. For instance, a minimum amount of 1 µg of DNA is needed to prepare a TruSeq PCR-free whole genome shotgun sequencing library (User manual, Illumina, San Diego, CA, USA), on the contrary to traditional Sanger sequencing where as little as 100 ng of DNA is needed as sequencing template (Platt et al., 2007). Accurate quantification of the nucleic acids is also important, since pooling different samples with different amounts of DNA for the same sequencing run (e.g. RAD-Seq) can cause bias in sequencing, resulting in samples with lower DNA concertation having lower sequencing coverage, thus reducing the reliability of the genotyping results and number of markers recovered (Davey et al., 2011; Fountain et al., 2016).

Previous studies of *Streptocarpus* species have been restricted to non-NGS genotyping method or traditional Sanger sequencing approaches (e.g. Harrison et al., 2005), which does not have such strict sample quality and quantity requirements. In order to successfully carry out NGS experiments in the *Streptocarpus* materials, the establishment of DNA and RNA extraction methods suitable for NGS experiments were seen as a prerequisite for this project.

2.1.2 Quality requirement of DNA and RNA for NGS experiments

The quality and quantity of the nucleic acid sample can be checked by spectrophotometer, electrophoresis, and fluorometer (Endrullat et al., 2016; Hart et al., 2016). Spectrophotometer is used to assess the purity of the samples. It measures and calculates the absorbance ratio at specific wave lengths, i.e. A260/A280 and A260/A230. For pure DNA and RNA, the A260/A280 ratio is 1.8 and 2.0, respectively. On the other hand, the A260/A230 ratio should be around 2.0 for both DNA and RNA samples (Endrullat et al., 2016). For NGS experiments, the A260/A280 ratio should be in the range of 1.8 – 2.0, and the A260/A230 ratio between 2.0 – 2.2. A lower A260/A280 ratio suggests contamination such as polysaccharides, phenols or ethylenediaminetetraacetic acid (EDTA), and a lower A260/A230 ratio suggests the presence of proteins and phenols (Endrullat et al., 2016).

The integrity (absence of degradation) of samples can be checked by gel electrophoresis and the Agilent TapeStation system (Hart et al., 2016). For genomic DNA, the gel electrophoresis should show a sharply defined band at high molecular weight without smearing (degradation). For total RNA, the gel should show two intact bands, representing 18S and 28S rRNA that comprises 80-90% of the total RNA (Buckingham and Flaws, 2007). The TapeStation system gives a quantitative measurement of the integrity. The DNA Integrity Number (DIN) and RNA Integrity Number Equivalent (RIN^{e}) are scales ranging from 1 (severely degraded) to 10 (highly intact), thus higher value suggest better sample quality that is suitable for NGS experiment (Hart et al., 2016). For NGS samples, DIN and RIN^{e} values above 7 are recommended (Keats et al., 2018).

The concentration of the samples can be measured by using fluorescent dye and fluorometer (O'Neill et al., 2011; Simbolo et al., 2013). The Qubit assay system utilises a fluorescent dye that binds to DNA or RNA specifically. Once bound, the dye emits fluorescence which the intensity is fluorometrically measured. Thus, it provides an accurate and specific measurement of DNA and RNA concentration with minimised interference of other contaminants (O'Neill et al., 2011; Simbolo et al., 2013).

2.1.3 Nucleic acid extraction methods used for Gesneriaceae species

Plant tissues can contain high contents of polysaccharide and phenolic components that co-precipitate with nucleic acids, making the extraction of high quality DNA and RNA extra difficult (Križman et al., 2006; Elshire et al., 2011). C*etyltrimethylammonium* bromide extraction (CTAB; Doyle and Doyle, 1987) is a commonly used method for DNA extraction from plants, where the CTAB molecules trap proteins and polysaccharides and separate them from the nucleic acids (Tan and Yiap, 2009). However, for Gesneriaceae samples, a modified CTAB protocol or other extraction techniques are known to be used for NGS, which includes the phenol purification step (Allen et al., 2006), and was used for DNA extraction for *Dorcoceras hygrometricum* genome sequencing (Xiao et al., 2015). DNeasy silica-membrane spin columns were used for the extraction of DNA from *Rhytidophyllum* samples for genotyping-by-sequencing experiments (Alexandre et al., 2015); here, the DNA is bound to silica-membranes while the contaminants pass through and washed away (Tan and Yiap, 2009). For the sequencing of the *Streptocarpus teitensis* chloroplast genome, a magnetic beads-based extraction method was used (Kyalo et al., 2018): in this method, the DNA molecules are bound to magnetic beads coated with ligands or biopolymers. Contaminants are washed away with wash buffer, while the magnetic beads, with their DNA load, are immobilised by a magnet, thus the nucleic acid purified (Tan and Yiap, 2009).

For RNA extraction, the Sigma Spectrum Plant Total RNA Kit was used for sample preparation for RNA-Seq of *S. rexii* (Chiara et al., 2013). This method is based on silica column purification, though the protocol has only being tested on vegetative tissues so far, i.e. leaves and cotyledons (Chiara et al., 2013), but not on floral tissues. In our research group, RNA extraction is frequently carried out using guanidium isothiocyanate-phenol-chloroform extraction (Ullrich et al., 1977; Chomczynski and Sacchi, 1987; Nishii et al., 2010a), followed by acidic phenol:chloroform (5:1) purification and a final clean-up with the PureLink RNA Mini Kit (Invitrogen, Waltham, MA, USA). This method has been tested for the extraction of both floral and vegetative tissues of the *S. grandis* for RNA-Seq.

In this chapter, I tested and compared the DNA extraction methods mentioned above, in order to find the optimal DNA extraction protocols for the sample preparation for whole genome sequencing and RAD-Seq. The existing RNA extraction protocol was also tested on the *S. rexii* materials to evaluate the efficiency for the RNA-Seq sample preparation.

2.2 Materials and methods

2.2.1 Plant materials

All plant materials were grown and maintained in the Royal Botanic Garden Edinburgh (RBGE) research glasshouses. *Streptocarpus rexii* (RBGE accession 20150819) and *Streptocarpus grandis* (RBGE accession 20150821) were used for optimising the nucleic acid extraction methods. The materials were sown and grown from seeds. All samples were collected from young actively growing leaf or cotyledons with the length within 5 cm, except for *S. grandis* which only a single enlarged cotyledon can be used (Figure 2.1).

For DNA extraction, the leaf materials were collected from the proximal part of developing phyllomorphs, which is the area around the actively dividing basal meristem and groove meristem tissue: For *S. grandis*, the tissue was collected from the only phyllomorph, i.e. macrocotyledon (Figure 2.1 a). For *S. rexii*, young actively developing phyllomorphs were used, with the midrib length roughly 5 cm and smaller (Figure 2.1 b). To ensure uniform sample sizes, the lid of a sterile 2 ml Eppendorf tube was used to punch out discs of leaf tissue (Kim et al., 1997). The midrib and vein tissues were avoided, as they cause difficulty for grinding (Figure 2.1 c - f). For DNA extraction, the collected leaf discs were frozen immediately in liquid nitrogen after collection to prevent DNA degradation.

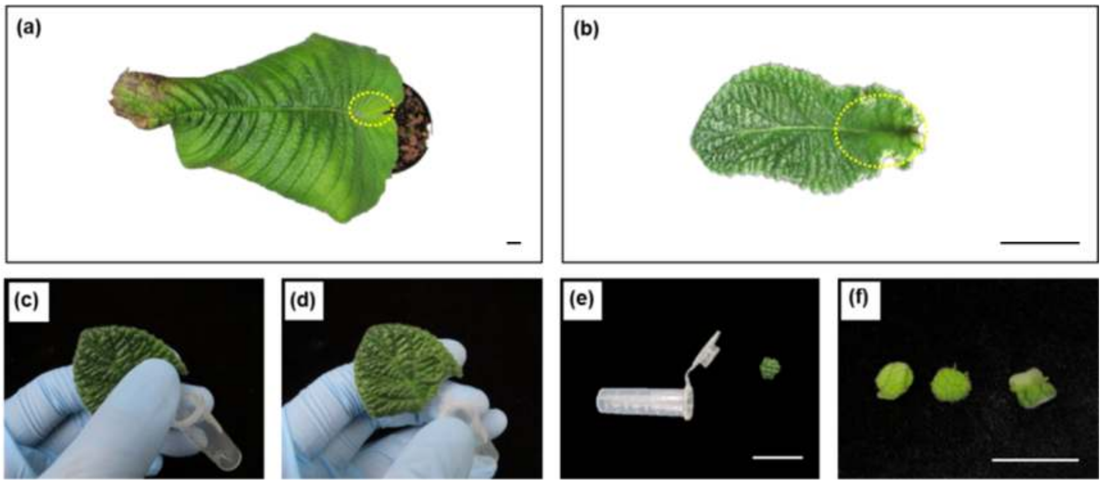


Figure 2.1 Standardisation of sampling of leaf disc material from *Streptocarpus* for DNA extraction. **(a)** Sampling area for *S. grandis*; tissue for DNA extraction was collected from the proximal area of the leaf (yellow dashed circle). **(b)** Sampling area for *S. rexii*; the tissue was collected from the proximal area of young actively growing leaves of 5 cm in length and smaller (yellow dashed circle). **(c)** – **(f)** Collection of leaf disc samples, **(c)** Leaf discs were punched out from leaves with the lid of a 2 ml Eppendorf tube. **(d)** The punched-out leaf disc was equal to the size of the lid. **(e)** The leaf disc was separated from the lid and collected for later DNA extraction. **(f)** The process was repeated until the desired amount of tissue was collected. Bars = 2cm.

For RNA extraction, leaf material was collected as described above. Root material was collected from leaf cuttings grown in perlite for about 2 weeks, and prior to RNA extraction, the roots were dug out and the perlite thoroughly washed off with tap water. Flower buds (length 0.5 – 5 cm), open flowers (length about 5.5 cm) and developing fruits (length 2 – 5 cm) were collected directly from flowering plants (Figure 2.2). For each tissue types, roughly 1 – 1.5 g (fresh weight) of materials were collected for 6 – 12 tubes of RNA extraction reactions (see section 2.2.3). All materials were frozen immediately in liquid nitrogen after collection.

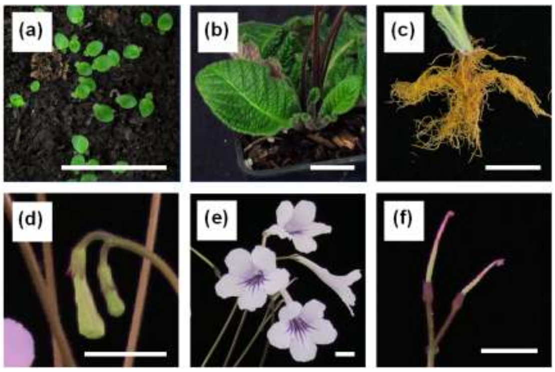


Figure 2.2 Different types of tissue of *S. rexii* used for RNA extraction and RNA-Seq. **(a)** seedlings approximately 30 days after sowing **(b)** Leaf tissues, collected from young developing leaves smaller than 1 cm in length to medium sized leaves up to 5 cm in length **(c)** actively growing roots from young adventitious plantlets **(d)** floral buds, 1-5 mm in length **(e)** open flowers **(f)** developing fruits. Bars = 2 cm.

2.2.2 DNA extraction protocols

Tissue grinding

For the testing of different extraction protocols, 1 to 2 tubes of extractions of each method were performed. For each tube of extraction, 2 – 4 leaf discs collected from young actively growing leaves were used. The tissues were ground using Eppendorf tube and pellet pestle (Sigma-Aldrich, Merck, Darmstadt, Germany) with liquid nitrogen.

For DNA extraction for whole genome sequencing, since a larger amount of tissues were required (i.e. 32 leaf discs for *S. grandis*, and 168 leaf discs for *S. rexii*), the tissues were ground using mortar and pestle with liquid nitrogen.

DNA extraction

(1) CTAB method (for detailed lab protocol format see Appendix 2.1)

The solutions and reagents required included 4% CTAB solution (100 mM Tris HCl pH 8.0, 1.4 M NaCl, 20 mM EDTA, 4% CTAB), β-mercaptoethanol (Sigma-Aldrich), chloroform:isoamyl alcohol (24:1), isopropanol (Sigma-Aldrich), wash buffer (10 mM ammonium acetate in 76% ethanol), and TE buffer (10 mM Tris, 1 mM EDTA, pH 8.0, Sigma-Aldrich).

2 µl β-mercaptoethanol and a 2% of PVPP were freshly added into 1 ml of 4% CTAB solution immediately before the start of the experiment. The solution was preheated to 65°C before tissue grinding. The ground leaf tissue was transferred to an 2 ml Eppendorf tube containing 1 ml of the pre-heated CTAB solution, and the mixture incubated in a 65°C heat-block for 60 minutes. During incubation, the mixture was occasionally mixed by inversion of the tube. After incubation, 500 µl chloroform:isoamyl alcohol (24:1) was added and the tube placed on an orbital shaker at minimum speed for 30 minutes. The tube was then centrifuged at 11,000 rpm for 10 minutes and the aqueous phase (~700 µl) transferred to a new 1.5 ml Eppendorf tube. The chloroform:isoamyl alcohol steps were repeated, with an equal amount (~700 µl) of chilled isopropanol added to the aqueous phase, and was mixed by inversion. The sample was then stored at -20°C overnight. The next day, the sample was centrifuged at 8,000 rpm for 10 minutes for pelleting the precipitate. The supernatant was discarded, and 500 µl of wash buffer added to the tube. The tube was shaken vigorously and incubated at room temperature for 30 minutes, followed by centrifuging at 8,000 rpm for 10 minutes. The supernatant was discarded and the DNA pellet dried using a SpeedVac concentrator (Thermo Fisher Scientific, Waltham, MA, USA) for 10 to 15 minutes. The dry pellet was finally dissolved in 100 µl TE buffer, and stored at -20°C.

(2) ChargeSwitch gDNA Plant Kit (for detailed lab protocol format see Appendix 2.2)

Two protocols based on the ChargeSwitch gDNA Plant Kit (Thermo Fisher Scientific) were tested; one following the manufacturer's instructions, and the other with modifications on the incubation time, which the lysis step was extended to 60 minutes and the rest of the incubation steps extended to 30 minutes (named the ChargeSwitch Kit$^{\text{Extended time}}$ protocol). The protocol is as follows: The ground leaf tissue was transferred to a 2 ml Eppendorf tube containing 1 ml of L18 lysis buffer. The sample was vortexed and incubated at room temperature for 1 hour. 100 µl of 10% SDS buffer was then added to the tube and incubated at room temperature for 30 minutes. 400 µl of N5 precipitation buffer (pre-chilled) was then added, and the sample incubated in ice for 30 minutes. The sample was centrifuged at maximum speed for 5 minutes, and the lysate transferred to a clean 2 ml Eppendorf tube. 100 µl of D1 detergent was added to the tube, followed by 40 µl of resuspended ChargeSwitch Magnetic Beads, and mixed by gentle pipetting. The mixture was incubated at room temperature for 30 minutes, followed by the use of the MagnaRack$^{\text{TM}}$ (Thermo Fisher Scientific) to pelletise the magnetic beads, thus separating the beads from the solution. The solution was discarded, and 1 ml of W12 wash buffer added to the tube and mixed with the magnetic beads by gentle pipetting. The MagnaRack$^{\text{TM}}$ was used again to remove the wash buffer, and the washing step was repeated once. After discarding the wash buffer, 150 µl of E6 elution buffer was added and well mixed with the magnetic beads by gentle pipetting. The mixture was incubated at room temperature for 30 minutes, and then the MagnaRack^{TM} was used to separate the beads from the DNA-containing elution buffer. The DNA elution was transferred to a clean 1.5 ml Eppendorf tube and stored in -20°C.

(3) DNAzol method (for detailed lab protocol format see Appendix 2.3)

The solutions and reagents required include the Plant DNAzol^{TM} reagent (Thermo Fisher Scientific), 100% ethanol, 75% ethanol, and TE buffer (10 mM Tris, 1 mM EDTA, pH 8.0, Sigma-Aldrich).

The ground leaf tissue was added to an Eppendorf tube containing 300 µl DNAzol reagent and the mixture vigorously shaken. The mixture was incubated with constant shaking for 30 minutes, followed by the addition of 300 µl chloroform and mixed vigorously, and shaken again for 5 minutes. The tube was then centrifuged at 10,000 rpm for 10 minutes, and the viscous supernatant transferred to a clean 1.5 ml Eppendorf tube. 225 µl of 100% ethanol were added to the tube and mixed by inverting the tube 6 to 8 times for the precipitation of the DNA. The mixture was incubated at room temperature for 5 minutes, followed by centrifugation at 7,000 rpm for 4 minutes. The supernatant was discarded, and the precipitated DNA was washed with freshly prepared wash buffer (contains 1 volume of DNAzol with 0.75 volumes of 100% ethanol). 300 µl of the prepared wash buffer was added to the tube containing the DNA pellet, and the tube was vortexed. The sample was kept at room temperature for 5 minutes, and then centrifuged at 7,000 rpm for 4 minutes. The wash buffer was discarded, and the pellet dissolved in 70 µl TE buffer and stored at -20°C.

(4) DNeasy Plant Mini Kit (for detailed lab protocol format see Appendix 2.4)

Two protocols based on the DNeasy Plant Mini Kit (Qiagen, Hilden, Germany) were tested; one following the manufacturer's instructions and the other with an extended incubation time of 30 minutes (DNeasy Kit$^{\text{Extended time}}$ protocol). The DNeasy Kit$^{\text{Extended time}}$ protocol was as follows: The ground leaf tissue was mixed with 400 µl of AP1 buffer in an 1.5 ml Eppendorf tube and mixed by vortexing. The mixture was incubated at 65°C for 30 minutes, with occasional inversion 2 to 3 times. 130 µl of P3 buffer was added and mixed, and the tube placed on ice for 5 minutes. The sample was then centrifuged at 13,000 rpm for 5 minutes, and the lysate transferred to a QIAshredder Mini Spin Column provided in the kit. The column was centrifuged at 13,000 rpm for 2 minutes, and the flow-through transferred to a clean 1.5 ml Eppendorf tube. 1.5 volumes of AW1 buffer were added to the flow-through and mixed well by pipetting. 650 µl of the mixture was then transferred to a DNeasy Mini Spin column, followed by centrifugation at ≥8,000 rpm for 1 minute, and the flow-through discarded. The step was repeated with the remaining mixture until all was processed.

500 µl of AW2 buffer was added to the column for washing, then centrifuged at ≥8,000 rpm for 1 minute and the flow-through discarded. The washing step was repeated twice. After discarding the flow-through from the second wash, the column was transferred to a clean 2 ml Eppendorf tube. 100 µl of AE buffer was added to the column and incubated at room temperature for 5 minutes to elute the DNA. The column was centrifuged at ≥8,000 rpm for 1 minute to collect the DNA solution.

RNase A treatment and phenol:chloroform:isoamyl alcohol purification

RNase A treatment and phenol:chloroform:isoamyl alcohol purification (PCI; 25:24:1) was performed for samples extracted using CTAB and ChargeSwitch methods. First, the volume of the DNA elution was adjusted to 300 µl by adding TE buffer. This is then followed by adding 2 µl of 4 mg / ml RNase A (#12091-021, Thermo Fisher Scientific; 1 / 5 dilution of the original stock using auto-claved distilled water) to the DNA and mixed by repeated tube inversions, and the sample was incubated at room temperature for 5 to 10 minutes. The RNase A reaction was stopped by adding 300 µl of PCI (pH 8.0) and mixed on an orbital shaker for 30 minutes. The sample was centrifuged at 11,000 rpm for 10 minutes and the aqueous phase transferred to a clean 1.5 ml Eppendorf tube (ap. 250 µl). The PCI step was repeated once. 0.1× volumes of 3 M sodium acetate (NaOAc, about 25 µl) were added to the sample, followed by 2.5× volumes of 100% ethanol and the solution mixed. The sample was kept at -20°C overnight for DNA precipitation. The sample was then centrifuged at 11,000 rpm for 10 minutes. The supernatant was discarded and 1 ml of 70% ethanol was added to the pellet and left for 30 minutes for washing. The sample was centrifuged at 11,000 rpm for 10 minutes. The supernatant was discarded and the pellet dissolved in TE buffer. The DNA eluate was incubated in a heat block at 65°C to help dissolve the DNA. The dissolved DNA was used for quality check.

In addition, the ChargeSwitch$^{\text{extended time}}$ protocol was eventually used for the DNA extraction for whole genome sequencing (detail extraction protocol described in Appendix 2.6).

2.2.3 RNA extraction protocol

RNA was extracted using the TRIzol reagent (Thermo Fisher Scientific) following the manufacturer’s manual with modifications. The RNA was then further purified using phenol:chloroform (5:1, pH 4.3 – 4.7, Sigma-Aldrich) and PureLink RNA Mini Kit (Thermo Fisher Scientific). The protocol is described in brief below and in detail in Appendix 2.7:

TRIzol extraction

The tissue samples for RNA extraction were freshly collected from healthy growing plants and frozen in liquid nitrogen immediately after collection. The tissue was ground in liquid nitrogen using a pestle and mortar. The ground tissue was transferred to a 1.5 ml Eppendorf tube containing 1 ml of TRIzol reagent. The sample was incubated at room temperature with constant gentle shaking on an orbital shaker for 50 minutes to 1 hour. The sample was then centrifuged at 11,000 rpm and 4°C for 10 minutes, and the supernatant transferred to a clean 1.5 ml Eppendorf tube. 200 μl of chloroform (BDH, VWR International, Radnor, PA, USA) were added to the sample and mixed by shaking the tube vigorously, and the mixture kept for 2 to 3 minutes before being centrifuged at 11,000 rpm and 4°C for 10 minutes. The aqueous phase was transferred to a clean 1.5 ml Eppendorf tube, and 500 μl of ice-cold isopropanol (Sigma-Aldrich) added for RNA precipitation. The sample was stored at -20°C for over 1 hour or overnight, followed by centrifugation at 11,000 rpm and 4°C for 10 minutes. After removing the supernatant, the pellet was dissolved in 50 μl of diethyl pyrocarbonate (DEPC) treated-water for preliminary quality check.

Phenol:chloroform (5:1) solution treatment

DEPC water was added to the RNA sample, or by RNA extracts from the same tissue type was combined, to make up the total volume of RNA extract to 300 μl per tube. 300 μl of phenol:chloroform (5:1, pH 4.3 - 4.7, Sigma-Aldrich) solution was then added to the sample and mixed by shaking the tube vigorously. The sample was centrifuged at 11,000 rpm for 10 minutes, and the aqueous phase transferred to a clean 1.5 ml Eppendorf tube. This step was repeated once. 300 μl of ice-cold isopropanol was added to the sample, and the sample stored at -80°C overnight for RNA precipitation. After the overnight incubation, the sample was centrifuged at 11,000 rpm for 10 minutes. The supernatant was discarded. This was followed by adding 500 μl of 75% ethanol to the sample for washing. The sample was centrifuged at 9,000 rpm for 5 min. The supernatant was discarded and the remaining liquid carefully removed with a pipette. The pellet was dissolved in 100 μl of DEPC water for preliminary quality checks.

PureLink RNA Mini Kit purification

The RNA sample was further purified using the PureLink RNA Mini Kit (Thermo Fisher Scientifc) with modifications of the original protocol. The lysis buffer was first prepared by adding 4 μl of β-mercaptoethanol (Sigma-Aldrich) into 400 μl of Lysis Buffer provided in the kit. The prepared lysis buffer was then added to the RNA sample, mixed by vortexing and incubated for 3 minutes. 200 μl of ethanol was added to the sample, and the sample mixture was transferred to the spin cartridge and centrifuged at 11,000 rpm for 1 minute. The flow-through was discarded and 600 μl of Wash Buffer I were added to the cartridge. The sample was centrifuged at 11,000 rpm for 1 minute. The flow-through was discarded and the collection tube (tube below the spin cartridge) was replaced with a clean one. 400 μl of Wash Buffer II was added to the spin cartridge and the sample centrifuged at

11,000 rpm for 1 minute. The flow-through was discarded, and the Wash Buffer II step was repeated once. The spin cartridge was centrifuged at 11,000 rpm for 2 minutes, and the column of the spin cartridge was transferred to a recovery tube. 40 µl of RNase-free water was added to the centre of the column and incubated at room temperature for 10 minutes to elute the RNA. The column was centrifuged at 11,000 rpm for 2 minutes and the collected RNA was used for final quality checks.

2.2.4 Quantification and quality control of the nucleic acid samples

For both DNA and RNA, the absorbance ratios (i.e. A260/A280 and A260/A230) of the extracted samples were measured using the NanoVue^{TM} Plus spectrophotometer (GE Healthcare, Chicago, IL, USA). The concentration measured by the spectrophotometer was also recorded. The final concentration was determined using the Qubit dsDNA HS Assay Kit or the Qubit RNA HS Assay Kit with the Qubit 2.0 fluorometer (Thermo Fisher Scientific). For the Qubit assay, 1/2 dilutions of the samples were first prepared and 2 µl of the dilution were loaded on the machine according to the manufacturer's manual. The integrity of the samples was first checked by agarose gel electrophoresis with 1% agarose gels at 100 volts for 45 minutes. Finally, the DIN and RIN^e values of the DNA and RNA samples were assessed using Agilent TapeStation system (Agilent Genomics, Santa Clara, CA, USA) installed in Edinburgh Genomics (The University of Edinburgh, Edinburgh, UK). The DIN and RIN^e values were only checked for the selected samples used for whole genome sequencing and RNA-Seq.

2.3 Results

2.3.1 Testing different DNA extraction protocols

The different DNA extraction protocols were first tested on fresh *S. grandis* material. The CTAB method resulted in an overall medium quality DNA (A260/A280 ratio around 2.0; A260/A230 ratio around 1.1 – 1.2) and medium quantity (4,000 – 5,000 ng of total DNA) with little difference between using 2 or 4 leaf discs (Table 2.1). However, the gel electrophoresis result showed intensely smeared bands at around 500 to 1,000 bp and below 200 bp (Figure 2.3 a). The ChargeSwitch Kit extraction results varied greatly depending on the protocol used: the original ChargeSwitch Kit protocol gave poor DNA quality and quantity (A260/A280 = 4.450; A260/A230 = 0.640; 660 ng of total DNA). By extending the incubation time, the ChargeSwitch Kit$^{\text{Extended time}}$ protocol gave the best DNA extraction results among the methods tested (A260/A280 = 2.144; A260/A230 = 1.739; 37,950 ng of total DNA). The gel electrophoresis of the ChargeSwitch Kit$^{\text{Extended time}}$ protocol showed a single intact band of DNA (Figure 2.3 b).

The DNAzol protocol extracted a large amount of DNA but of poor quality: The amount extracted DNA was high and ranged from around 18,000 to 26,000 ng with 2 and 4 leaf discs respectively, and the A260/A280 ratio was around 1.7, but the A260/A230 ratio was very low (< 0.8 for both 2 and 4 leaf discs). The gel electrophoresis showed a single band of DNA (Figure 2.3 c). The DNeasy Kit yielded the least usable DNA in terms of quality and quantity, with highly variable A260/A280 ratio (1.5 and 4.5 for 2 and 4 leaf discs respectively) and a much lower A260/A230 ratio compared the the requirement (0.017 and 0.018 respectively). The total amount of DNA extracted was also low (50 ng for both 2 and 4 leaf discs). The result improved slightly after the incubation time was extended but still showed poor quality and quantity (A260/A280 = 2.700; A260/A230 = 0.953; 2,075 ng of total DNA). The gel electrophoresis did not show any visible band (Appendix 2.8).

The two overall best methods (CTAB method and ChargeSwitch Kit) were also tested on the *S. rexii* material (Table 2.1 lower part). The performance of the CTAB protocol was similar to that in *S. grandis*, generating a medium quality and quantity of DNA (A260/A280 = 2.218; A260/A230 = 1.271; 3,050 ng of total DNA). The gel electrophoresis pattern is cleaner, with a single sharp high-molecular-weight band and only little smearing below 200 bp (Figure 2.3 b, left). The original ChargeSwitch protocol produced a poor quality and low quantity of extracted DNA similar to that in *S. grandis*. The ChargeSwitch Kit$^{\text{Extended time}}$ protocol again gave the best quality DNA (A260/A280 around 1.8; A260/A230 around 1.7 to 1.9) and quantity depending on the use of 2 leaf discs (3,750 ng total DNA) or 4 leaf discs (6,225 ng of total DNA). The gel electrophoresis result shows a single band with only weak smearing below 200 bp (Figure 2.3 b, right).

Table 2.1 Quality check results of DNA extracted from *S. grandis* and *S. rexii* using different protocols

Species	Starting material	Extraction method	A260/A280	A260/A230	Elution volume†	Total DNA extracted as measured by NanoVue (ng)
S. grandis	2 leaf discs	CTAB	2.062	1.250	50	4,125
S. grandis	4 leaf discs	CTAB	2.073	1.137	50	4,975
S. grandis	4 leaf discs	ChargeSwitch Kit	4.450	0.640	150	660
S. grandis	4 leaf discs	ChargeSwitch Kit$^{\text{Extended time}}$	2.144	1.739	150	37,950
S. grandis	2 leaf discs	DNAzol	1.718	0.202	70	26,757
S. grandis	4 leaf discs	DNAzol	1.734	0.734	70	18,165
S. grandis	2 leaf discs	DNeasy Kit	1.500	0.018	100	50
S. grandis	4 leaf discs	DNeasy Kit	4.500	0.017	100	50
S. grandis	2 leaf discs	DNeasy Kit$^{\text{Extended time}}$	2.700	0.953	100	2,075
S. rexii	4 leaf discs	CTAB	2.218	1.271	50	3,050
S. rexii	4 leaf discs	ChargeSwitch Kit	5.462	0.640	150	615
S. rexii	2 leaf discs	ChargeSwitch Kit$^{\text{Extended time}}$	1.805	1.689	150	3,750
S. rexii	4 leaf discs	ChargeSwitch Kit$^{\text{Extended time}}$	1.886	1.976	150	6,225

† The volume for elution varied among the different DNA extraction protocols

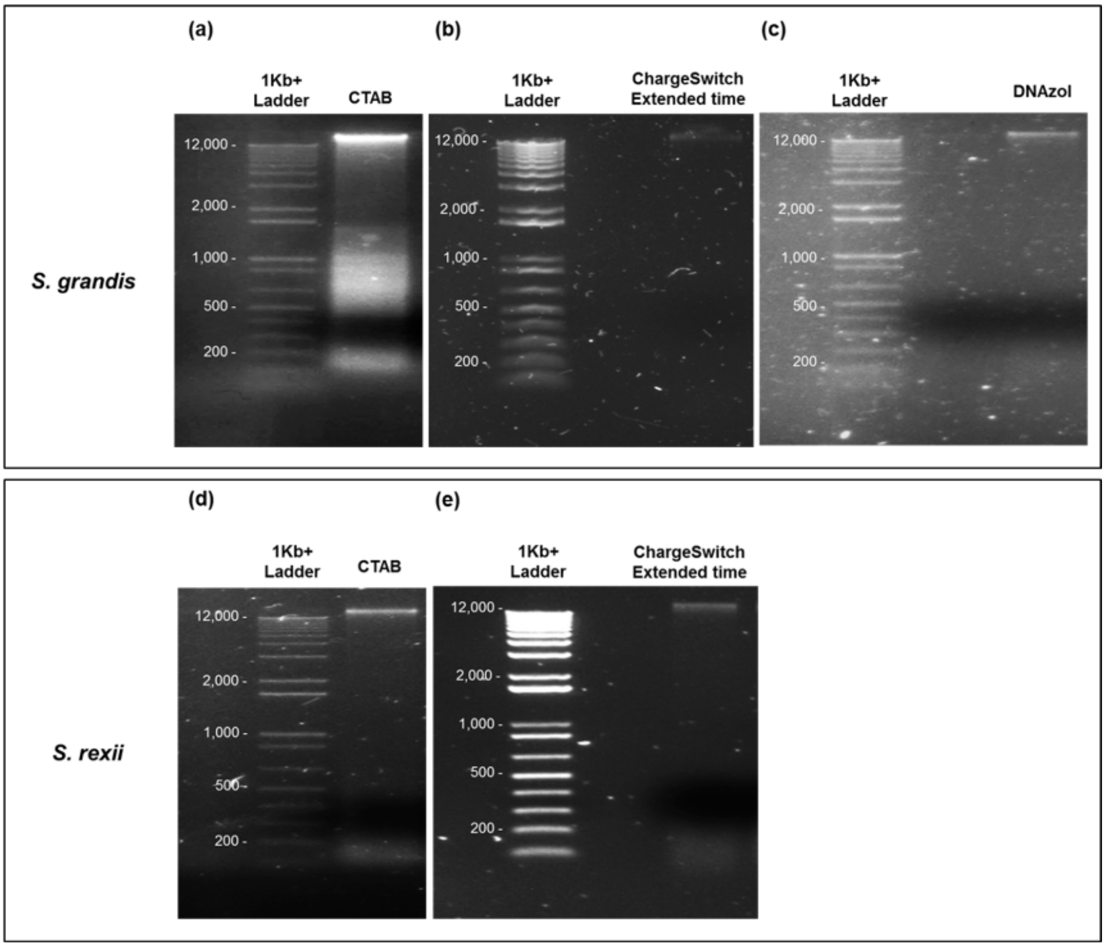


Figure 2.3 Gel electrophoresis results of DNA extracted using different extraction protocols. (a) *S. grandis* CTAB extraction. (b) *S. grandis* ChargeSwitch$^{\text{Extended time}}$ extraction. (c) *S. grandis* DNAzol extraction. (d) *S. rexii* CTAB extraction. (e) *S. rexii* ChargeSwitch$^{\text{Extended time}}$ extraction. Numbers beside the 1Kb+ Ladder indicates the molecular weight in base pairs.

To further purify the extracted DNA, RNase A treatment and PCI purification were performed on the extracted DNA using the CTAB and ChargeSwtich Kit$^{\text{Extended time}}$ methods in both *S. grandis* and *S. rexii* (Table 2.2). In *S. grandis*, the CTAB-extracted DNA had an A260/A280 ratio of 2.125 and an A260/A230 ratio of 1.708 prior to the treatment (Table 2.2). After the RNase A and phenol purification, the A260/A280 ratio was improved to 1.826 but the A260/A230 ratio decreased to 1.613, though the recovery rate of the DNA was about one fifth after the treatment (from 16,100 ng of DNA prior to 3,000 ng DNA post purification). For the ChargeSwtich Kit$^{\text{Extended time}}$ method, the A260/A280 and A260/A230 ratios were 2.144 and 1.739 respectively prior to the purification treatment. After the treatment, the values improved to 1.897 and 2.395 respectively.

The total amount of DNA extracted measured by the Qubit assay showed great discrepancies in values compared to NanoVue spectrophotometry (Table 2.2). In the CTAB extraction, the DNA measured by Qubit was 482 ng in total, which is only about 16% of the value measured by NanoVue (3,000 ng). In the ChargeSwitch Kit$^{\text{Extended time}}$ extraction, the Qubit measurement was 518 ng of total DNA, which was only about 3% of the value from NanoVue (17,130 ng). Gel electrophoresis indicated that RNase A and phenol purification successfully removed the smearing observed in both extraction methods prior purification (Figure 2.4 a).

For the *S. rexii* samples, the DNA quality also improved after PCI treatment (Table 2.2). In the CTAB extraction, prior to the treatment the DNA sample had A260/A280 and A260/A230 ratios of 2.218 and 1.271, respectively. After the treatment, the A260/A280 ratio decreased to 1.562, and the A260/A230 ratio increased to 1.444. The ChargeSwitch Kit$^{\text{Extended time}}$ method gave the best result; prior and after the treatment the A260/A280 and A260/A230 ratios were 1.886 and 1.728, and 1.976 and 2.188, respectively.

In the CTAB extraction the Qubit measurement for total extracted DNA (182 ng) was only about 14% of that measured by NanoVue (1,300 ng). In the ChargeSwitch Kit$^{\text{Extended time}}$ extraction, the amount measured by Qubit (328 ng) was about 21% of the NanoVue measurement (1,520 ng, Table 2.2). The smearing observed prior to the treatment was also removed in both extractions (Figure 2.4 b).

Table 2.2 Quality check results of the *S. grandis* and *S. rexii* DNA before and after RNase A treatment and phenol purification.*

Species	Extraction method*	A260/ A280	A260/ A230	Elution volume†	Total DNA extracted as measured by NanoVue (ng)	Concentration as measured by Qubit (ng/µl)	Total DNA extracted as measured by Qubit (ng)
S. grandis	CTAB	2.125	1.708	50	16,100	N/A	N/A
S. grandis	CTAB + RNase A + PCI	1.826	1.613	15	3,000	32.1	482
*S. grandis*º	ChargeSwitch Kit$^{\text{Extended time}}$	2.144	1.739	150	37,950	N/A	N/A
S. grandis	ChargeSwitch Kit$^{\text{Extended time}}$ + RNase A + PCI	1.897	2.395	20	17,130	25.9	518
*S. rexii*º	CTAB	2.218	1.271	50	3,050	N/A	N/A
S. rexii	CTAB + RNase A + PCI	1.562	1.444	15	1,300	12.1	182
*S. rexii*º	ChargeSwitch Kit$^{\text{Extended time}}$	1.886	1.976	150	6,225	N/A	N/A
S. rexii	ChargeSwitch Kit$^{\text{Extended time}}$ + RNase A + PCI	1.728	2.188	20	1,520	16.4	328

N/A - The values were not measured; PCI - phenol:chloroform:isoamyl alcohol treatment; * - starting material in all cases was 4 leaf discs. º - values taken from first experiment for comparison.

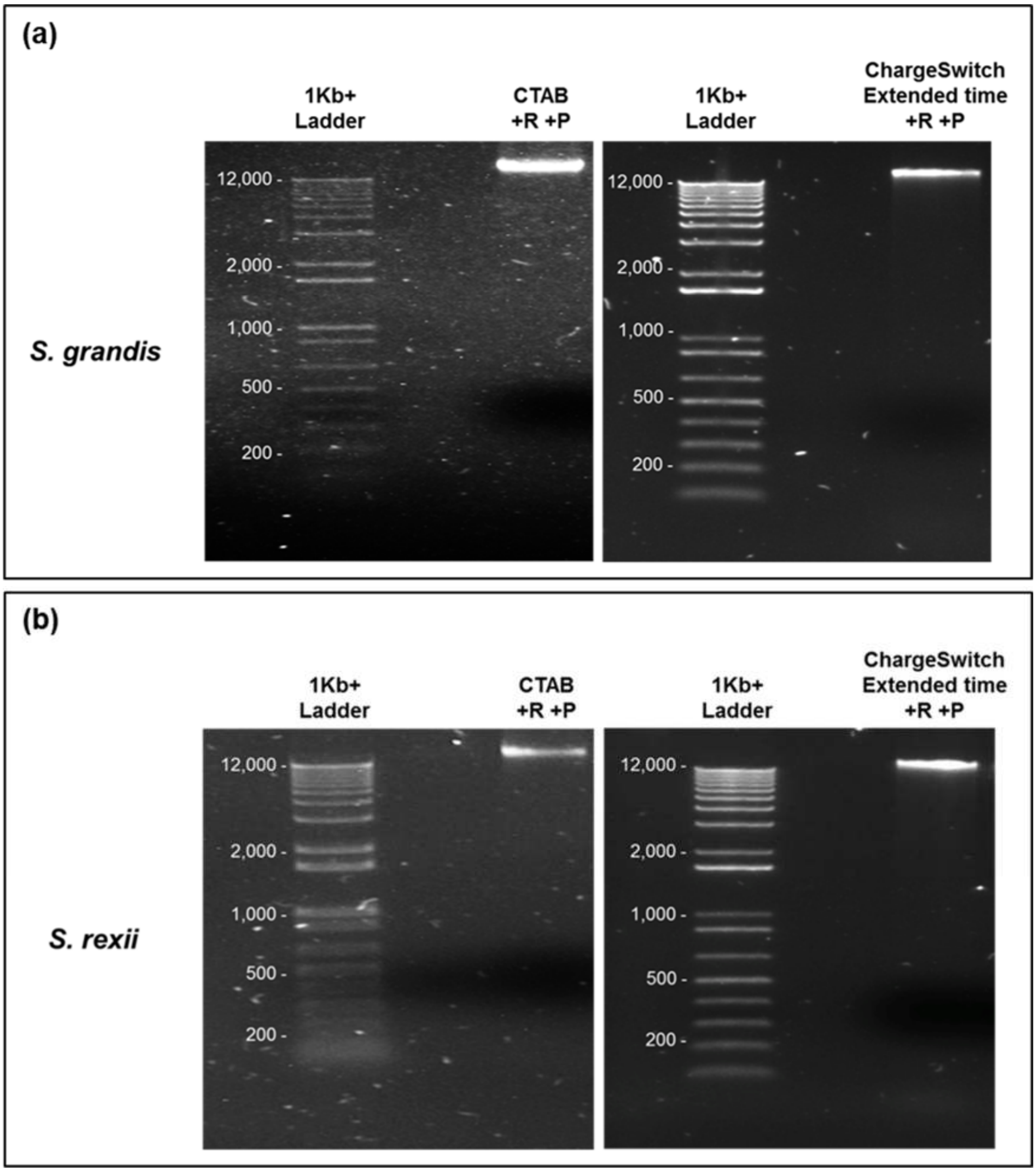
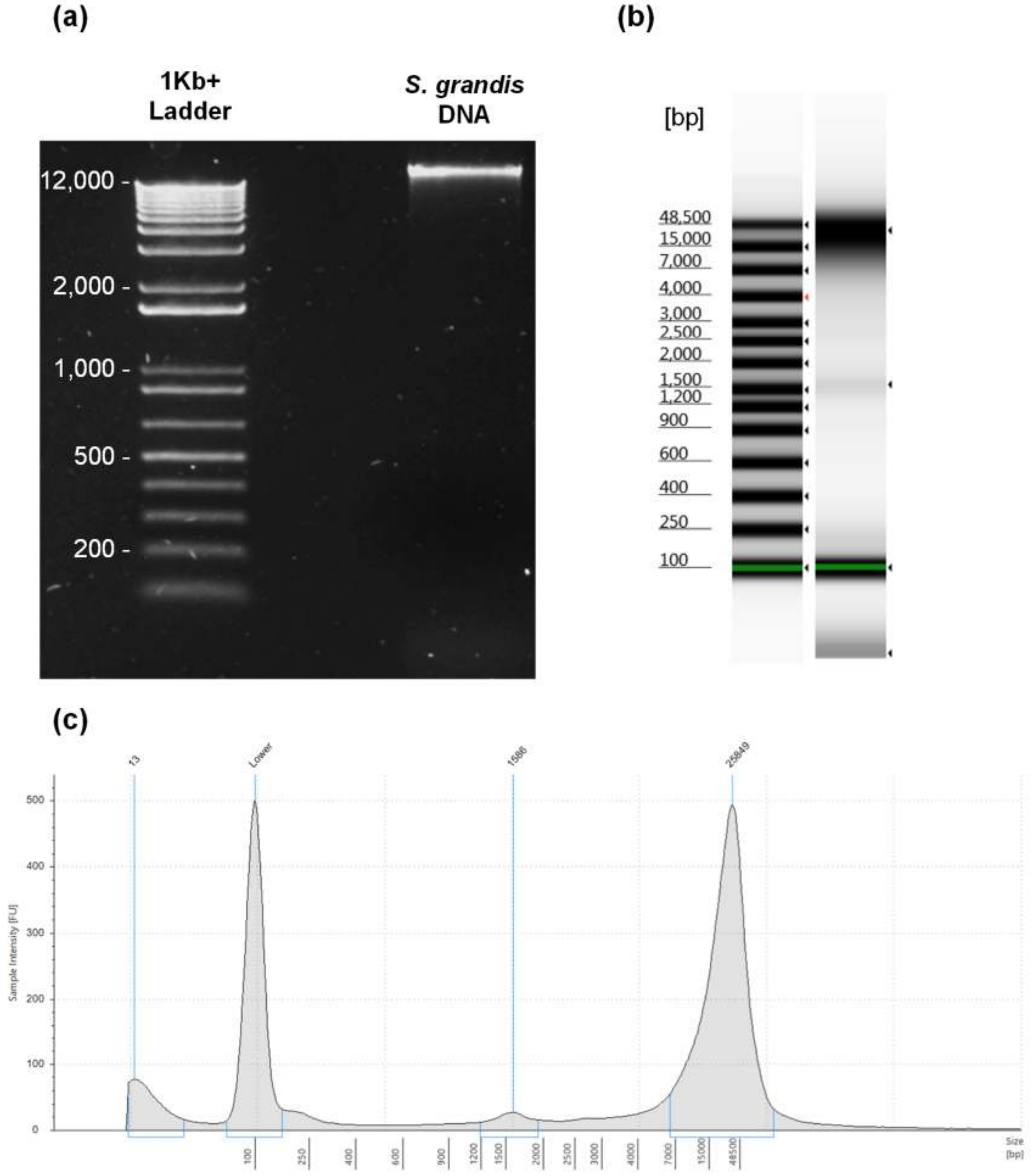


Figure 2.4 Gel electrophoresis results of the extracted DNA after RNase A treatment and phenol purification. **(a)** *S. grandis*. **(b)** *S. rexii*. Numbers beside the 1Kb+ Ladder indicates the molecular weight in base pairs. +*R* RNase A treatment, +*P* Phenol purification.

2.3.2 DNA extraction for whole genome sequencing of *S. grandis* and *S. rexii*

The ChargeSwitch Kit$^{\text{Extended time}}$ with RNase A treatment and phenol purification protocol (Appendix 2.6) was used to extract the DNA samples required for the whole genome shotgun sequencing of *S. grandis* and *S. rexii*. For the *S. grandis* extraction, 8 tubes of ChargeSwitch reactions (totally 32 leaf discs) were processed and the DNA combined. The extracted DNA had an A260/A280 ratio of 1.887, and an A260/A230 ratio of 1.879. The concentration measured using the Qubit assay was 37.1 ng/µl, and the total amount of DNA extracted was 2.7454 µg (37.1 ng/µl × 74 µl). On average, each ChargeSwitch reaction provided approximately 343 ng of DNA. The gel electrophoresis showed a single high-molecular-weight DNA band (Figure 2.5 a). The sample had a DIN value of 7.8 and successfully passed the quality control test required (Figure 2.5 b and c), and was used for the whole genome sequencing library preparation.

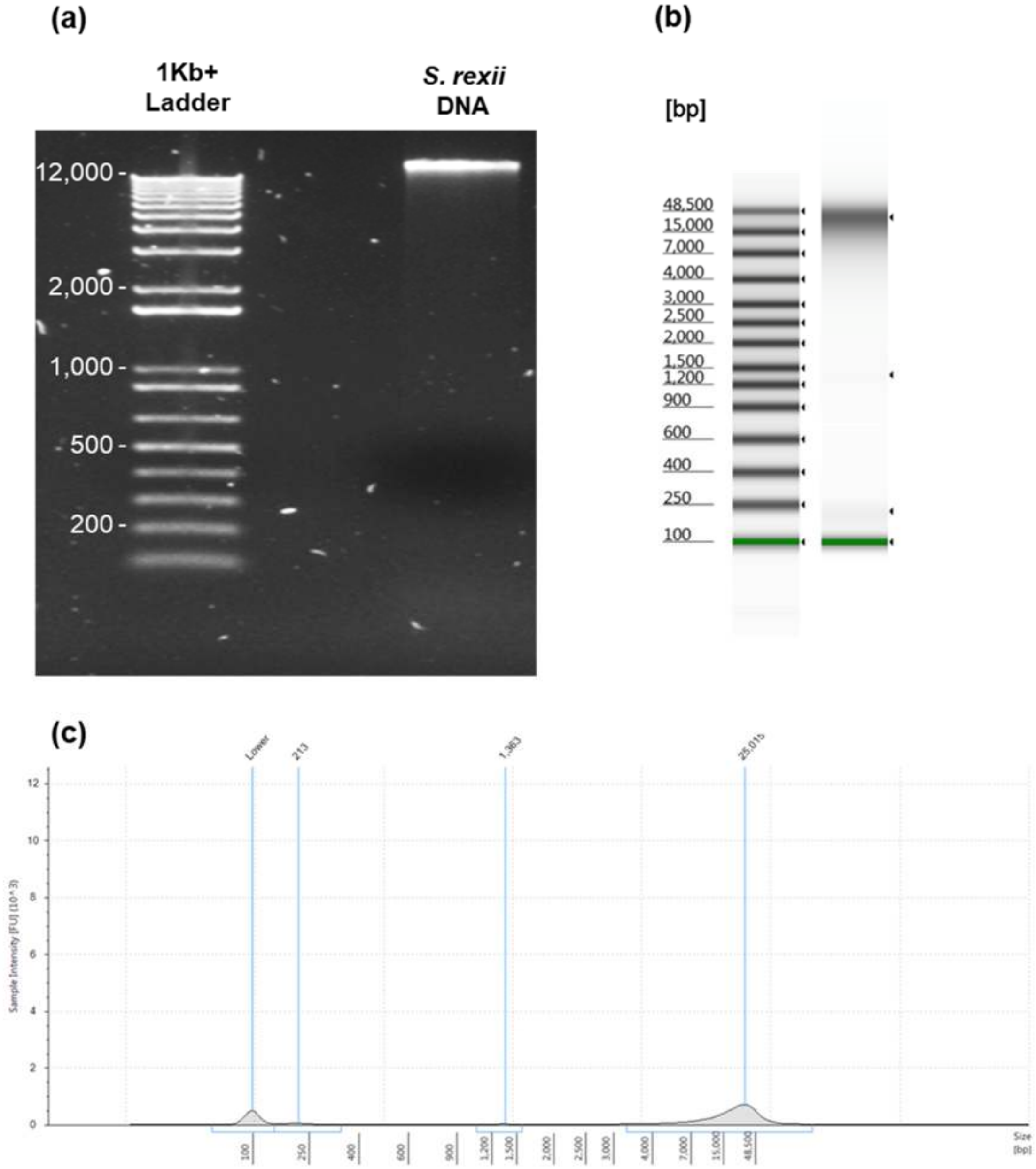
The same method was applied for the extraction of *S. rexii* DNA. 42 tubes of ChargeSwitch reactions were carried out and combined (about 168 leaf discs in total). The extracted DNA had an A260/A280 value of 1.898 and an A260/A230 value of 1.915. The Qubit concentration was 20 ng/µl, and the total amount of DNA extracted 9.565 µg (20 ng/µl × 478 µl, Table 2.4). On average, each ChargeSwitch reaction produced about 227 ng of DNA. The gel electrophoresis of the sample showed a single intact band (Figure 2.6 a). The TapeStation system gave a DIN value of 7.6 and a clear electropherogram (Figure 2.6 b, c). The sample successfully passed the quality control requirement and was used for the whole genome sequencing library preparation.



(d)

A230	A260	A280	A260/280	A260/230
8.933	16.792	8.892	1.887	1.879

Figure 2.5 Gel electrophoresis and TapeStation quality check results of the *S. grandis* DNA sample used for whole genome sequencing. **(a)** Gel electrophoresis image. **(b)** Gel image of the TapeStation run. **(c)** The electropherogram of the samples from the TapeStation run. **(d)** NanoVue measurement results.



(d)

A230	A260	A280	A260/280	A260/230
1.159	2.224	1.171	1.898	1.915

Figure 2.6 Gel electrophoresis and TapeStation quality check results of the *S. rexii* DNA sample used for whole genome sequencing. **(a)** Gel electrophoresis image. **(b)** Gel image of the TapeStation run. **(c)** The electropherogram of the samples from the TapeStation run. **(d)** NanoVue measurement results.

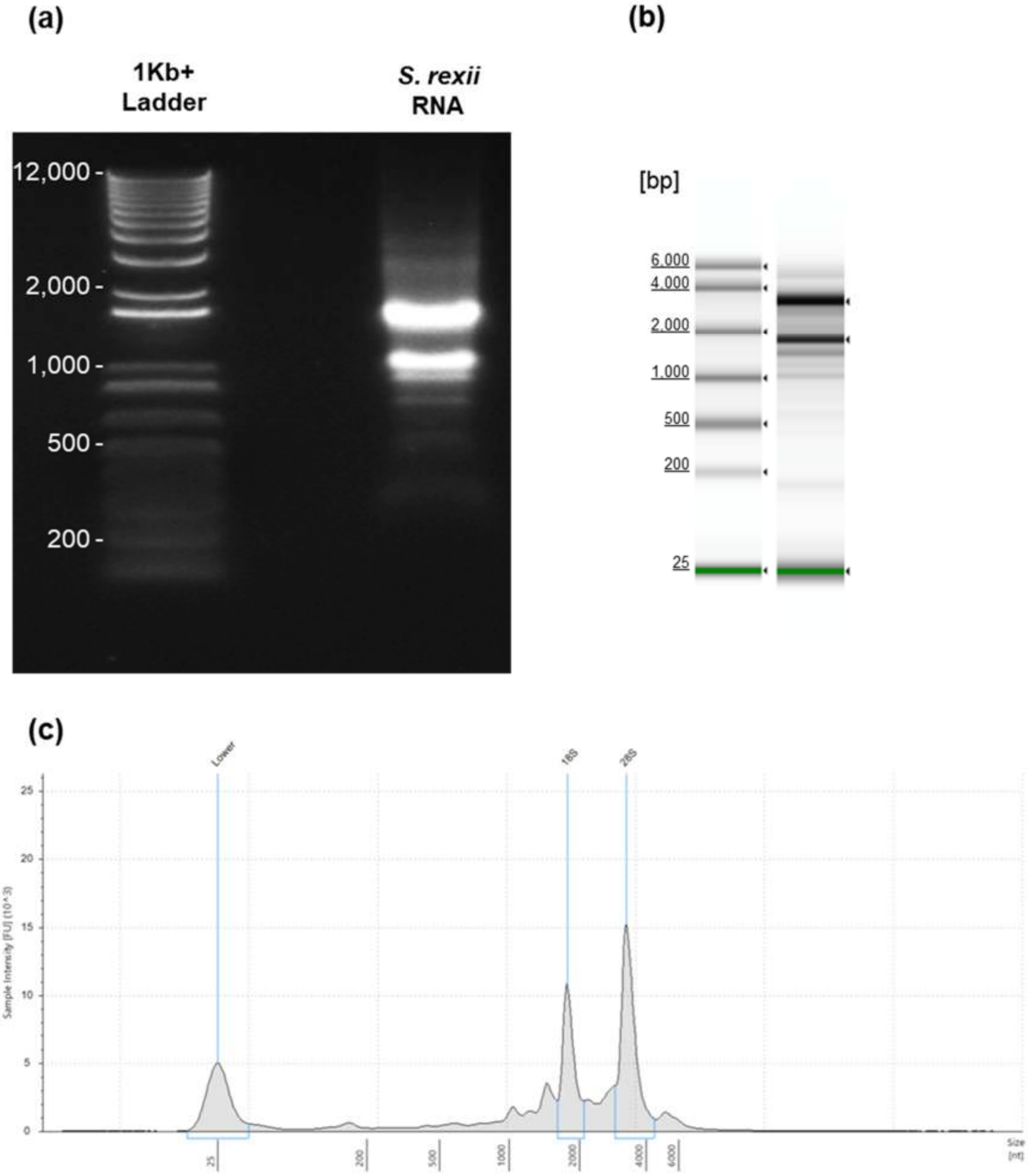
2.3.3 RNA extraction for RNA-Seq of *S. rexii*

RNA extraction was performed on different tissues of *S. rexii* (Table 2.3). Among these extractions, the seedling and flower tissues gave the best quality RNA, with an A260/A280 ratio around 2.0 and an A260/A230 ratio above 2.0. Leaves and root tissue extractions have a good A260/A280 ratio, but both showed lower than expected A260/A230 ratios (1.227 and 1.454, respectively). The floral bud and fruit showed poor quality RNA, with low A260/A280 and A260/A230 ratios (Table 2.3). In total, 3,936 ng of RNA were extracted according to the NanoVue measurement.

Table 2.3 Quality check results of the RNA extractions from different *S. rexii* tissues

Species	Tissue type	A260/ A280	A260/ A230	Elution volume (µl)	Total RNA extracted measured by NanoVue (ng)
S. rexii	Seedlings	2.114	2.070	30	901
	Leaves	2.245	1.227	30	152
	Roots	2.299	1.454	30	409
	Buds	1.147	1.146	20	1,179
	Flowers	1.981	2.304	20	1,017
	Fruits	1.478	0.831	15	278
Total				145	3,936

The resulting RNA samples were combined, and the quality and quantity measured again. The A260/A280 ratio was 2.117, and the A260/A230 ratio was 1.822. The concentration measured by Qubit was 602.5 ng/µl, and the total amount measured as 80,132 ng of RNA. The value is about 20-fold higher than the previous NanoVue measurements combined (totally 3,936 ng of RNA, Table 2.5). The gel electrophoresis showed a clear 18S and 28S rRNA banding pattern (Figure 2.7 a). The sample had a RIN^{e} value of 8.6 (Figure 2.7 b, c). This sample successfully passed the quality test and was used for library preparation for RNA-Seq.



(d)

A230	A260	A280	A260/280	A260/230
0.290	0.523	0.253	2.117	1.822

Figure 2.7 Gel electrophoresis and TapeStation quality check results of the *S. rexii* RNA sample for RNA-Seq. **(a)** Gel electrophoresis image. **(b)** Gel image of the TapeStation run. **(c)** The electropherogram of the samples from the TapeStation run. **(d)** NanoVue measurement results.

2.4 Discussion

2.4.1 Comparisons between different DNA extraction methods

Among all the DNA extraction protocols tested, the ChargeSwitch gDNA Plant Kit with extended incubation time (ChargeSwitch Kit$^{\text{Extended time}}$) followed by RNase A treatment and phenol purification gave the best DNA extraction results (Table 2.4). This protocol gave an intact high-molecular-weight DNA band, a reasonable DIN value, and absorbance ratios within the recommended ranges. The method was successfully applied for the extraction of the whole genome sequencing samples and passed the quality check. Thus, a method was established to successfully prepare NGS-grade quality DNA from *S. grandis* and *S. rexii*.

Extensive modifications of the protocol for the supplier of the ChargeSwitch Kit were required for maximal results, as the original protocol only extracted low quality and quantity of DNA from both *Streptocarpus* species (Table 2.1). One possibility is that the amount of leaf tissue used per extraction, 4 leaf discs, exceeds the suggested starting material amount (i.e. 100 mg, ChargeSwitch gDNA Plant Kit manual). The average weight per leaf disc of *S. grandis* and *S. rexii* was around 30 to 50 mg (Appendix 2.9), larger than the suggested 100 mg if four leaf discs were combined. Similar results were observed in CTAB extractions, which the DNA yield from two and four leaf discs were very similar (Table 2.1). Yet, since four leaf discs still gave higher DNA yield, four instead of two leaf discs was chosen for the rest of the DNA extraction testing. It should be noted that the original ChargeSwitch protocol has not been tested using appropriate amount of leaf disc yet (e.g. 2 or 3 leaf discs, Table 2.1).

In ChargeSwitch protocol, the potential limitation on starting material was overcame by extending the incubation time from the original 1 minute up to 60 minutes at the lysis step and 30 minutes at the rest of the incubation steps. After the modification, the quality of the DNA improved and the total DNA extracted increased. In addition, RNase A treatment and phenol purification were shown to be crucial for the sample preparation. Similar results were observed in CTAB extraction, which ribosomal RNA-like smear presented in the gel electrophoresis result prior to the treatment (Figure 2.3 a). The NanoVue A260/A280 ratio is also higher than 2.0 prior to RNase A, suggesting the presence of RNA in the sample. After the RNase A and phenol treatments, the smear disappeared and the NanoVue values improved (Figure 2.4 and Table 2.2).

In general, longer incubation time and phenol purification is required for extracting high quality DNA from the *Streptocarpus* materials. Even with these modifications, the performance of the CTAB extraction protocol is still unstable (e.g. varying total DNA yield and A260/A230 ratio; Table 2.1 and 2.2). There are many factors that may contribute to the final extraction quality and quantity, for instance, the age and general condition of the leaf material used, or high levels of phenolic compounds may present in the leaves tissues (Inoue et al., 1982; 1984; Sheridan et al., 2011). The *Streptocarpus* species, *S. dunnii* and *S. saxorum*, are known to have the phenolic compounds quinones (Hook et al., 2014). Furthermore, phenolic glucosides are known to present in many Gesneriaceae species that might be related to anti-fungal and anti-microbial activity (Verdan and Stefanello, 2012). It is not known whether these compounds accumulate in *S. grandis* and *S. rexii*, and whether the presence of these compounds affects the extraction quality or not. But in general phenolic and polysaccharides can make the DNA extracts appeared as viscous, glue-like, and brown in colour (Zhang et al., 2000; Healey et al., 2014).

Interestingly, DNA yield difference was found between *S. rexii* and *S. grandis* samples when both were extracted using ChargeSwitch$^{\text{extended time}}$ protocol (Table 2.2). *S. rexii* tissues tend to yield less DNA (totally ~9,500 ng DNA from 168 leaf discs; 56.5 ng per leaf disc) comparing to the *S. grandis* (totally ~2,700 ng DNA from 32 leaf discs; 84.3 ng per leaf disc). Since leaf tissues of similar age and properties (i.e. proximal leaf, near the actively dividing basal and groove meristem) were used in all extractions, it is unlikely that the difference was caused by sampling bias. One possible explanation is that the vein tissues of *S. grandis* are easier to remove comparing to the *S. rexii* due to their larger size. The vein tissues are difficult to grind, and if the majority of the vein tissues were excluded in *S. grandis* extractions, it may contributed better grinding and higher DNA recovery. Another possibility is the two species may share dissimilar physiological properties, such as different secondary metabolites, that may affect the purity of extracted DNA.

The established ChargeSwitch Kit$^{\text{Extended time}}$ protocol is more similar to the protocol used in Kyalo et al. (2018) for the sequencing of the *S. teitensis* chloroplast genome, where the DNA was extracted using the magnetic bead-based MagicMag Genomic DNA Micro Kit (Sangon Biotech Co.). On the other hand, while the CTAB protocol and the DNeasy Kit were reported to have worked for the DNA extraction from *Dorcoceras* and *Rhytidophyllum* species, respectively (Xiao et al., 2015; Alexandre et al., 2015), they failed to do so on the *Streptocarpus* materials in the present study. It is known that the CTAB method can remove neutral-pH polysaccharides, but cannot separate acidic polysaccharides from the DNA (Tan and Yiap, 2009). It is possible that the polysaccharides of the other Gesneriaceae genera have different pH properties to those in *Streptocarpus*, and thus CTAB protocol could worked but not in *Streptocarpus*.

A disadvantage of the ChargeSwitch Kit extraction protocol is the high unit cost per reaction; according to the results obtained here, each ChargeSwitch reaction extracted about 200 – 300 ng of DNA from 4 leaf discs, and each reaction costs about £3.00. For a genotyping experiment which requires the extraction of hundreds of samples with at least a minimum yield of 500 ng of DNA each (for e.g. RAD-Seq), the cost for the extraction kit itself would be over 1,000 pounds for 200 samples. In addition, the CTAB-extracted-DNA was successfully digested when tested with restriction enzyme, which is a key step during RAD-Seq library preparation (Baird et al., 2008; Peterson et al., 2012). Thus, the modified CTAB method was later chosen to be used for the DNA extraction of the mapping population for RAD-Seq experiments (Table 2.4 and Appendix 2.5, see details in Chapter 5). On the other hand, the ChargeSwitch Kit$^{\text{Extended time}}$ protocol remained the best method for DNA extraction and was used to prepare the DNA samples for whole genome shotgun sequencing.

Table 2.4 Comparison of the DNA extraction results for different methods for the *Streptocarpus* leaf material

Protocol	Results	Usage in this thesis
ChargeSwitch Kit$^{\text{Extended time}}$ + RNase A treatment + phenol purification (Appendix 2.6)	A260/A280 1.7 to 1.9 A260/A230 ~2.0 Extracts ~300 ng DNA from 4 leaf discs	Used to extract DNA from *S. grandis* and *S. rexii* for whole genome sequencing (see Chapter 3)
Modified CTAB method + RNase A treatment + phenol purification (Appendix 2.5)	A260/A280 1.5 to 1.9 A260/A230 < 2.0 Extracts ~500 ng DNA from 4 leaf discs	Used to extract DNA from the mapping population for RAD-Seq (see Chapter 5)
ChargeSwitch Kit	A260/A280 > 4 A260/A230 < 1 Very low DNA recovery	N/A
DNAzol	A260/A280 ~1.7 A260/A230 < 1 High DNA recovery	N/A
DNeasy Kit$^{\text{Extend time}}$	A260/A280 2.7 A260/A230 < 1 Low DNA recovery	N/A
DNeasy Kit	Highly variable A260/A280 Very low A260/A230 Very low DNA recovery	N/A

2.4.2 Revaluating the RNA extraction protocol

The RNA extraction protocol established for RNA-Seq of *S. grandis* (Appendix 2.7) was shown to be suitable for the extraction from *S. rexii* materials (Figure 2.7), although quality variation was observed among the different tissue types (Table 2.3). The same observation was made in other crop species, such as strawberries and cardamom, in which the RNA extracted from fruits are particularly low in quality even when using optimised protocols (Nadiya et al., 2015; Christou et al., 2014). In *S. rexii*, our method was most effective for the extraction of RNA from seedlings and open flowers, the products of which had reasonable absorbance ratios. The method was applicable to the developing leaves and root tissues, but the extracted RNA had low A260/A230 ratios. On the other hand, the method only resulted in RNA of poor quality from the floral buds and developing fruits (Table 2.3). This implies that current protocols cannot fully remove the polysaccharide and phenolic contaminants. Some studies suggested that CTAB-based methods produce high quality RNA in the presence of high concentration of PVP and β-mercaptoethanol (Nadiya et al., 2015; Sánchez et al., 2016). Other methods, such as the RNeasy Kit, may be an option to be tested for future extractions. Nevertheless, our combined RNA samples passed the NGS quality requirements, indicating that the current protocol is suitable for our material.

2.4.3 Discrepancies between quantification by spectrophotometer and Qubit assay

The DNA concentrations measured by Qubit assay were significantly lower than the measurements obtained from the NanoVue spectrophotometer (Table 2.2). It is frequently reported that the NanoVue overestimates DNA concertation, as the spectrophotometer cannot distinguish target nucleic acids from the accompanying contaminants that also absorb lights of 260 nm wavelength (O'Neill et al., 2011; Simbolo et al., 2013). On the other hand, Qubit assay has been demonstrated to be more accurate, and the measurement result is closer to that obtained from highly-specific PicoGreen nucleic acid quantification method (O'Neill et al., 2011; Garcia-Elias et al., 2017). This suggests that there may still be contamination remained in the ChargeSwitch-extracted DNA samples, thus causing the overestimation of DNA concentration in NanoVue measurement.

However, NanoVue may have underestimated the RNA concentration of our RNA extractions, which the NanoVue measurement was 20 folds lower than the concentration obtained from Qubit assay (Table 2.3). This is unlikely to be an experimental error, as it was observed repeatedly in other *Streptocarpus* RNA extractions (> 25 species; unpublished data). It was reported that the secondary structure of RNA may prevents UV absorbance thus lowering the NanoVue measured quantity, and the solution is to denature the RNA sample at 70°C for 2 minutes prior to the measurement, which may enhance the accuracy and increase the measured concentration by up to 25% (Aranda et al., 2009). Still, this does not explain the 20 folds difference observed in our measurements. Overestimation of the RNA concentration by Qubit assay has so far only been reported in non-peer reviewed technical reports (Fischer et al., 2016). Comparisons of multiple quantification methods (e.g. qPCR and PicoGreen) on quantifying serial diluted samples may be needed to determine which methods can more accurately reflect the actual RNA concentration (Aranda et al., 2009; Garcia-Elias et al., 2017).

2.4.4 Conclusion

Nucleic acid sample preparation is the first step of any sequencing experiment. Here the optimised DNA and RNA extraction methods for *Streptocarpus* materials were found, which were successfully used to extract DNA and RNA suitable for NGS experiments. The ChargeSwitch Kit$^{\text{Extended time}}$ protocol was selected for DNA extraction for whole genome shotgun sequencing, and a modified CTAB method with RNase A treatment and phenol purification was chosen for the DNA extraction for RAD-Seq experiments. For RNA extraction of sufficient quality for RNA-Seq for *S. rexii*, the protocol set up for *Streptocarpus* materials by the Royal Botanic Garden Edinburgh Gesneriaceae research group was found suitable. These methods were applied in sequencing experiments reported in the following chapters of this thesis.

Chapter 3 Building genome resources – Genome sequencing and *de novo* genome assembly of *Streptocarpus rexii* and *S. grandis*

3.1 Introduction

3.1.1 Genome size and chromosome count of *S. rexii* and *S. grandis*

Plant genome size and complexity vary greatly among species. For example, the smallest angiosperm genome currently known is that of the *Genlisea margaretae* approximately 63 mega base pairs (Mbp) (Greilhuber et al., 2006), while the largest genome reported so far is that of the octoploid lily *Paris japonica* with about 148.8 giga base pairs (Gbp) (Pellicer et al., 2010). Large and polyploid genomes are more costly in terms of sequencing to the same depth of coverage comparing to smaller genome, and are more difficult to assemble comparing to smaller and diploid genomes (Li and Harkess, 2018). The *Streptocarpus* materials chosen in this study, *S. rexii* and *S. grandis*, have medium sized genomes among angiosperms. The monoploid genome contents (1C value) are 0.95 pg and 1.289 pg respectively (Möller, 2018), which correspond to an estimated genome size of 929.1 Mbp for *S. rexii*, and 1,260.6 Mbp for *S. grandis* (Cavaller-Smith, 1985). *Streptocarpus rexii* and *S. grandis* both belong to the subgenus *Streptocarpus*, and are both diploid with 16 pairs of chromosomes (2n = 32; Lawrence et al., 1939).

3.1.2 Currently available genome resources for Gesneriaceae and Lamiales

Genome assemblies are available for one Gesneriaceae and several Lamiales species. For the Gesneriaceae family, the genome of *Dorcoceras hygrometricum* was sequenced and consists of 520,969 scaffolds, with a total span of 1,548 Mbp and an N50 value of 110,988 bp (Xiao et al., 2015). In the order Lamiales, genomes such as that of sesame (*Sesamum indicum*, Pedaliaceae; Wang et al., 2016) and spotted monkey flower (*Mimulus guttatus*; Hellsten et al., 2013. now *Erythranthe guttata*, Phrymaceae; Nesom, 2012) are available. Sesame has a high quality genome with 17 chromosomes and a total span of about 270 Mbp (Wang et al., 2016). The *E. guttata* genome, a recently established model organism, consists of 1,507 scaffolds, with a total span of 312.7 Mbp and an N50 value of 21.2 Mbp. However, due to the distant phylogenetic relationship of these species with *Streptocarpus*, their genomes may show little homology to that of *Streptocarpus*, and thus may not be a suitable reference sequence for this study.

3.1.3 Whole genome sequencing and its applications

A major application of a genome assembly is to serve as reference sequence for the analyses of other NGS data, including RNA-Seq and RAD-Seq (Ekblom and Wolf, 2014). Eventhough genome-scale screening can be conducted without a reference (Davey et al., 2011; Haas et al., 2013), analyses using a reference sequence can improve the results by recover genes or genotypes with lower sequencing coverage, and to correct sequencing errors (Lu et al., 2013; Haas et al., 2013; Florea and Salzberg, 2013). The combination of *de novo* assembly and reference-guided assembly of RNA-Seq data was shown to provide the best quality transcriptome in terms of both transcript length and number (Lu et al., 2013). For the analysis of RAD-Seq data, by mapping the reads to the reference genome, the chance of genotyping error is decreased as the required depth of coverage for correct genotyping is lower (Fountain et al., 2016). Reference-guided analysis can also increase the number of markers recovered and improve the resolution of the resulting genetic map (Shafer et al., 2016).

Reference genomes for the *Streptocarpus* species are thus invaluable for this study. It can improve the resolution and reduce the chances of error in RNA-Seq and RAD-Seq data analyses, and the assembly itself can provide sequence information at the targeted genetic regions that QTL mapping identifies. To obtain the genome assemblies of both *S. rexii* and *S. grandis* is also useful, since the comparative information may help identifying sequence differences of developmental regulating genes, untranslated introns or promoter regions, or for designing fine-mapping markers for further study once a genetic region of interest is identified. Therefore, the genomes of both *S. rexii* and *S. grandis* will be sequenced and analysed here.

3.1.4 Genome sequencing and assembly strategy

The output of different NGS technologies varies in read length, amount of data output, sequencing error rate, and cost (Goodwin et al., 2016; Van Dijk et al., 2018). Third generation sequencing technologies are suitable for resolving complex genomes with many repetitive elements, but they are more costly and have lower throughput per run thus reducing the depth of coverage of the target genome. Approaches such as mate-pair libraries (Illumina, San Diego, CA, USA), chromosome conformation capture (Dovetail Genomics, Santa Cruz, CA, USA) and optical mapping (Bionano Genomics, San Diego, CA, USA) are suitable for the production of chromosome-level scaffoldings based on preliminary genome assemblies (Jiao et al., 2017), and are thus not considered in this study. On the other hand, the Illumina sequencing-by-synthesis approach provides higher data output per unit cost of sequencing (Goodwin et al., 2016), allowing higher sequencing coverage for the medium-sized genome of *Streptocarpus* (Sims et al., 2014). By using the Patterned Flow Cell technology such as the HiSeq 4000 and HiSeq X, up to 650 to 900 Gbp of data can be generated in a sequencing experiment from a single flow cell with 8 lanes (Product specification, Illumina). By sequencing the *Streptocarpus* genome in a single lane, the 80 to 112 Gbp data generated can provide an approximately 100× sequencing depth for the ~1 Gbp genome of *Streptocarpus* (Goodwin et al., 2016).

Since no reference genome exists of a species closely related to *Streptocarpus*, *de novo* assembly is required to produce a draft genome from the sequencing data. *De novo* assembly describes the process of reconstructing contiguous sequences from shorter nucleotide fragments ('reads' from sequencing experiments) without the guidance of a reference (Ekblom and Wolf, 2014). Many difficulties can be encountered throughout this process, such as sequencing errors, low depth of coverage, repetitive elements or cross-species contamination. Hence, the aim during *de novo* assembly is to minimise these errors (Ekblom and Wolf, 2014; Laurence et al., 2014; Sims et al., 2014; Compeau et al., 2017). This can be achieved through a cautious choice of assembly tools, assembly parameters, and stringent quality control and filtering (e.g. Baker, 2012; Ekblom and Wolf, 2014; Dominguez Del Angel et al., 2018). Bioinformatics tools for genome assembly using NGS data have also been developed and are readily available. For example, SOAPdenovo (Luo et al., 2012) and ABySS (Simpson et al., 2009; Jackman et al., 2017) are commonly used for assembling medium to large-sized genomes without the requirement of enormous amounts of computing resources. SOAPdenovo was used for the *de novo* assembly of the *D. hygrometricum* and *Nicotiana tabacum* genomes (Sierro et al., 2014; Xiao et al., 2015), while ABySS has also been proven to be able to deal with highly complex plant genomes such as that of bread wheat (*Triticum aestivum*) (IWGSC, 2014).

Both SOAPdenovo and ABySS assemblers use the de Bruijn graph method (DBG) for *de novo* genome assembly (Li et al., 2012). In brief, the DBG approach starts by breaking down the sequencing reads into a series of "*k*-mers" (substrings of the original read with a length of *k*). A de Bruijn graph is then constructed from these *k*-mers, where each *k*-mer represents a node and the overlapping region between *k*-mers are indicated by arrows (called directed edges). By walking through the directed edges, clipping off stranded node tips, and resolving the graph structures such as bubbles and low coverage nodes, a long-contiguous "contig" is reconstructed (Figure 3.1; Li 2012; Luo et al., 2012; Jackman et al., 2017; Compeau et al., 2017). Usually, this is followed by using the read-pair information, possibly from the paired-end library or an additional long-insert library (Goodwin et al., 2016). The reads/contigs from the same read pair are joined together, and the gaps between the sequences are estimated from the insert size of the library, and replaced with "N" bases, thus forming the "scaffolds", meaning noncontiguous sequences with gaps of known length (Van Dijk et al., 2018).

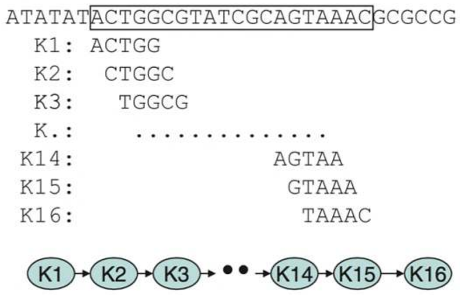


Figure 3.1 A simplified example of a DBG method assembly. In this example the target is to obtain the 20 bp-length genomic region on the top sequence. Sixteen *k*-mers can be obtained with the *k* value of 5 bp. The *k*-mers were used to construct the de Bruijn graph (bottom graph), where each node represents one *k*-mer, and the arrows (directed edges) indicate the direction of the assembly. Each pair of connected nodes differs only by one nucleotide at the beginning and at the end. By walking through the directed edges from the beginning to the end, the original sequence is reconstructed (modified from Li et al., 2012).

In practice, the *k* value of the *k*-mers is a major parameter to be tested and optimised for different assemblers (Li et al., 2012; Compeau et al., 2017; Mapleson et al., 2017). Larger *k* values retain more unique *k*-mers, which helps to resolve low-complexity short tandem repeats and reconstruct longer contigs; smaller *k* values may not span across the regions, but can recover assemblies with lower sequencing depth (Li et al., 2012; Compeau et al., 2017). In any case, the *k* value should be larger than half of the read length, and it should be an odd number to avoid sequence palindromes (Simpson et al., 2009; Reuter et al., 2015). The optimal *k* value can be determined by checking the shape of the *k*-mer histograms and by comparing assemblies reconstructed with iterative *k* values (Mapleson et al., 2017). Additionally, *k*-mer information is useful for the estimation of genome properties, i.e. genome size, heterozygosity, repeat content (Marçais and Kingsford, 2011; Mapleson et al., 2017; Vurture et al., 2017).

Another common problem associated with whole genome assembly is cross-species contamination. Biological samples are usually contaminated with traces of bacteria, algae, fungi, symbionts, and arthropods from the surrounding environment or introduced during nucleic acid sample preparation (Laurence et al., 2014; Laetsch and Blaxter, 2017). NGS technologies cannot distinguish between these and the target sequences and this can result in contaminant sequences being misassembled as part of the target genome of interest (Merchant et al., 2014; Laetsch and Blaxter, 2017). This error may lead to overestimating the genome size, overestimating gene copy number, or erroneous biological conclusions such as horizontal-gene-transfers (Koutsovoulos et al., 2016). Hence, a crucial step for genome analysis is to identify and remove these non-target contaminants (Kumar et al., 2013; Laurence et al., 2014; Ekblom and Wolf, 2014). Assessing the GC content of the raw reads

(Andrews, 2010) or the assembled contigs is a common way to check for contaminants that have different GC ratios to the target species (Ekblom and Wolf, 2014). Removing contigs and scaffolds smaller than a given size threshold also greatly helps removing potential contaminants or misassemblies that tend to be small in size (Koutsovoulos et al., 2016). Software packages like Blobtools (Kumar et al., 2013; Laetsch and Blaxter, 2017) can be used to visualise the GC content of each scaffold against its sequence coverage. Together with BLAST-assigned taxonomical information, the taxon-annotated GC-coverage plot (blobplot) is a useful approach to remove cross-species contamination from the assemblies.

The quality of the assemblies should be compared quantitatively. Tools such as Quast can be used to assess the assembly metrics including total assembly size, N50, L50, GC content, and percentage of N bases (Gurevich et al., 2013). The N50 and L50 values represent a measurement of the contiguity and length of the assembled contig; N50 is defined as a length L so that the summation of all contigs with length $\geq L$ is at least half the length of the total assembly; L50 is defined as the number of contigs required to meet above criteria (Gurevich et al., 2013; Ekblom and Wolf, 2014). The GC content, as described above, can be a measurement of potential contaminant species. The percentage of N bases represent the proportion of gaps in the assembly (Ekblom and Wolf, 2014).

Another frequently used assembly quality metric is the completeness of house-keeping genes that are highly conserved across most of the organisms (Parra et al., 2007; Simão et al., 2015). Software packages such as BUSCO searches for the Benchmarking Universal Single Copy Orthologs in a genome assembly, and the percentage of BUSCO completeness provides a rough estimation of the completeness of the assembly (Simão et al., 2015; Koutsovoulos et al., 2016). There is no standard value for the BUSCO completeness of a draft genome, although a BUSCO completeness of ~90% can generally be considered as a good assembly (M Blaxter, personal communication). Another way to compare two genome assemblies is through genome-to-genome alignment, where the similarity between two sequences can be compared and the alignment can be visualised as a dot plot to identify local sequence rearrangement (Delcher et al., 2002; Marçais et al., 2018; Cabanettes and Klopp, 2018).

In addition to the nuclear genome assembly, whole genome sequencing data are usually accompanied by a high proportion of reads derived from the organellar genomes, i.e. plastid and mitochondria (McPherson et al., 2013). These genomes are much smaller in size and are typically circular. The plastid genomes are known to range from about 120 to 160 Kbp (Wicke et al., 2011), and the mitochondrial genomes vary much more in size and typically range between 200 to 750 Kbp (Gualberto et al., 2014; Dierckxsens et al., 2017). The organellar genome reads can be assembled, and the produced genomes could provide potential resources for barcoding, phylogeny or population genetics research (Jansen et al., 2006; Shaw et al., 2007; Zhao et al., 2018). So far in the genus *Streptocarpus*, only the plastid genome of *S. teitensis* has been assembled and annotated (Kyalo et al., 2018). In this study, the organellar genomes of the *S. rexii* and *S. grandis* will also be assembled and characterised.

In summary, whole genome data analysis requires cautious optimisation and quality control during assembly. In this chapter I attempt to assemble the first whole genome sequence for *S. rexii* and *S. grandis* in the genus *Streptocarpus*. The resulting nuclear and organellar genomes will serve as reference sequences for later chapters, as well as provide invaluable resources for future studies.

3.2 Materials and methods

3.2.1 Plant materials

The *Streptocarpus rexii* (accession 20150819*A) and *Streptocarpus grandis* (accession 20150821*A) were used as the materials for whole genome shotgun sequencing (Detail information of the materials are in Appendix 3.1). Both accessions are direct descendants of inbred lineages (*S. rexii*: selfed F2, *S. grandis*: selfed F3) that are later used for genetic studies in Chapter 4 and Chapter 5. Plants of both species were grown in the research glasshouse of the Royal Botanic Garden Edinburgh.

3.2.2 DNA extraction, library preparation and genome sequencing

Approaches to the DNA extraction and quality assessment of *Streptocarpus* are described in Chapter 2 (see also Appendix 2.6). In brief, the plant DNA was extracted using ChargeSwitch gDNA Plant Kit (Thermo Fisher Scientific), followed by RNase A treatment and phenol:chloroform:isoamyl alcohol (25:24:1) purification before finally eluted in TE buffer. The extracted DNA samples were submitted to Edinburgh Genomics (University of Edinburgh, UK) for library preparation and whole genome shotgun sequencing. Short-insert paired-end libraries were prepared using the TruSeq DNA PCR free Library Prep Kit (Illumina, San Diego, CA, USA). For *S. rexii*, two libraries with insert sizes of 350 bp and 550 bp were prepared. For *S. grandis*, one library with an insert size of 350 bp was prepared. Paired-end reads of length 150 bp were generated from the libraries. The *S. rexii* library with insert size 550 bp was sequenced on HiSeq 4000, and both libraries with insert sizes of 350 bp were sequenced on the HiSeq X (Illumina). All libraries were sequenced in individual lanes to ensure maximum coverage. The read data was returned in fastq format and the software FastQC v.0.11.7 (Andrews, 2010) was used to evaluate the quality of the reads and the results were summarised using MultiQC v.1.5 (Ewels et al., 2016).

3.2.3 Assembly and analysis of the organellar genome

The plastid and mitochondrial genomes were assembled using the software NOVOPlasty v.2.6.5 (Dierckxsens et al., 2017). In brief, this software takes a seed sequence and carries out seed-and-extend algorithm for assembly, and can incorporate a reference sequence to resolve the structure of repetitive regions. For both *Streptocarpus* species, the HiSeq X data were used as the input. The input file for the assembler was first prepared according to Dierckxsens et al. (2017), which a subset of 20 million unprocessed read pairs was prepared by extracting the first 20 million reads from both forward and reverse read files of the paired-end data, as described in Box 3.1.

Box 3.1 Extracting 20 million read pairs from the whole genome sequencing data

```
head -n 80000000 <(gunzip -c [READ1.fq.gz]) > read1_20M.fq
head -n 80000000 <(gunzip -c [READ2.fq.gz]) > read2_20M.fq
```

Note: As each read entry has four lines of information in fastq format, a total number of 20 × 4 = 80 million lines were specified.

The *S. rexii* and *S. grandis* plastid genomes were assembled with the *D. hygrometricum* plastid sequence (153,493 bp, GenBank accession: JN107811; Zhang et al., 2012) as the seed and reference in NOVOPlasty. The "Genome range" parameter value was set to 120,000 to 200,000 bp, so that a genome larger than that of *D. hygrometricum* was considered by the software. The rest of the parameters remained unchanged as default. The detailed configuration file for the assembler is shown in Box 3.2.

The assembly of the mitochondrial genome was carried out in a step-wise approach. For the *S. rexii* mitochondrion, the reads were first mapped to the *D. hygrometricum* mitochondrial genome (510,519 bp, GenBank accession: JN107812; Zhang et al., 2012) to identify a mitochondrion-specific read that was later used as the seed sequence to initiate the assembly. The mapped reads were BLAST searched against the nucleotide (nt) database on the NCBI webpage (Altschul et al., 1990) and the BLAST report checked to ensure that the read only matched plant mitochondrial but not plastid sequences, for confirming that the origin of the read is a mitochondrion-specific region. Finally, the following read sequence was chosen as the seed for assembling the *S. rexii* mitochondrial genome:

CATAAGGGCCATGCGGACTTGACGTCATCCCCACCTTCCTCCAGTATATC
ACTGACAGTCCTTCGTGAGTGCGGCACGCACCTTTTTCTTTCTTTTGGAGCTGTTT
TGTCGGGGCGTACTAAACCCACTACGTACCACACCACCGGGCAG

The *D. hygrometricum* mitochondrial sequence was used as reference, and the "Genome range" parameter value was set to 150,000 to 700,000 bp, so that a genome size larger than the *D. hygrometricum* mitochondrion was considered. The assembled *S. rexii* plastid genome was provided to the NOVOPlasty software for filtering out plastid reads from the raw data (parameter "Chloroplast sequence"), so the software can assemble the mitochondrial genome without incorporating the plastid sequences. The assembly began with the default *k*-mer size of 39, and gradually increased by 10 bp intervals if the mitochondrial assembly was not circularised. Finally, the circularised *S. rexii* mitochondrial assembly was created with a *k*-mer size of 79. The detailed configuration file for the assembler is shown in Box 3.3.

The same procedure was repeated to obtain the *S. grandis* mitochondrial genome. The following sequence was chosen as the seed to initiate the assembly:

GGCCATGCGGACTTGACGTCATCCCCACCTTCCTCCAGTATATCACTGAC
AGTCCTTCGTGAGTGCGGCACGCACCTTTTTCTTTCTTTTGGAGCTGTTTTGTCGG
GGCGTACTAAACCCACTACGTACCACACCACCGGGCAGATCGCC

The other parameters were the same as those used for the *S. rexii* mitochondrial assembly, with the *S. grandis* plastid genome provided for the "Chloroplast sequence" parameter to filter out plastid reads. Finally, the circularised genome was assembled with a *k*-mer value of 59 (Box 3.3).

Box 3.2 NOVOPlasty configuration file and commands for the plastid genome assembly of *Streptocarpus* species.

```
# For the config file
Project:
-----------------------
Project name          = Plastid_assembly
Type                  = chloro
Genome Range          = 120000-200000
K-mer                 = 39
Max memory            =
Extended log          = 0
Save assembled reads  = no
Seed Input            = [INPUT_SEED.fa]
Reference sequence    = [INPUT_REFERENCE.fa]
Variance detection    = no
Heteroplasmy          =
Chloroplast sequence  =

Dataset 1:
-----------------------
Read Length           = 150
Insert size           = 350
Platform              = illumina
Single/Paired         = PE
Combined reads        =
Forward reads         = [READ1_20M.fastq]
Reverse reads         = [READ2_20M.fastq]

Optional:
-----------------------
Insert size auto      = yes
Insert Range          = 1.8
Insert Range strict   = 1.3

# To execute the software for the assembly
perl novoplasty.pl –c [CONFIG_FILE]
```

Box 3.3 NOVOPlasty parameter settings for the mitochondrial genome assembly. The main differences to the plastid assembly are (1) Type, (2) Genome range, (3) Chloroplast sequence, as the plastid sequence should be provided for the mitochondrial assembly.

```
# For the config file
Project:
-----------------------
Project name          = Mitochondrial_assembly
Type                  = mito_plant
Genome Range          = 150000-700000
K-mer                 = [K-MER_SIZE]
Max memory            =
Extended log          = 0
Save assembled reads  = no
Seed Input            = [INPUT_SEED.fa]
Reference sequence    = [INPUT_REFERENCE.fa]
Variance detection    = no
Heteroplasmy          =
Chloroplast sequence  = [ASSEMBLED_PLASTID_GENOME.fa]

Dataset 1:
-----------------------
Read Length           = 150
Insert size           = 350
Platform              = illumina
Single/Paired         = PE
Combined reads        =
Forward reads         = [READ1_20M.fastq]
Reverse reads         = [READ2_20M.fastq]

Optional:
-----------------------
Insert size auto      = yes
Insert Range          = 1.8
Insert Range strict   = 1.3

# To execute the software for the assembly
perl novoplasty.pl –c [CONFIG_FILE]
```

The annotation of the organellar genome assemblies was carried out with the webtool GeSeq v.1.50 (Tillich et al., 2017). This tool provides visualisation of the genome annotation using OGDRAW v.1.2 (Lohse et al., 2007; 2013) and generates an annotated GenBank annotation file. The annotation of the plastid genome was carried out under default parameters plus enabling the plastid-specific functions, i.e. HMMER profile search (Wheeler and Eddy, 2013) and the coding gene sequence and rRNA BLAT search (Kent, 2002) using the MPI-MP plastid references (Tillich et al. 2017). Plastid genomes of model plant species were also included as references for the BLAT search for gene prediction (Kent, 2002), which the *A. thaliana* (GenBank accession: NC_000932; Sato et al., 1999), *Nicotiana tabacum* L. (accession: NC_001879; Shinozaki et al., 1986), and *Solanum lycopersicum* L.

(accession: AC_000188; Kahlau et al., 2006) were chosen. tRNAscan-SE v.2.0 was enabled for tRNA annotation (Lowe, 1997; Lowe and Chan, 2016).

The annotation of the mitochondrial genome was also carried out under default settings. tRNAscan-SE v.2.0 was used for tRNA search, and the mitochondrial genomes of model species were included for the BLAT searches for gene prediction, including *A. thaliana* (accession: NC_001284; Unseld et al., 1997), *N. tabacum* (accession: NC_006581; Sugiyama et al., 2004), and *S. lycopersicum* (accession: NC_035963; Mueller et al., 2005).

The assembled plastid / mitochondrial sequences were aligned to compare the genome structures between *S. rexii* and *S. grandis*. The alignment was done using the "progressiveMauve alignment" function in the program Mauve v.2.4.0 under default settings (Darling et al., 2004; 2010). The aligned sequences were further checked for identity and similarity by the Sequence Identity And Similarity webtool (last update 20 September 2017, http://imed.med.ucm.es/Tools/sias.html), and uncorrected distance between sequences by the EMBOSS DISTMAT v.6.6.0.0 (http://www.hpa-bioinfotools.org.uk/pise/distmat.html).

3.2.4 *De novo* assembly of the nuclear genome

A summarised flowchart of the whole process of genome assembly and filtering is in Appendix 3.2, and each step is described in detail below.

Preliminary S. rexii *genome assembly*

To assess the assembler performance and the overall genome properties, a preliminary genome assembly was performed using the *S. rexii* HiSeq 4000 reads. Prior to the assembling, the reads were quality checked and adapter trimmed (removed) using Trimmomatic v.0.36 (ILLUMINACLIP: TruSeq3-SE.fa:2:30:10 LEADING:20 TRAILING:20 SLIDINGWINDOW:30:20 AVGQUAL:20 MINLEN:51; Bolger et al., 2014). The trimmed reads were then assembled using SOAPdenovo2 with a *k* value of 55 (Luo et al., 2012).

A second preliminary assembly attempt was taken to compare the assembly performance between SOAPdenovo2 and ABySS2 (Simpson et al., 2009; Jackman et al., 2017). The Hiseq 4000 and Hiseq X reads of *S. rexii* were used but without trimming, as the assemblers can automatically exclude lowly supported (likely error) *k*-mers (Simpson et al., 2009; Jackman et al., 2017). The assemblers ABySS2 v.2.0.2 (Simpson et al., 2009; Jackman et al., 2017) and SOAPdenovo2 (Luo et al., 2012) were compared for their product assemblies' metrics. A *k* value of *k* = 77 was chosen as the value represents about the half value of the read length (150 bp). The rest of the assembly parameters remained unchanged as default.

The quality metrics of all the assemblies were assessed using QUAST v.4.6.3 (Gurevich et al., 2013). The remapping rate was calculated by mapping the reads onto the genome using the BWA v.0.7.17 "mem" function (Li and Durbin, 2009). The average depth of coverage was calculated based on the remapped BAM file via the "genomecov" function in BEDtools (Quinlan and Hall, 2010).

K-*mer size optimisation*

The optimal *k*-mer value was estimated using the software Kmer Analysis Toolkit KAT v.2.3.4 (Mapleson et al., 2017). This software was used generate *k*-mer distribution histograms from the raw reads, using *k* values between 77 bp and 147 bp with a 10 bp interval. The generated histograms were compared to check the number of distinct *k*-mers and the frequency of *k*-mer occurrence. In addition, the *k*-mer counting result was used to estimate the genome properties i.e. genome size, heterozygosity and repeat content, using the webtool GenomeScope (Vurture et al., 2017).

Genome assembly of S. rexii *and* S. grandis

ABySS2 v.2.0.2 (Simpson et al., 2009; Jackman et al., 2017) was used for the final assembling of *S. rexii* and *S. grandis* genomes. For *S. rexii*, both the HiSeq 4000 and HiSeq X data were used and the assembly was carried out with *k* values between *k* = 77 and *k* = 147, with *k* = 10 intervals. For *S. grandis*, the HiSeq X data was used and the assembly was carried out with *k* values between *k* = 107 and *k* = 137, with *k* = 10 intervals. The Bloom filter was enabled to reduce the maximum memory usage (parameter B, H, and kc). The commands used for the assembly are given in Box 3.4.

Box 3.4 Parameter setting for the genome assembly of *Streptocarpus* species using ABySS2 and SOAPdenovo2

```
## Genome assembly using ABySS2 with two paired-end libraries
# The Bloom filter was enabled to reduce the memory usage
abyss-pe -j j=[NO._THREADS] k=[K_VALUE] name=[OUTPUT_NAME] \
    lib='[LIBRARY_A] [LIBRARY_B]' \
    LIBRARY_A='[LIB_A_READ1] [LIB_A_READ2]' \
    LIBRARY_B='[LIB_B_READ1] [LIB_B_READ2]' \
    B=20000M H=3 kc=3

## Genome assembly using SOAPdenovo2
# The configuration file for SOAPdenovo2

max_rd_len=150
[LIB]
reverse_seq=0
asm_flags=3
rd_len_cutoff=100
rank=1
pair_num_cutoff=3
map_len=32
q1=[LIB_A_READ1]
q2=[LIB_A_READ2]

[LIB]
reverse_seq=0
asm_flags=3
rd_len_cutoff=100
rank=1
pair_num_cutoff=3
map_len=32
q1=[LIB_B_READ1]
q2=[LIB_B_READ2]

# Assemble the genome using SOAPdenovo2
SOAPdenovo-63mer all -s [CONFIG_FILE] -o [OUTPUT_NAME] \
    -K [K_VALUE] -p [NO._THREADS]

## Calculating the remapping rate using BWA and samtools
# Index the genome assembly
bwa index [GENOME_ASSEMBLY]
# Remap the raw reads to the assembly and write the output in BAM format
bwa mem -t [NO._THREADS] [GENOME_ASSEMBLY] [READ1] [READ2] \
    | samtools view -Sb \
    | samtools sort -O bam -o [OUTPUT_NAME.bam]
# Calculate the mapping rate
Samtools flagstat [OUTPUT_NAME.bam]

## Calculating the average depth of coverage of the genome
# Computes the depth of coverage of the genome
bedtools genomecov [OUTPUT_NAME.bam] > [OUTPUT_FILE]
# Calculate the average depth of coverage
awk '{ total += $2 } END { print total/NR }' [OUTPUT_FILE]
```

3.2.5 Post-assembly filtering and assessment of assembly quality

For all the produced assemblies, scaffolds smaller than 500 bp were removed. QUAST v.4.6.3 (Gurevich et al., 2013) was used to calculate the assembly quality metrics. The remapping rate was calculated by mapping the reads onto the genome using the BWA v.0.7.17 “mem” function (Li and Durbin, 2009). The average depth of coverage was calculated based on the remapped BAM file via the “genomecov” function in BEDtools (Quinlan and Hall, 2010).

The software Blobtools v.1.0 (Kumar et al., 2013; Koutsovoulos et al., 2016; Laetsch and Blaxter, 2017) was used to remove potential contaminant scaffolds from the assemblies. For the preparation of the Blobtools input file, (1) the coverage file (.cov file) was prepared by aligning raw reads to the assembly using BWA “mem” (Li and Durbin, 2009) with default settings. SAMtools v.1.7 (Li et al., 2009) was then used to process the SAM format to BAM format. The resultant bam file was converted to coverage file (.cov file) using the Blobtools *map2cov* function; (2) the hits file containing the taxonomic information on the scaffolds was prepared by performing local BLAST searches of the draft genome against the NCBI nt database (downloaded on 26-Mar-2018). BLASTn megablast was performed using BLAST+ v2.7.1+ (Camacho et al., 2009) with an E-value threshold of 1e-25; (3) the assembly file was the genome assembly generated from the ABySS assembler in fasta format.

With the three input files, Blobtools was used to construct the blobplot for both *S. rexii* and *S. grandis* assemblies. The detailed commands are provided in Box 3.5.

Box 3.5 Commands for Blobtools for blobplot generation

```
# Preparing the .cov file
bwa index -p [INDEX_PREFIX] [GENOME_ASSEMBLY.fasta]
bwa mem [INDEX_PREFIX] [READ1.fastq.gz] [READ2.fastq.gz] \
    | samtools view –Sb \
    | samtools sort –O bam –o [FINAL_BAM.bam]
blobtools map2cov –b [FINAL_BAM.bam] –o [FINAL_COV_FILE]

# Preparing the hits file
blastn -task megablast –db nt –evalue 1e-25 –culling_limit 5 \
    -query [GENOME_ASSEMBLY.fa] \
    -outfmt '6 qseqid staxids bitscore std \
             sscinames sskingdoms stitle' \
    -out [OUTPUT_HIT_FILE]

# Run blobtools create to create blobDB data structure
blobtools create –x bestsum \
    -i [GENOME_ASSEMBLY.fasta] -c [FINAL_COV_FILE] \
    -t [OUTPUT_FIT_FILE] \
    -o [BLOBDB_OUTPUT_PREFIX]

# Run blobtools view on the output blobDB.json
# to create the summary table
```

```
blobtools view –i [BLOBDB]

# Run blobtools blobplot on the output blobDB.jason
# to create the blobplot
blobtools blobplot –i [BLBODB] –o [OUTPUT_GRAPH_PREFIX]
```

After identifying the scaffolds of potential contaminant origin (which were labelled as non-plant origin by the Blobtools analysis), they were removed, and only the scaffolds labelled as ‘Streptophyta’ or ‘no-hit’ (unidentifiable in the searched database) were kept. The filtered assemblies were analysed with Blobtools again to ensure the complete removal of the contaminants. The detailed commands including the filtering of the assembly are shown in Box 3.6.

Box 3.6 Commands for filtering genome assemblies based on blobplot results

```
## Filtering the genome assembly
# Retrieve the fasta entries that were labelled as Streptophyta
# or no-hit in the summary table
awk '$6=="Streptophyta" || $6=="no-hit"' [BLOB_TABLE] | \
    awk {print $1} > Sequence.to.keep

# Retrieve the actual fasta sequence from the unfiltered
# genome assembly
perl -ne 'if(/^>(\S+)/){$c=$i{$1}}$c?print:chomp;$i{$_}=1 if @ARGV'
    Sequence.to.keep [GENOME_ASSEMBLY.fasta] > \
    [FILTERED_GENOME.fasta]

# Retrieve the Blast result from the original hit file, keeping
# only the filtered assemblies
for i in $(cat Sequence.to.keep);do
    awk -v num="$i" '$1==num' [OUTPUT_HIT_FILE] >> [FILTERED_HIT]
done

# Retrieve the coverage information from the original cov file,
# keeping only the filtered assemblies
for i in $(cat Sequence.to.keep);do
    awk -v num="$i" '$1==num' [FINAL_COV_FILE] >> [FILTERED_COV]
done

## Generate the blobplot for filtered assembly
blobtools create –x bestsum \
    -i [FILTERED_GENOME.fasta] -c [FILTERED_COV] -t [FILTERED_HIT]\
    -o [FILTERED_BLOBDB_PREFIX]
blobtools view –i [FILTERED_BLOBDB]
blobtools blobplot –i [FILTERED_BLBODB] \
    –o [FILTERED_OUTPUT_PREFIX]
```

The completeness of the genome assemblies, in terms of the presence of core genes (essential for biological functions), was assessed using BUSCO v.3 (Simão et al., 2015). The core genes for the plant dataset were downloaded from the BUSCO webpage

(Embryophyta_odb9), which a total number of 1,440 core genes were used as reference for the search (Box 3.7).

Box 3.7 Commands for BUSCO analysis for completeness of core genes of the assemblies

```
python BUSCO.py -i [GENOME.fasta] -o [OUTPUT_NAME] -m geno \
    -cpu [NO._THREADS] -l embryophyta_odb9/
```

For genome-to-genome alignment, the webtool D-GENIES v.1.1.1 was used (Cabanettes and Klopp, 2018) which carries out alignments using Minimap v.2 (Li, 2018), and the results were visualised as dot plots. The alignment was carried out between (1) *S. rexii* and *S. grandis*, (2) *S. rexii* and *D. hygrometricum*, and (3) *S. grandis* and *D. hygrometricum* (GCA_001598015; Xiao et al., 2015), all under default parameters.

3.3 Results

3.3.1 Quality check of the whole genome shotgun sequencing reads

The number of reads obtained from the sequencing experiments are summarised in Table 3.1. For *S. rexii*, about 401.8 million read pairs (c. 803 million reads) and 260.4 million read pairs (c. 520 million reads) were obtained from HiSeq X and HiSeq 4000 sequencing, respectively. For *S. grandis*, almost 452.7 million read pairs (c. 905 million reads) were obtained from the HiSeq X sequencing.

The obtained reads had an average Phred quality score above Q30 except for the region at the end of read 2 (Figure 3.2). On the other hand, the GC distribution of the reads indicated that both *S. rexii* datasets had a GC content distribution which deviates from expected. The major peak was at around 39% GC, but a smaller fraction of the reads had a higher GC content of around 67%-71% (Figure 3.2 a and b). The GC content of *S. grandis* appeared as normal distribution, with a central peak at around 39% (Figure 3.2 c).

Table 3.1 Amount of read counts generated from the whole genome shotgun sequencing

	S. rexii		***S. grandis***
	HiSeq X	**HiSeq 4000**	**HiSeq X**
Library insert size (bp)	350	550	350
Read length (bp)	150	150	150
Read pairs obtained	401,838,795	260,476,261	452,684,043

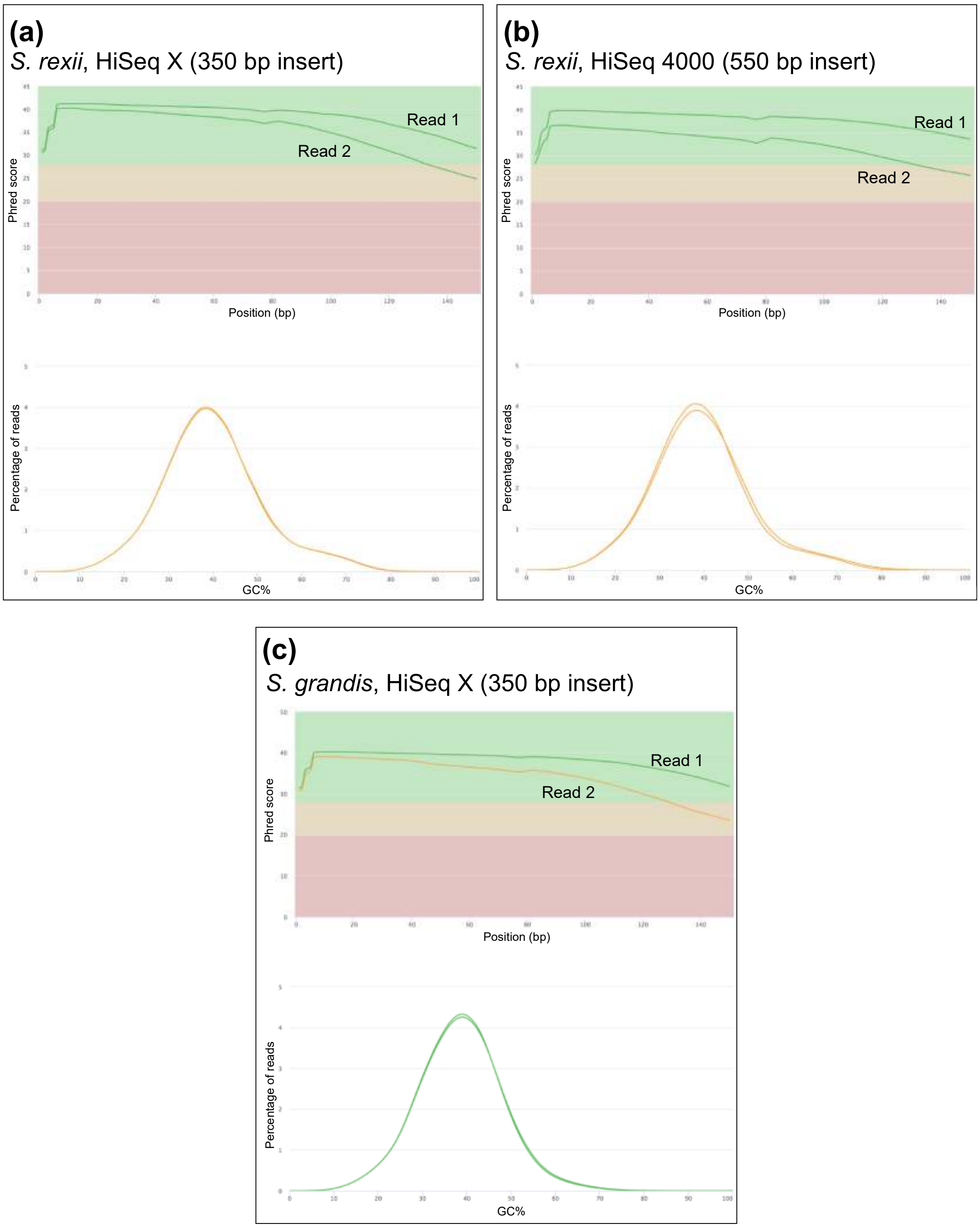


Figure 3.2 FastQC quality check of the reads generated from the three whole genome shotgun sequencing experiments. **(a)** *S. rexii* HiSeq X. **(b)** *S. rexii* HiSeq 4000. **(c)** *S. grandis* HiSeq X. The upper graph shows the mean quality score (Phred score), and the lower graph shows the distribution of the per read GC% content. The two lines represent the forward and reverse reads from the paired-end sequencing. The colours of the lines indicate pass (green) or warning (orange) of the quality check results.

3.3.2 Organellar genome assembly

The complete circular plastid genomes were assembled for *S. rexii* and *S. grandis*. For *S. rexii,* the assembled sequence had a total length of 152,724 bp, with 107 protein coding genes, 8 rRNA and 33 tRNA annotated sequences (Table 3.2 and Figure 3.3 a). The *S. grandis* plastid genome had similar metrics, and the assembly spanned 152,770 bp, and the annotation is identical to that of the *S. rexii* assembly (Table 3.2 and Figure 3.3 b). The two assembled sequences can be aligned where a large synteny block was identified that covers the entire plastid genome (Figure 3.4). In total, 267 SNPs and 63 gaps were found in the alignment. The two sequences shared 99.65% identity and similarity, and the uncorrected distance was 0.18.

Table 3.2 Metrics of the *S. rexii* and *S. grandis* chloroplast genome assemblies based on HiSeq X data.

	S. rexii	***S. grandis***
No. contigs	1	1
Total base pairs (bp)	152,724	152,770
No. protein coding genes	107	107
No. rRNA annotated	8	8
No. tRNA annotated	33	33

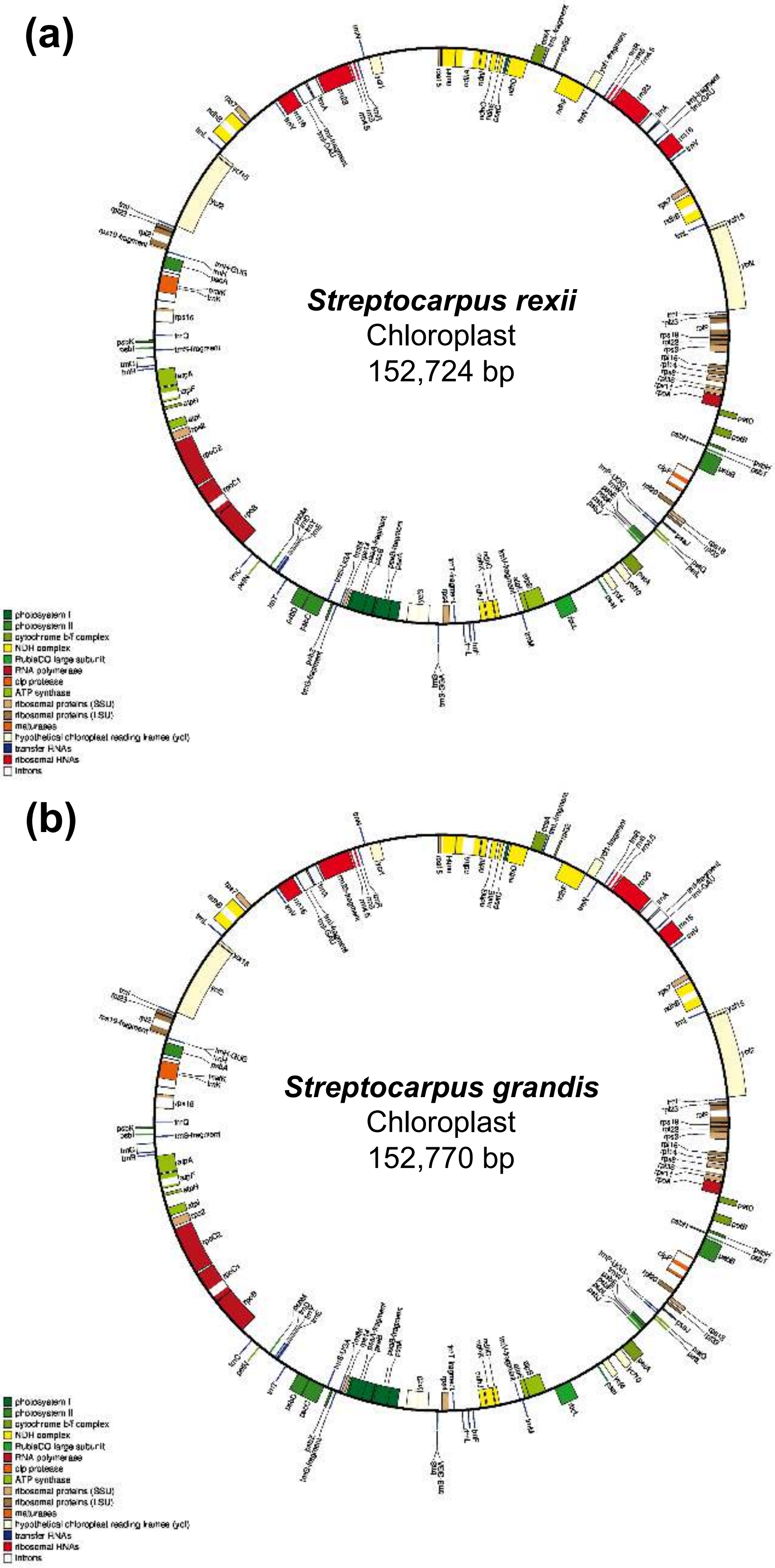


Figure 3.3 Gene map of the *Streptocarpus* plastid genome assemblies, with gene annotations in colour. (a) *S. rexii*. (b) *S. grandis*.

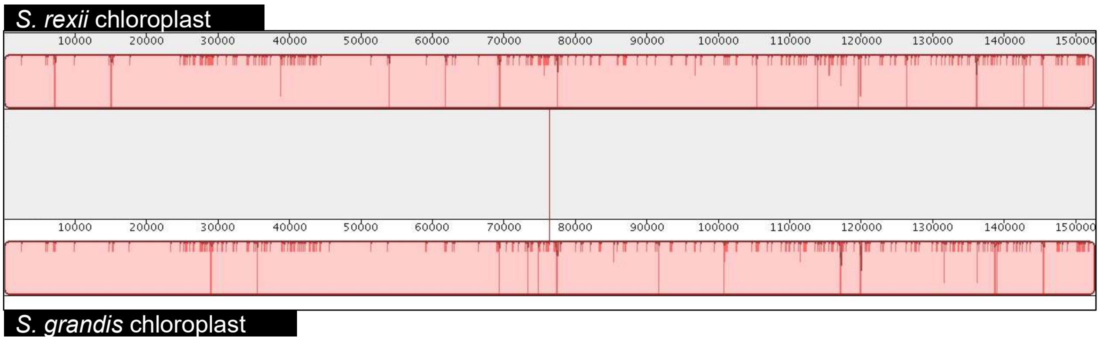


Figure 3.4 Synteny between the *S. rexii* and *S. grandis* plastid genome assemblies. The large red rectangles represent the synteny blocks identified by Mauve. The vertical lines inside the synteny blocks indicate the similarities between the two aligned sequences, where longer and darker lines suggest lower sequence similarity or gaps in the alignment. The numbers above the blocks indicate base pairs.

The complete mitochondrial genomes of *S. rexii* and *S. grandis* were also assembled. The *S. rexii* mitochondrial genome spanned 314,134 bp, with 72 protein coding genes, 2 rRNA and 17 tRNA annotated (Table 3.3 and Figure 3.5 a). The *S. grandis* mitochondria spanned 352,540 bp, with 79 protein coding genes, 3 rRNA and 17 tRNA annotated (Table 3.3 and Figure 3.5 b). Strangely, the mitochondrial complex III was only identified in *S. grandis* (Figure 3.5 b; red arrow) assembly but not in *S. rexii*. The synteny analysis suggested that the orientation of the identified synteny blocks were very different between the two assemblies, which 13 synteny blocks identified and they were different in terms of strand, position, and order (Figure 3.6). Among the aligned region, 1,080 SNPs and 289 gaps were found. Due to the fact that the sequence cannot really be aligned and the conserved region is too fragmented, the sequence identity, similarity and the uncorrected distance were not calculated.

Table 3.3 Metrics of the *S. rexii* and *S. grandis* mitochondrial genome assemblies based on HiSeq X data.

	S. rexii	***S. grandis***
No. contigs	1	1
Total base pairs (bp)	314,134	352,540
No. protein coding genes	72	79
No. rRNA annotated	2	3
No. tRNA annotated	17	17

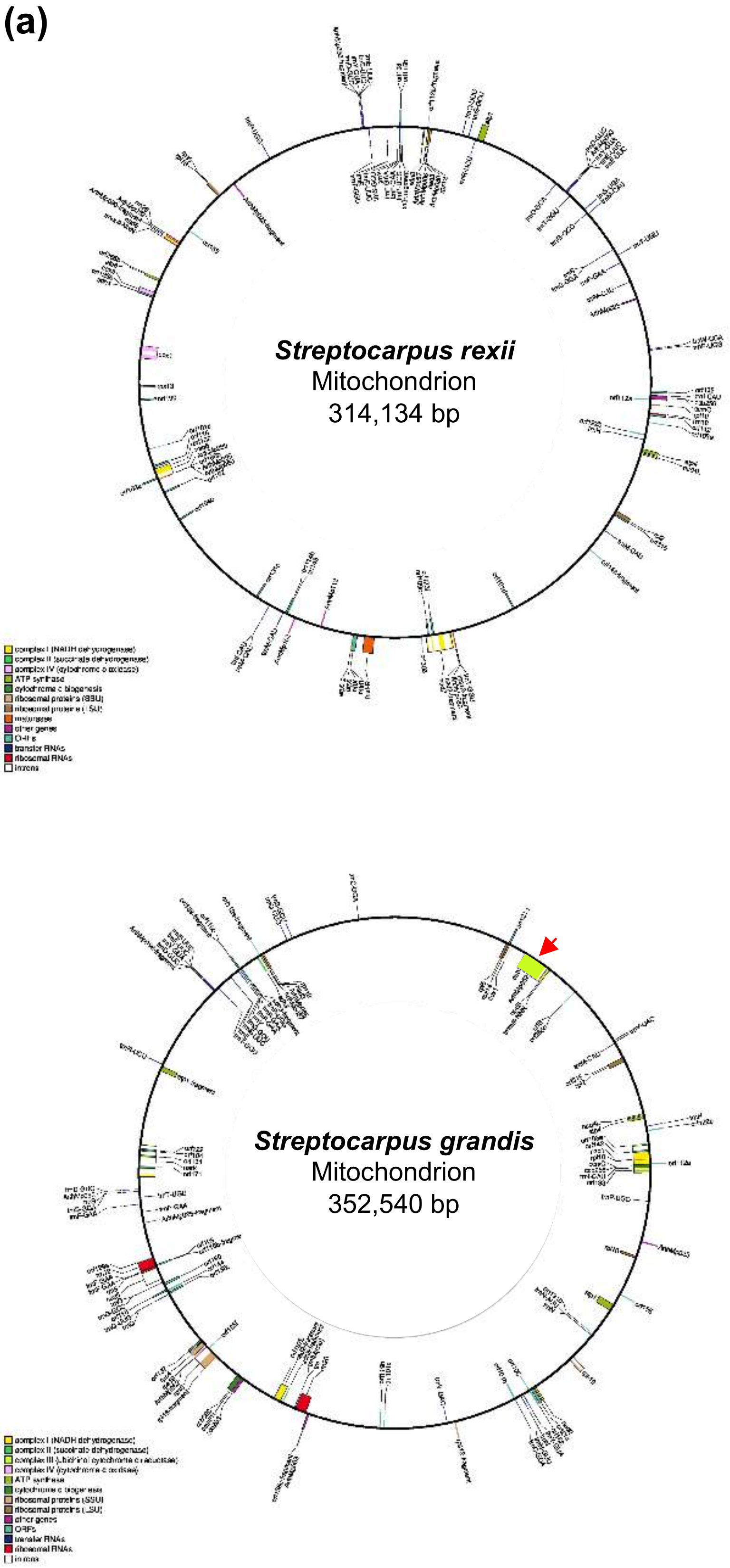


Figure 3.5 Gene map of the *Streptocarpus* mitochondrial genome assemblies, with gene annotations in colour. (a) *S. rexii*. (b) *S. grandis*. Red arrow: mitochondrial complex III.

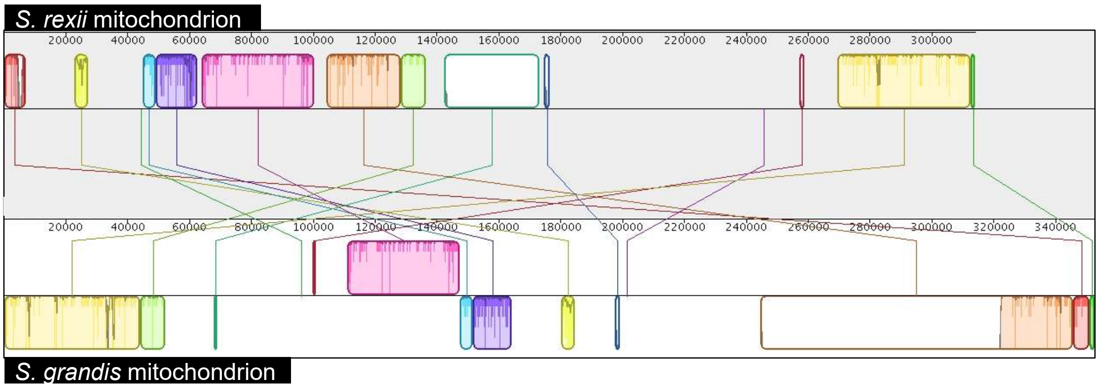


Figure 3.6 Synteny between the *S. rexii* and *S. grandis* mitochondrial genome assemblies. The rectangles of different colours represent the synteny blocks identified by Mauve. The vertical lines inside the synteny blocks indicate the similarities between the two aligned sequences, where longer and darker lines suggest lower sequence similarity or gaps in the alignment). The numbers above the blocks indicate base pairs.

3.3.3 Preliminary assembly tests for the plant nuclear genome

Among the initial 260,476,261 read-pairs, 257,355,603 were kept after quality and adapter trimming (98.8%). The trimmed reads showed an overall good quality with an average Phred quality score above Q30, but the abnormal GC content distribution remained the same (Figure 3.7).

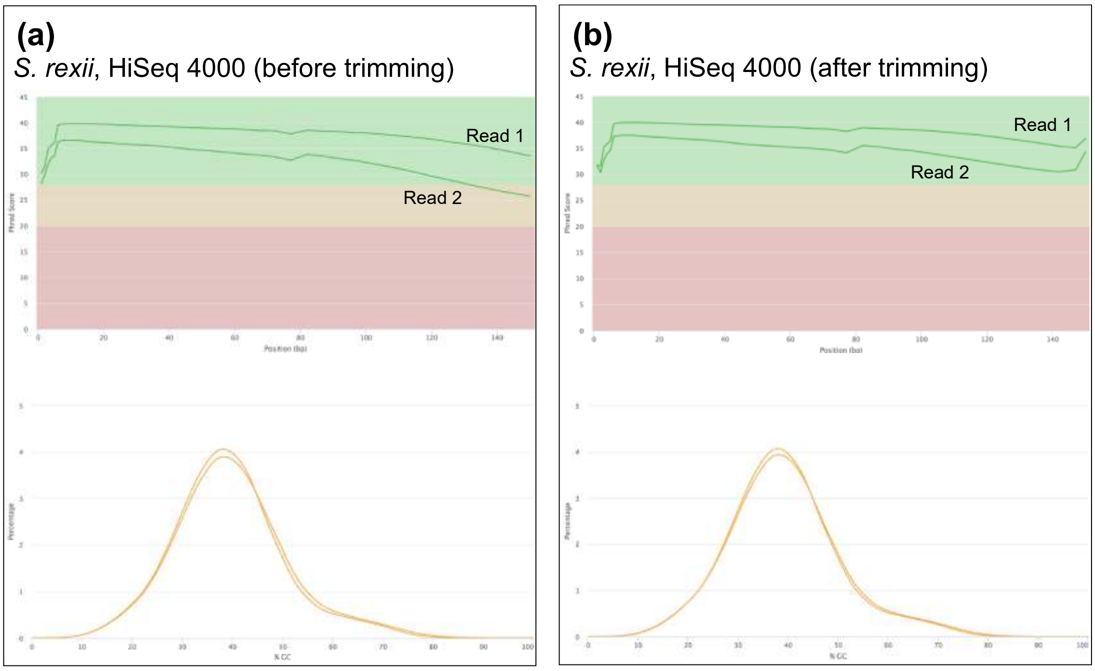


Figure 3.7 FastQC quality check of the *S. rexii* HiSeq 4000 reads. **(a)** Before quality check and adapter trimming. **(b)** After quality check and adapter trimming. The upper graph shows the mean quality score, and the lower graph shows the per sequence GC% content. The two

lines represent the forward and reverse reads from the paired-end sequencing. The colours of the lines indicate pass (green) or warning (orange) of the quality check result.

Assembly of the preprocessed reads resulted in 1,380,875 scaffolds (Table 3.4). The total span was 902,891,804 bp, and the longest scaffold was about 1.6 Mbp. After filtering out the small scaffolds shorter than 500 bp, 97,377 scaffolds were kept. The filtered assembly had a total span of 716,373,945 bp, an N50 value of 25,903 bp, and a L50 value of 5,871. About 64.7% of the input reads (166,613,685 out of 257,355,603 read pairs) were mapped to the draft genome assembly, and the mean depth of coverage of the assembled genome was 177×. However, the assembly contained a high proportion of Ns, i.e. 206,359,657 N bases which comprised about 29% of the assembly (Table 3.4,). In addition, the GC% distribution of the assembly indicated the presence of a group of scaffolds with an abnormally high GC content at 60% to 65% (Figure 3.8).

Table 3.4 Metrics of the *S. rexii* SOAPdenovo2 preliminary assembly on preprocessed reads

***S. rexii* SOAPdenovo2 preliminary assembly**	
Dataset used	HiSeq 4000 (preprocessed)
No. input read pairs	257,355,603
Assembler	SOAPdenovo2 v2.04-r240
Assembly parameter	$k = 55$
Assembly metrics	
Total no. scaffolds	1,380,875
Total span (bp)	902,891,804
No. scaffolds (≥ 500 bp)	97,377
Total span (≥ 500 bp) (bp)	716,373,945
Largest scaffold (bp)	1,647,276
N50 (bp)	25,903
L50	5,871
GC (%)	39.61
N base count (bp)	206,359,657
No. N bases per 100 kbp	28,342

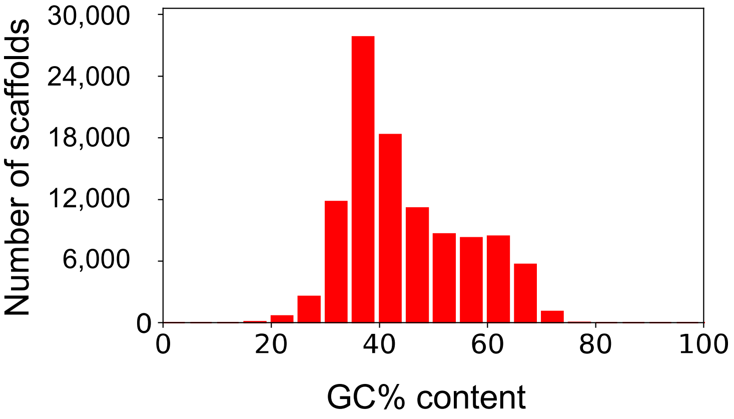


Figure 3.8 Distribution of the GC% of the assembled scaffolds in the *S. rexii* SOAPdenovo2 preliminary assembly

To further explore the possibility of improving the genome assembly, both *S. rexii* datasets generated from HiSeq 4000 and HiSeq X platforms were analysed using ABySS2 and SOAPdenovo2 (Table 3.5 and Figure 3.9). The ABySS2 assembly resulted in 8,985,595 scaffolds comparing to the 3,377,520 scaffolds of the SOAPdenovo2 assembly. The total span of the ABySS2 assembly was 1,610,664,622 bp, which is about 300 Mbp longer than the SOAPdenovo2 assembly (Table 3.5). After filtering out short scaffolds smaller than 500 bp, the metrics of the two assemblies were similar as shown in the cumulative plot and the Nx plot, except for that the ABySS2 assembly was about 200 Mbp shorter (Figure 3.9 a and b). After filtering, there were 191,825 and 271,821 scaffolds remaining in the ABySS2 and SOAPdenovo2 assemblies, respectively. The total span of the ABySS2 assembly was 623 Mbp, and for the SOAPdenovo2 assembly 830 Mbp. The SOAPdenovo2 assembly had the longest scaffold of 3,627,799 bp, compared to the ABySS2 assembly with 1,582,816 bp. The ABySS2 assembly showed a slightly better contiguity with a N50 value of 13,689 bp and a L50 value of 10,075. For the SOAPdenovo2 the N50 value was 8,462 bp and the L50 value 21,884. The presence of the double peaks in the GC% graph was again seen in both assemblies, with a smaller peak at around 60% to 70% GC (Figure 3.9 c).

A major difference between the two assemblies was in the number of N bases (Table 3.5). The SOAPdenovo2 assembly contained about 40 times more N bases than that of the ABySS2 assembly (40,191,947 to 1,482,696 N bases). On average 4,836 N bases were present in every 100 Kbp of the SOAPdenovo2 assembly, which was about 20 times higher than the value of the ABySS2 assembly (237 Ns per 100 Kbp).

Overall the two assemblies were similar in their metrics. Even though the SOAPdenovo2 assembly had a longer total span, it also tended to have a higher proportion of N bases, indicating that many of the assembled sequences were non-informative. In addition, the ABySS2 assembly showed a slightly better contiguity using the same datasets and parameter i.e. better N50 and L50 values (Table 3.5 and Figure 3.9 d). Thus, ABySS2 was chosen for further optimisation to generate the *S. rexii* draft genome using both the HiSeq 4000 and HiSeq X datasets.

Table 3.5 Metrics of the preliminary ABySS2 and SOAPdenovo2 *S. rexii* genome assemblies

	S. rexii **ABySS2** **Preliminary assembly**	***S. rexii*** **SOAPdenovo2** **Preliminary assembly 2**
Dataset used	HiSeq 4000 + HiSeq X	HiSeq 4000 + HiSeq X
No. input read pairs	662,315,056	662,315,056
Assembler	ABySS2 v2.0.2	SOAPdenovo2 v2.04-r240
Assembly parameter	*k* = 77	*k* = 77
Assembly metrics		
Total no. scaffolds	8,985,595	3,377,520
Total span (bp)	1,610,664,622	1,331,851,305
No. scaffolds (≥ 500 bp)	191,825	271,821
Total span (≥ 500 bp) (bp)	623,869,921	830,971,992
Largest scaffold (bp)	1,582,816	3,627,799
N50 (%)	13,689	8,462
L50	10,075	21,884
GC (%)	42.17	44.83
N base count (bp)	1,482,696	40,191,947
No. N bases per 100 kbp	237.66	4,836.74

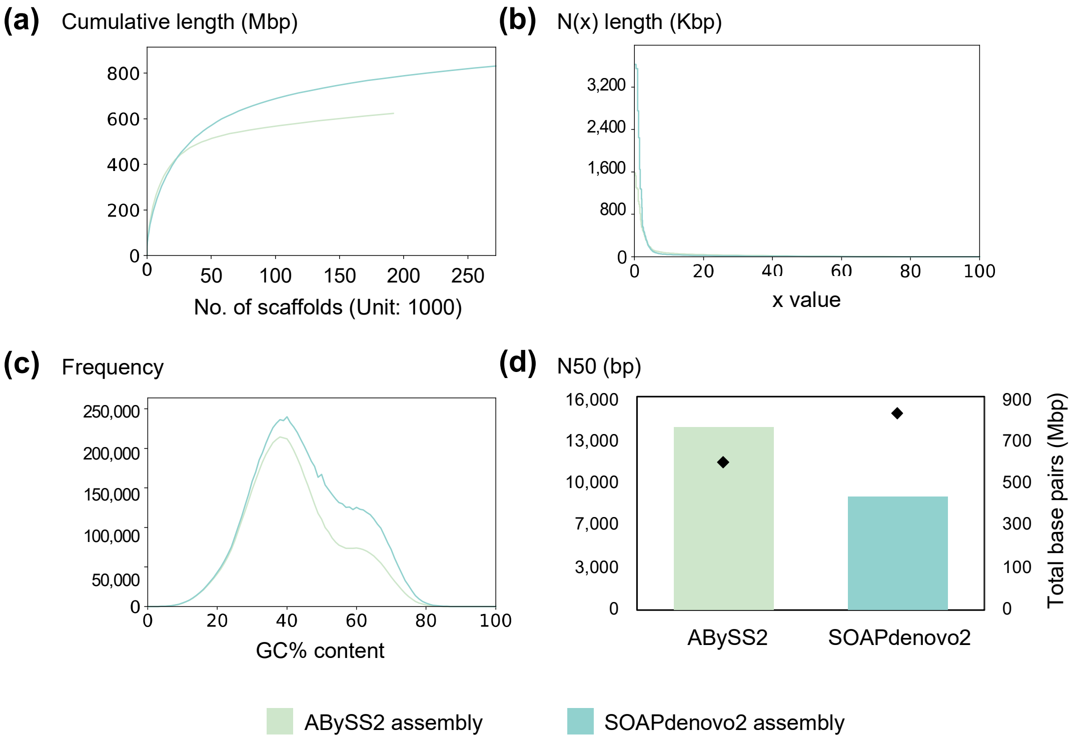


Figure 3.9 Comparisons between the ABySS2 and the SOAPdenovo2 assemblies (k = 77). (a) Cumulative length plot. (b) N(x) length plot. (c) GC% frequency distribution. (d) N50 value and total base pairs assembled. The bars indicate the N50 values and the black diamonds indicate the total length of the assemblies.

3.3.4 Assembly and quality control of the *S. rexii* draft genome

The k-mer histograms of the raw reads were first generated. The k-mer histograms of both datasets showed that the number of distinct k-mers kept increasing until k = 137, and the shape collapsed completely at k = 147 (Figure 3.10). From k-mer counting results, *S. rexii* was estimated to have a genome size between 542 Mbp and 710 Mbp. The estimated heterozygosity of the genome ranged between 0.03% and 0.13%, and the estimated repeat content was between 13% and 27% (Table 3.6).

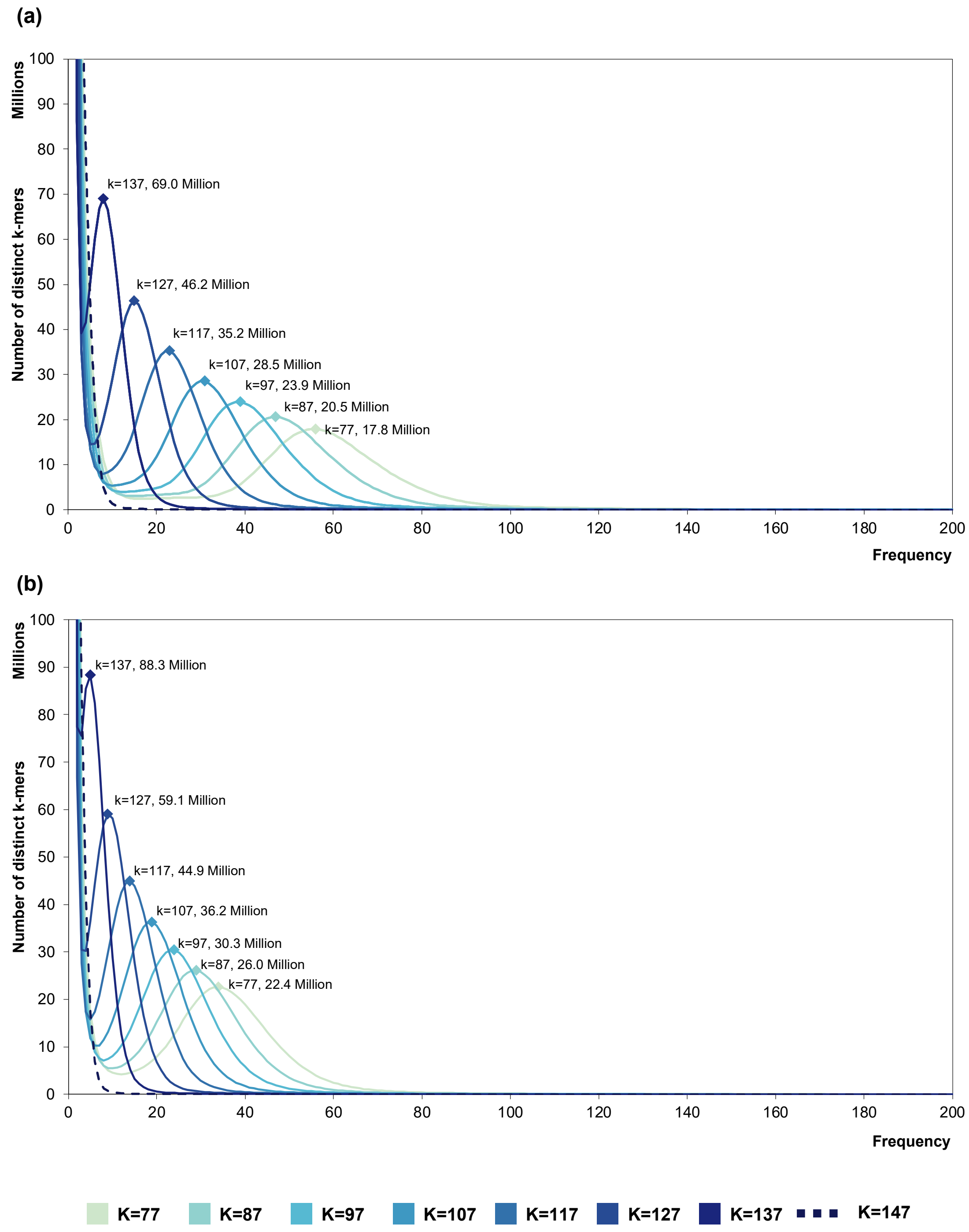


Figure 3.10 *K*-mer histograms of *S. rexii* sequencing data. (A) Histogram of HiSeq X dataset. (B) Histogram of HiSeq 4000 dataset.

Table 3.6 Estimation of *S. rexii* genome size, heterozygosity, and repeat content from *k*-mer count data, from HiSeq X (upper half) and HiSeq 4000 (lower half) data.

HiSeq X	***k* = 77**	***k* = 87**	***k* = 97**	***k* = 107**	***k* = 117**	***k* = 127**	***k* = 137**	***k* = 147**
Genome size (Mbp)	666	674	680	686	701	710	N/A	N/A
Heterozygosity (%)	0.04	0.04	0.05	0.04	0.03	0.03	N/A	N/A
Repeat content (%)	14.94	14.05	13.44	12.92	13.69	13.92	N/A	N/A
Model fit (%)	95.30	96.11	96.80	97.83	98.73	99.25	N/A	N/A
HiSeq 4000	***k* = 77**	***k* = 87**	***k* = 97**	***k* = 107**	***k* = 117**	***k* = 127**	***k* = 137**	***k* = 147**
Genome size (Mbp)	664	673	685	692	690	692	N/A	N/A
Heterozygosity (%)	0.06	0.05	0.05	0.05	0.07	0.04	N/A	N/A
Repeat content (%)	14.78	14.36	14.67	14.50	14.92	14.50	N/A	N/A
Model fit (%)	98.48	98.81	99.09	99.41	98.84	99.41	N/A	N/A

Note. N/A - The GenomeScope tool failed to fit the model to the *k*-mer distribution at *k* = 137 and *k* = 147

ABySS2 was used to assemble the *S. rexii* nuclear genome based on the *k* value range tested above. The HiSeq 4000 and HiSeq X datasets were combined to generate the assembly. After removing short scaffolds < 500 bp, the assembly metrics of the assemblies improved together with higher *k* values in general (Table 3.7 and Figure 3.11): *k* = 117 gave the maximum total span (total span = 637,572,950 bp), as well as the longest scaffold (1,380,762 bp). The best N50 and L50 values were achieved with *k* = 137, giving an N50 value of 35,890 bp and L50 value of 4,568 (Table 3.7 and Figure 3.11 b). Interestingly, a gradual disappearance of scaffolds with 60% - 70% GC content was observed in assemblies with higher *k* values, which in the assembly at *k* = 137 the small lump of high GC content almost disappeared completely (Figure 3.11 c). On the other hand, the assembly at *k* = 147 showed the worst metrics, with the smallest total span and worst contiguity (Table 3.7 and Figure 3.11 d).

Overall, the assembly at *k* = 137 showed the best contiguity metrics, and also a high remapping rate of 95.6% (633,607,562 out of 662,315,056 mapped read pairs), and the mean depth of coverage was 867×. Its shorter total span was possibly due to the absence of the suspiciously high-GC content sequences, i.e. may represent cross-species contamination, rather than the missing of the target plant nuclear genome (Figure 3.11 c). Thus, this assembly was chosen for further quality checks and filtering.

Table 3.7 Metrics of the ABySS2 genome assemblies of *S. rexii* based on HiSeq 4000 + HiSeq X data

	***S. rexii* ABySS2 *k* = 77**	***S. rexii* ABySS2 *k* = 87**	***S. rexii* ABySS2 *k* = 97**	***S. rexii* ABySS2 *k* = 107**
Total no. scaffolds	8,985,595	7,457,747	6,126,393	5,045,696
Total span (bp)	1,610,664,622	1,539,402,274	1,451,571,257	1,358,773,807
No. scaffolds (≥500 bp)	1,582,816	1,104,780	1,105,045	1,274,067
Total span (≥500 bp) (bp)	623,869,921	636,030,438	640,577,967	638,481,047
Longest scaffold (bp)	191,825	168,900	145,355	120,713
N50 (bp)	13,689	17,784	22,373	26,570
L50	10,075	8,227	6,925	6,026
GC (%)	42.17	41.55	40.95	40.3
N base count (bp)	1,482,696	1,197,906	1,018,561	943,364
No. N bases per 100 kbp	237.66	188.34	159.01	147.75

Table 3.7 continued	***S. rexii* ABySS2 *k* = 117**	***S. rexii* ABySS2 *k* = 127**	***S. rexii* ABySS2 *k* = 137**	***S. rexii* ABySS *k* = 147**
Total no. scaffolds	4,082,404	3,125,951	2,604,776	40,527,013
Total span (bp)	1,255,664,373	1,133,015,757	1,051,896,697	6,380,703,103
No. scaffolds (≥500 bp)	103,468	98,936	99,001	7,168
Total span (≥500 bp) (bp)	637,572,950	633,669,601	612,844,571	6,809,895
Longest scaffold (bp)	1,380,762	874,602	886,437	117,219
N50 (bp)	29,452	33,037	35,890	853
L50	5,525	4,984	4,568	1,661
GC (%)	39.75	39.33	38.48	40.06
N base count (bp)	979,516	855,495	961,842	10,110
No. N bases per 100 kbp	153.63	135.01	156.95	148.46

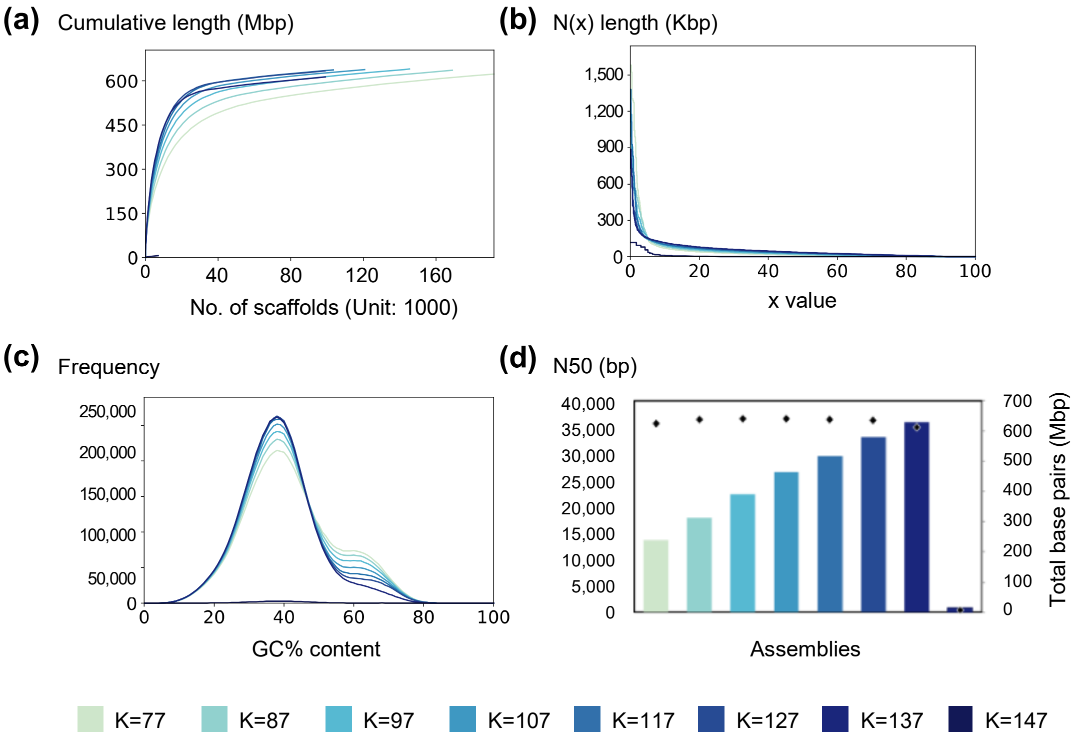


Figure 3.11 Comparisons of *S. rexii* assemblies generated from ABySS2 using different *k* values. All assemblies were generated from HiSeq 4000 and HiSeq X datasets combined. (a) Cumulative length plot. (b) N(x) length plot. (c) GC% frequency distribution. (d) N50 value and total base pairs assembled. The bars indicate the N50 values and the black diamonds indicate the total span of the assemblies.

The blobplot of the assembly at *k* = 137 indicated that the major source of contaminants was bacteria, and predominantly Proteobacteria and Actinobacteria (Figure 3.12). These bacterial scaffolds are the large red and green clusters on the right-hand-side of the plot, with a GC content ranging from 60% to 70%. This corresponded well with the suspiciously high GC distribution observed in previous figures (Figure 3.8, 3.9 c and 3.11 c). The Proteobacteria assemblies consisted of 1,300 scaffolds and spanned about 8.4 Mbp, while the Actinobacteria consisted of 1,753 scaffolds and spanned about 7.4 Mbp (Table 3.8). Other sources of contaminants identified included fungi (Basidiomycota and Ascomycota) and undefined Eukaryotic species. The Basidiomycota consisted of 32 scaffolds (spanning 87,109 bp), the Ascomycota 29 scaffolds (spanning 73,614 bp), and for undefined Eukaryota 28 scaffolds (spanning 144,514 bp). In total, 3,156 scaffolds (spanning c. 16.3 Mbp) were identified as potential contaminants. This comprised about 2.7% of the original assembly (Table 3.8). The complete list of the contaminant species identified is summarised in Appendix 3.3.

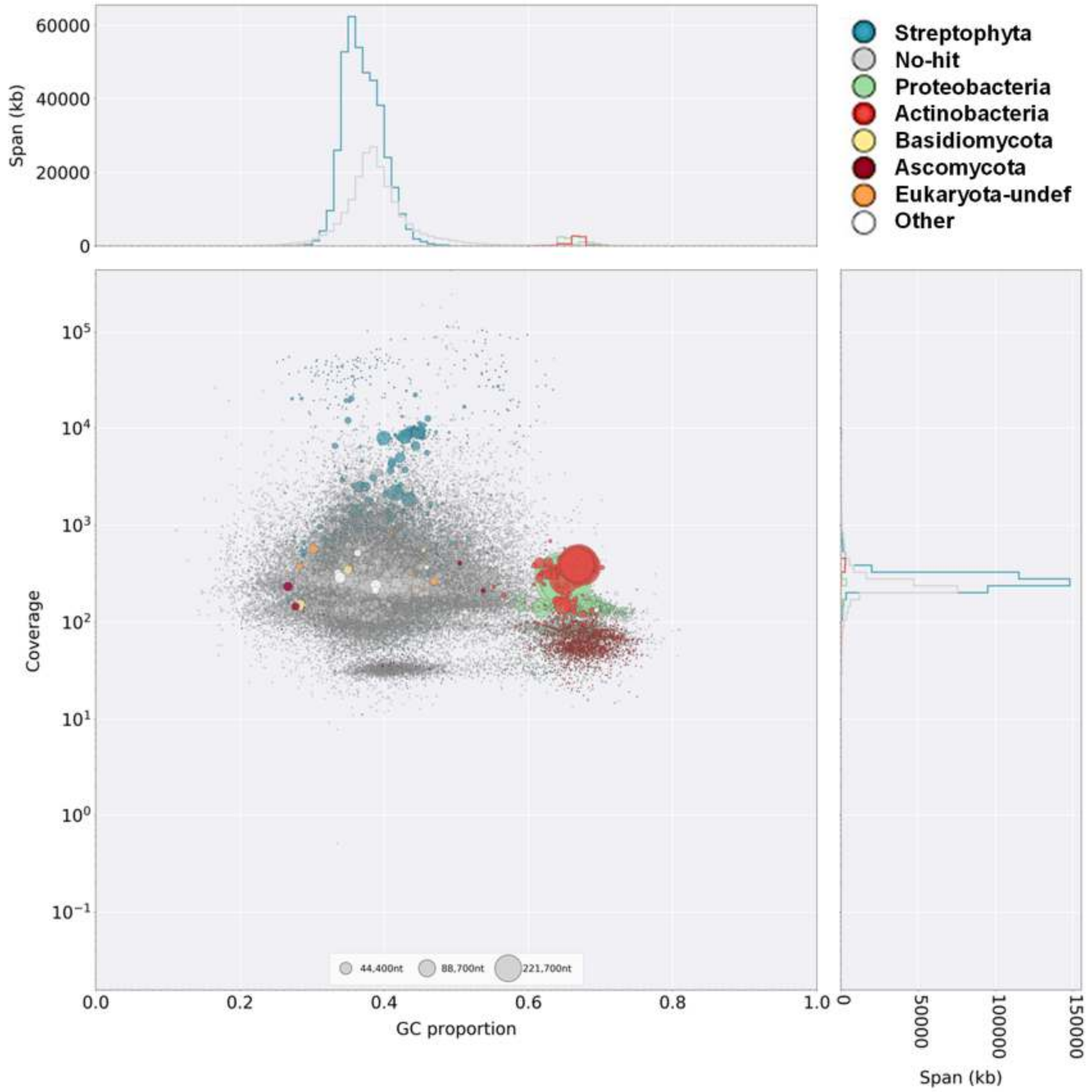


Figure 3.12 Blobplot of the *S. rexii* genome assembly (*k* = 137) before filtering. The colour code indicates the taxon assigned to the scaffolds (shown as circles). The size of the circles indicates the length of the scaffolds. The figure on the top shows the GC distribution of the assembly, and the figure on the right shows the coverage distributions.

Table 3.8 Summary of the potential contaminant scaffolds identified in the *S. rexii* genome assembly.

Category	**No. scaffolds**	**Total span (bp)**	**Total span (%)**	**Average scaffold length (bp)**
Actinobacteria	1,753	7,444,060	45.574	4,246
Proteobacteria	1,300	8,425,972	51.586	6,481
Basidiomycota	32	87,109	0.533	2,722
Ascomycota	29	73,614	0.451	2,538
Undefined Eukaryota	28	144,514	0.885	5,161
Arthropoda	6	4,165	0.025	694
Mucoromycota	4	2,056	0.013	514
Undefined bacteria	4	25,732	0.158	6,433
Undefined viruses	4	40,410	0.247	10,102
Chordata	3	42,696	0.261	14,232
Nematoda	3	2,005	0.012	668
Chlorophyta	2	1,068	0.007	534
Apicomplexa	1	38,912	0.238	38,912
Bacteroidetes	1	601	0.004	601
Undefined	1	556	0.003	556
Unresolved	1	517	0.003	517
TOTAL	3,172	16,333,987	100.000	5,149

After removing the contaminant scaffolds (keeping only the Streptophyta sequences and "no-hit" scaffolds), the filtered assembly had 95,845 scaffolds remaining with a total span of 596.6 Mbp (Table 3.9). The N50 value after filtering was 35,609 bp, which was only slightly lower than the unfiltered assembly. The average GC content after filtering was 37.75% (Table 3.9). The GC distribution graph shows that the high-GC peak disappeared after filtering (Figure 3.13 c). The unfiltered and filtered genomes had similar BUSCO completeness percentages of approximately 88%. Interestingly, after filtering, the BUSCO completeness was slightly higher (87.3% *versus* 88.8%) (Table 3.9). Finally, the blobplot of the filtered genome assembly showed a cleaner pattern without bacterial scaffolds (Figure 3.14). At this point, the finalised *S. rexii* draft genome assembly was generated in this study.

Table 3.9 Metrics of the unfiltered and filtered *S. rexii* genome assemblies

	S. rexii **ABySS2** ***k*** **= 137 unfiltered**	***S. rexii*** **ABySS2** ***k*** **= 137 filtered**
Assembly metrics		
Total no. scaffolds	2,604,776	95,845
Total span (bp)	1,051,896,697	596,583,869
No. scaffolds (≥500 bp)	99,001	95,845
Total span (≥500 bp) (bp)	612,844,571	596,583,869
Largest scaffold (bp)	886,437	421,987
N50 (bp)	35,890	35,609
L50	4,568	4,571
GC (%)	38.48	37.75%
N base count (bp)	961,842	907,432
No. N bases per 100 kbp	156.95	152.10
Genome completeness		
BUSCO completeness (%)	87.3	88.8
No. complete BUSCOs	1,257	1,322
No. fragmented BUSCOs	41	43
No. missing BUSCOs	183	161

Note. A total of 1,440 BUSCOs were searched

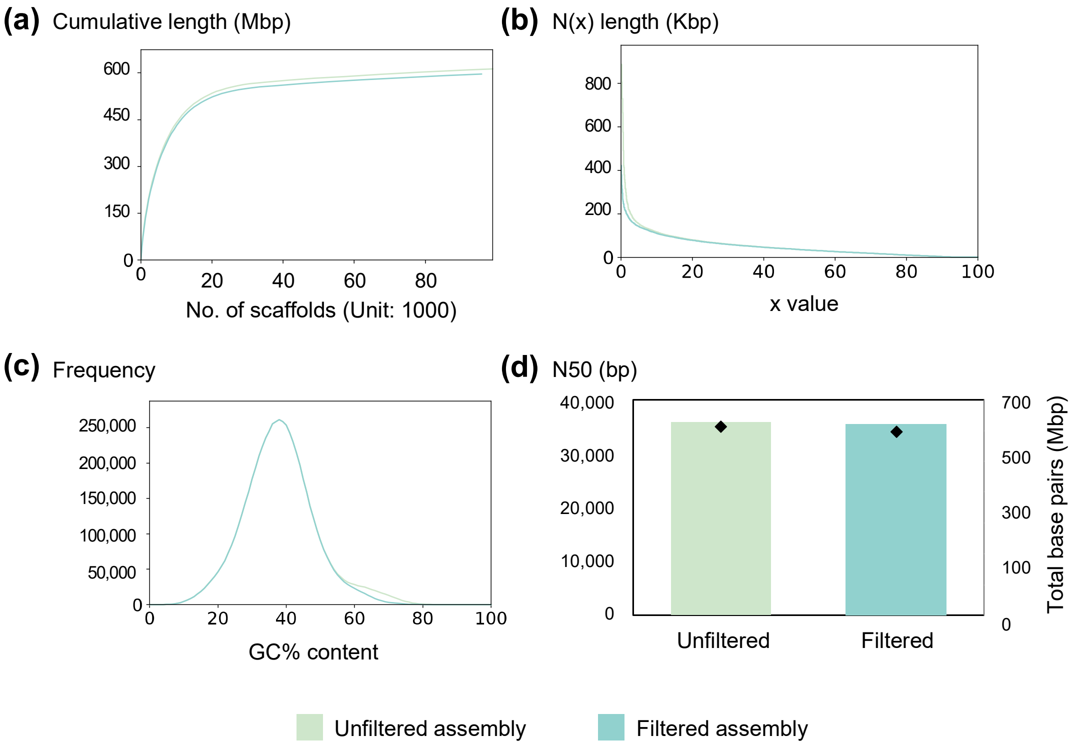


Figure 3.13 Comparisons of the unfiltered and filtered *S. rexii* genome assemblies. (a) Cumulative length plot. (b) N(x) length plot. (c) GC% frequency distribution. (d) N50 value and total base pairs assembled. The bars indicate the N50 value and the black diamonds indicate the total span of the assemblies.

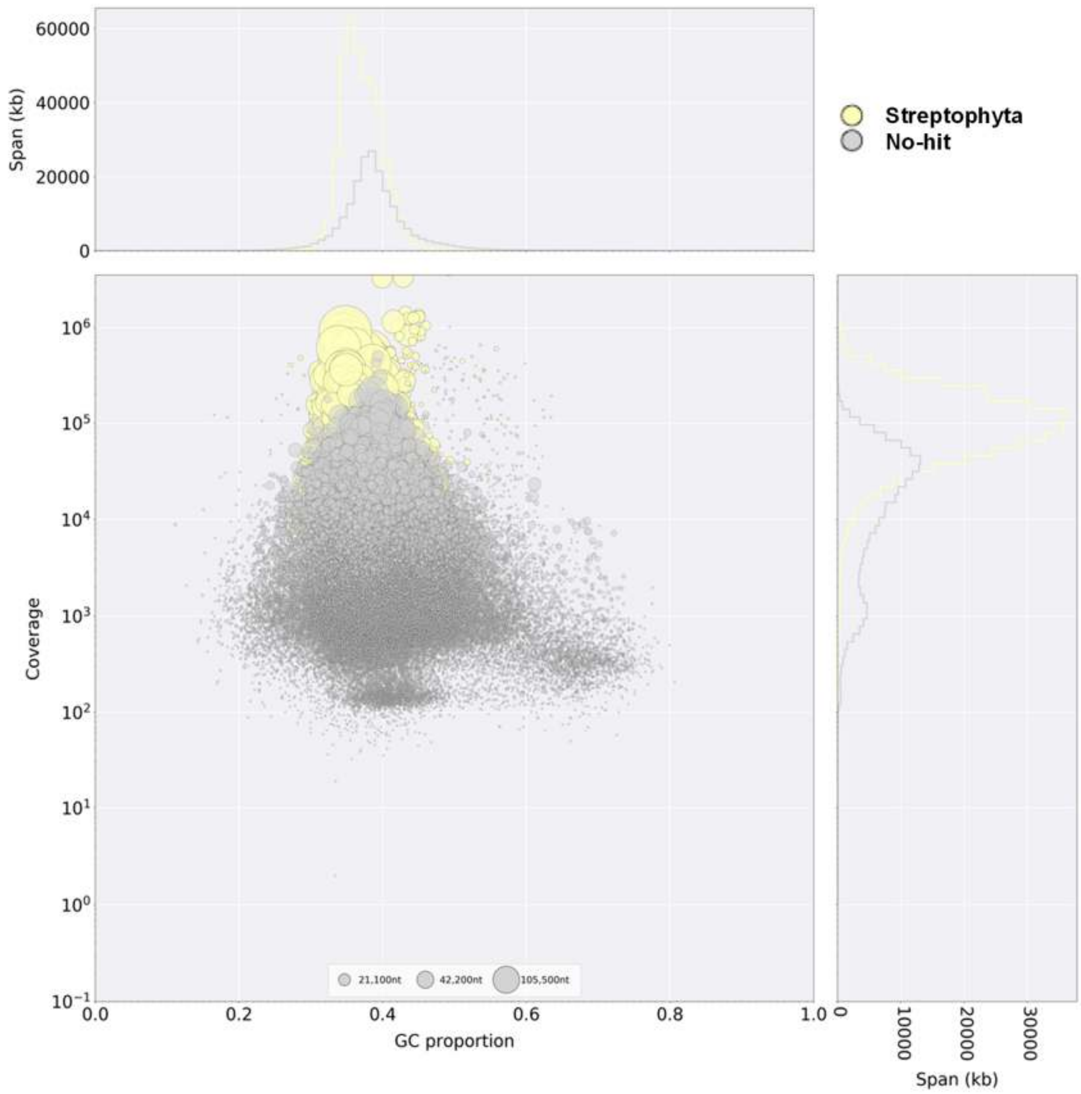


Figure 3.14 Blobplot of the *S. rexii* genome assembly (*k* = 137) after filtering. The colour code indicates the taxon assigned to the scaffolds (shown as circles). The size of the circles indicates the length of the scaffolds. The figure on the top shows the GC distribution of the assembly, and the figure on the right shows the coverage distributions.

3.3.5 Assembly and quality control of the *S. grandis* draft genome

The data generated from the HiSeq X sequencing experiments was used for the assembly. First, *k*-mer histograms were analysed, which suggested that optimal *k* values for the assembly were probably around 127 to 137, as these two values gave the highest number of distinct *k*-mers (Figure 3.15). The estimated genome size was about 990 Mbp to 1,003 Mbp, with 0.02% - 0.03% heterozygosity, and 12% - 15% repeat content (Table 3.10).

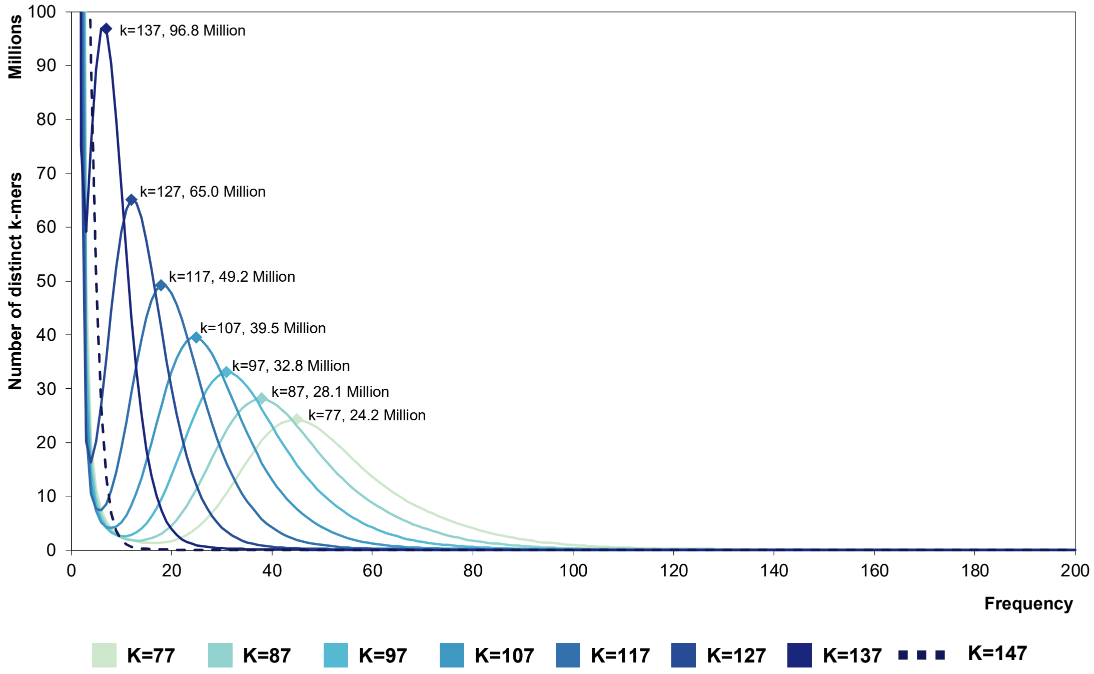


Figure 3.15 *K*-mer histogram of the *S. grandis* sequencing data generated from the HiSeq X experiment.

Table 3.10 Estimation of *S. grandis* genome size, heterozygosity, and repeat content from *k*-mer count data based on the HiSeq X dataset.

HiSeq X	***k* = 77**	***k* = 87**	***k* = 97**	***k* = 107**	***k* = 117**	***k* = 127**	***k* = 137**	***k* = 147**
Genome size (Mbp)	974	985	992	997	1,003	996	N/A	N/A
Heterozygosity (%)	0.07	0.02	0.02	0.02	0.03	0.03	N/A	N/A
Repeat content (%)	16.84	16.34	15.22	13.94	12.56	14.85	N/A	N/A
Model fit (%)	94.82	95.64	96.84	97.92	98.79	98.65	N/A	N/A

Note. N/A - The GenomeScope tool failed to fit the model to the *k*-mer distribution at *k* = 137 and *k* = 147

Based on the *k*-mer histogram results, the genome assembly was carried out at *k* values ranging from 107 to 137 (Table 3.11, Figure 3.16). The assembly with *k* = 127 resulted in the overall best assembly. After removing short scaffolds (< 500 bp), the assembly had 738,916 scaffolds and a total span of around 845 Mbp (Table 3.11, Figure 3.16 a), and had the highest N50 value (31,553 bp) and lowest L50 value (7,246) among the four assemblies. On the other hand, the assembly with *k* = 137 was highly fragmented, with 276,087 scaffolds and an N50 value of 7,271 bp (Table 3.11, Figure 3.16). The GC plot showed a normal distributed single peak centred at 36% to 38% GC content (Figure 3.16 c). As the assembly at *k* = 127 gave the best results for other parameters, this assembly was chosen for contaminant identification and further quality checks. This assembly also had a high remapping rate of 99.4% (450,658,777 mapped among 452,684,043 read pairs), and a depth of average coverage of 140×.

Table 3.11 Metrics the ABySS2 *S. grandis* genome assemblies based on HiSeq X dataset.

	S. grandis **ABySS2** ***k*** **= 107**	***S. grandis*** **ABySS2** ***k*** **= 117**	***S. grandis*** **ABySS2** ***k*** **= 127**	***S. grandis*** **ABySS2** ***k*** **= 137**
Total no. scaffolds	3,397,906	2,794,551	2,252,545	4,790,734
Total span (bp)	1,287,619,502	1,253,615,696	1,211,718,685	1,533,120,309
No. scaffolds (≥500 bp)	128,374	121,887	128,388	276,087
Total span (≥500 bp) (bp)	795,183,443	824,378,614	844,897,632	682,013,812
Largest scaffold (bp)	526,028	530,645	738,916	391,136
N50 (bp)	26,343	30,389	31,553	7,271
L50	8,065	7,310	7,246	17,910
GC (%)	38.25	38.31	38.33	36.70
N base count (bp)	721,914	738,806	859,536	1,969,208
No. N bases per 100 kbp	90.79	89.62	101.73	288.73

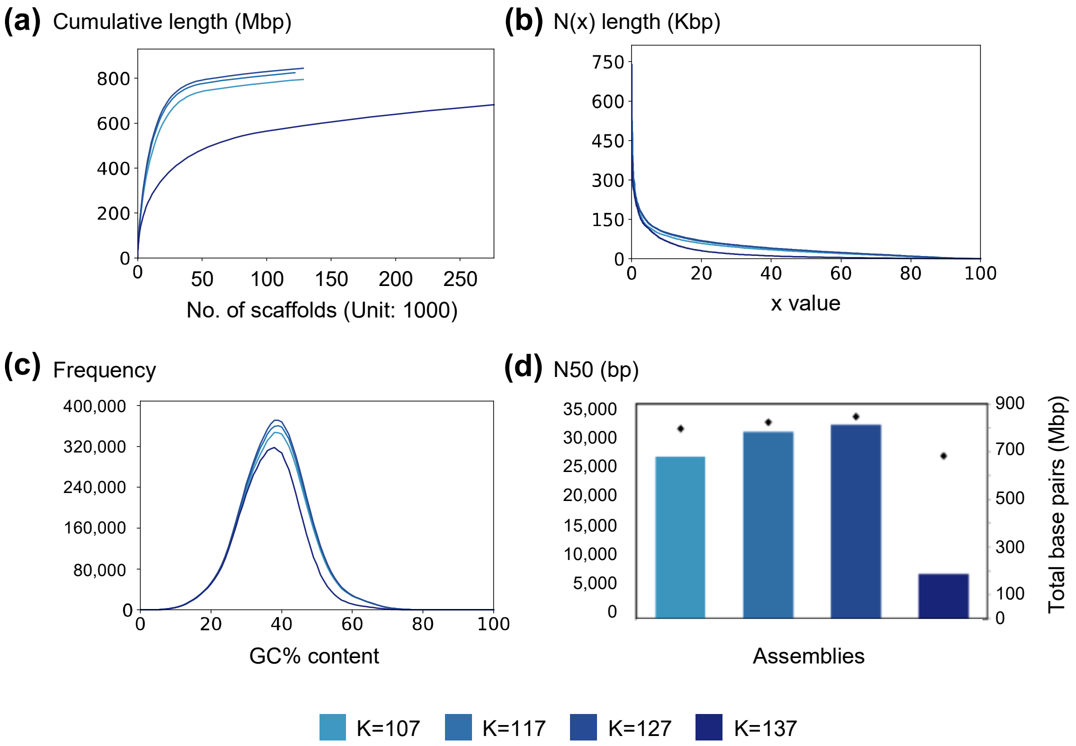


Figure 3.16 Comparisons of *S. grandis* assemblies generated from ABySS2 using different *k* values. All assemblies were generated from HiSeq X dataset. (A) Cumulative length plot. (B) N(x) length plot. (C) GC% frequency distribution. (D) N50 value and total base pairs assembled. The bars indicate the N50 value and the black diamonds indicate the total span of the assemblies.

Analysis of the blobplot revealed 437 potential contaminant scaffolds, with a total span of about 1.6 Mbp (Figure 3.17 and Table 3.12). The major source of contamination was from Proteobacteria (Figure 3.17 green circles), which contained 426 scaffolds spanning 1.4 Mbp (Table 3.12). A detailed list of the contaminant species identified are summarised in Appendix 3.4.

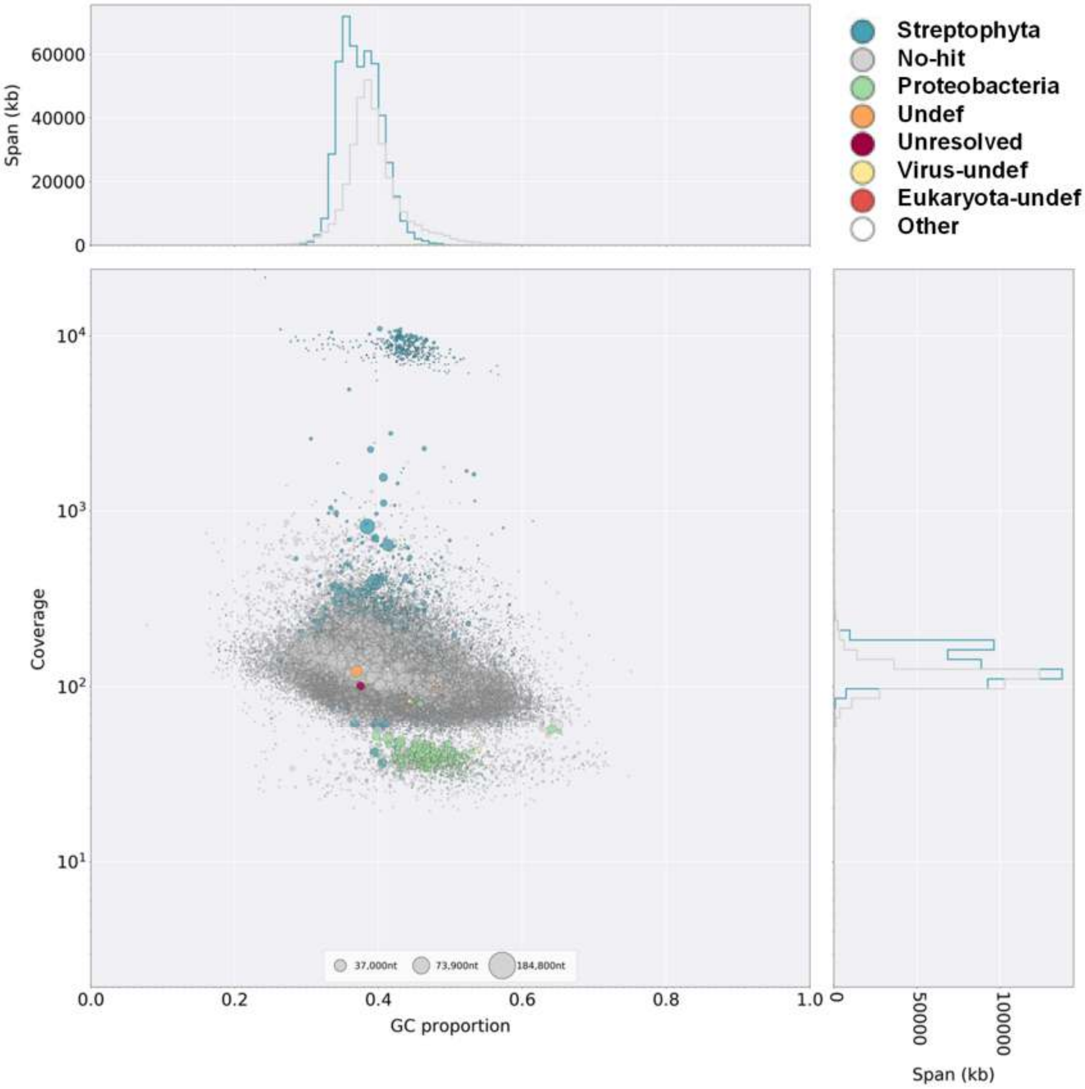


Figure 3.17 Blobplot of the *S. grandis* genome assembly (*k* = 127) before filtering. The colour code indicates the taxon assigned to the scaffolds (shown as circles). The size of the circles indicates the length of the scaffolds. The figure on the top shows the GC distribution of the assembly, and the figure on the right shows the coverage distributions.

Table 3.12 Summary of the potential contaminant scaffolds identified in *S. grandis* genome assembly based on HiSeq X dataset.

Category	No. scaffolds	Total span (bp)	Total span (%)	Average scaffold length (bp)
Proteobacteria	426	1,484,488	94.68	3,484.7
Undefined Viruses	3	15,357	0.98	5,119
Undefined Eukaryota	3	4,149	0.26	1,383
Undefined	2	42,177	2.69	21,088.5
Unresolved	1	17,138	1.09	17,138
Chordata	1	3,881	0.25	3,881
Undefined bacteria	1	734	0.05	734
TOTAL	437	1,567,924	100.00	3,587.9

The identified contaminant sequences were removed, and only the Streptophyta sequences and "no-hit" scaffolds were kept. Since there were very little contaminations, the metrics of the filtered and unfiltered genomes were nearly identical (Table 3.13, Figure 3.18). The filtered assembly consisted of 127,951 scaffolds with a total span of 843.3 Mbp (Table 3.13, Figure 3.18). The N50 and L50 was 31,638 bp and 7,221, respectively, and the average GC proportion was 38.31% (Table 3.13). Both filtered and unfiltered assemblies had similar BUSCO completeness of 88.5%. Inspection of the post-filtering blobplot confirmed that the bacterial and other contaminant sequences were effectively removed (Figure 3.19). Here the finalised *S. grandis* genome assembly with the data generated in this study was obtained.

Table 3.13 Metrics of the unfiltered and filtered *S. grandis* genome assemblies based on HiSeq X dataset. A total of 1,440 BUSCOs were searched.

	S. grandis **ABySS2** ***k*** **= 127 unfiltered**	***S. grandis*** **ABySS2** ***k*** **= 127 filtered**
Assembly metrics		
Total no. scaffolds	2,252,545	127,951
Total span (bp)	1,211,718,685	843,329,708
No. scaffolds (≥500 bp)	128,388	127,951
Total span (≥500 bp) (bp)	844,897,632	843,329,708
Largest scaffold (bp)	738,916	738,916
N50 (bp)	31,553	31,638
L50	7,246	7,221
GC (%)	38.33	38.31
N base count (bp)	1,232,681	874,532
No. N bases per 100 kbp	101.73	100.37
Genome completeness		
BUSCO completeness (%)	88.5	88.6
No. complete BUSCOs	1,275	1,276
No. fragmented BUSCOs	214	36

No. missing BUSCOs	128	128

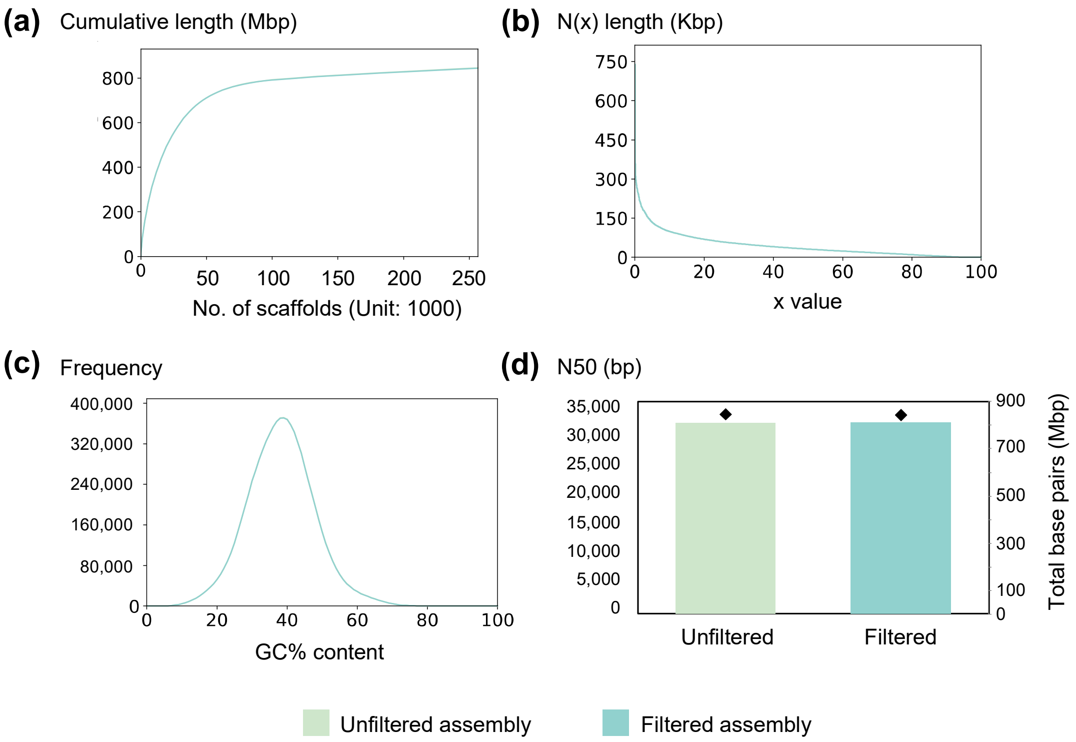


Figure 3.18 Comparison of the unfiltered and filtered *S. grandis* genome assemblies based on HiSeq X data. (a) Cumulative length plot. (b) N(x) length plot. (c) GC% frequency distribution. (d) N50 value and total base pairs assembled. The bars indicate the N50 value and the black diamonds indicate the total span of the assembly. Because the two assemblies are nearly identical, in (a) – (c) the two lines overlap completely and cannot be distinguished.

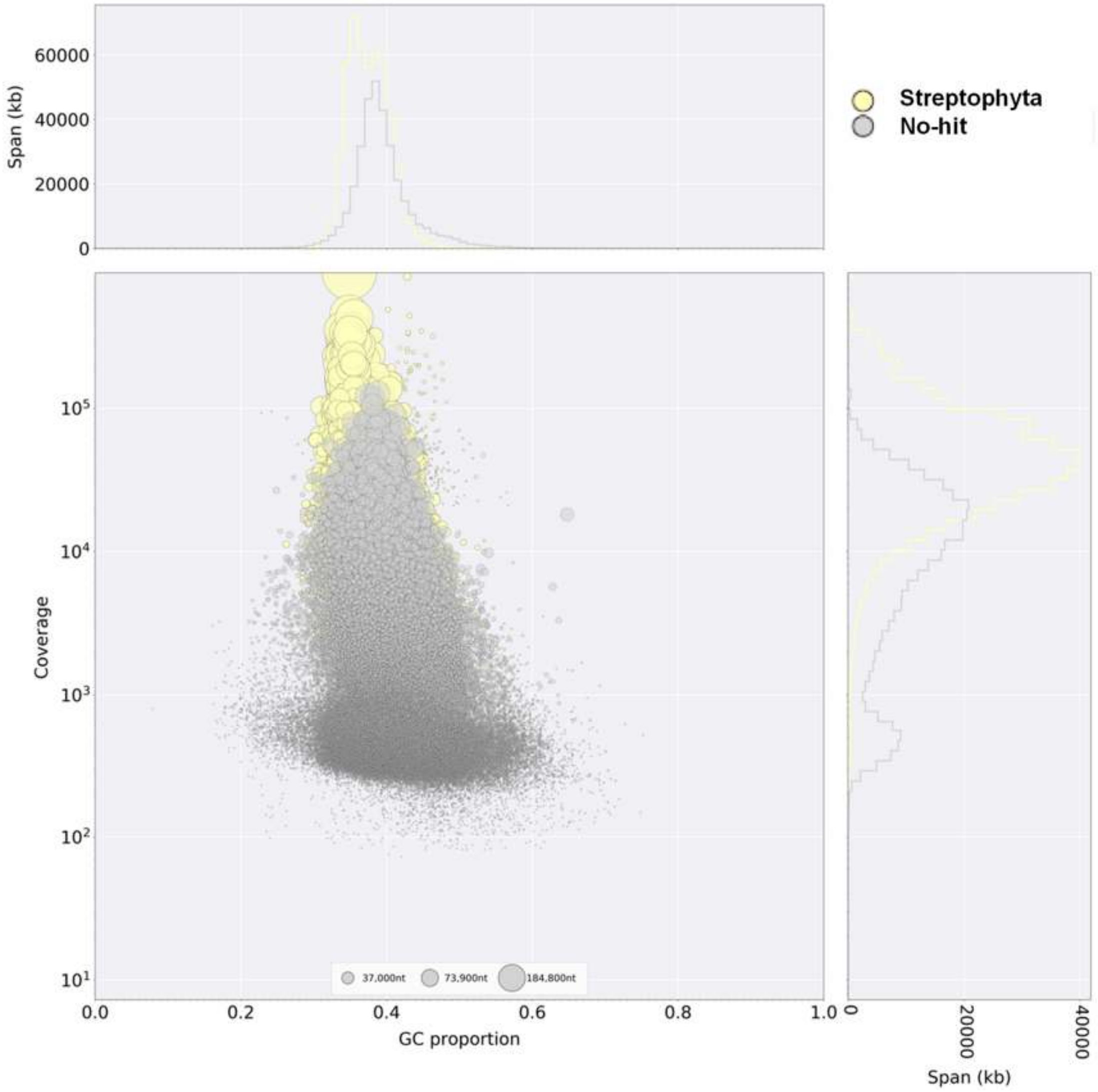


Figure 3.19 Blobplot of the *S. grandis* genome assembly (*k* = 127) based on HiSeq X data after filtering. The colour code indicates the taxon assigned to the scaffolds (shown as circles). The size of the circle indicates the length of the scaffolds. The figure on the top shows the GC distribution of the assembly, and the figure on the right shows the coverage distributions.

A rough comparison between the filtered *S. rexii* and *S. grandis* genome assemblies was made by mapping the *S. grandis* to the *S. rexii* assembly (Figure 3.20). The two assemblies were rather dissimilar with only about 20.7% (about 174 Mbp) of the assemblies matching. Among these, 12.7% (about 107 Mbp) showed an above 75% similarity. Also, a large proportion of the *S. grandis* assemblies were not identified in the *S. rexii* assembly (Figure 3.20 a). However, this is still relatively similar comparing to the results where the two *Streptocarpus* genomes were aligned to the *D. hygrometricum* genome (Figure 3.20 b and c). In these comparisons, the *S. rexii* and *S. grandis* assemblies matched only about 3%

to the *D. hygrometricum* genome. This indicated that the *Streptocarpus* assemblies were very different from the *Dorcoceras* assembly, and the two *Streptocarpus* assemblies might also be dissimilar between themselves, but less so.

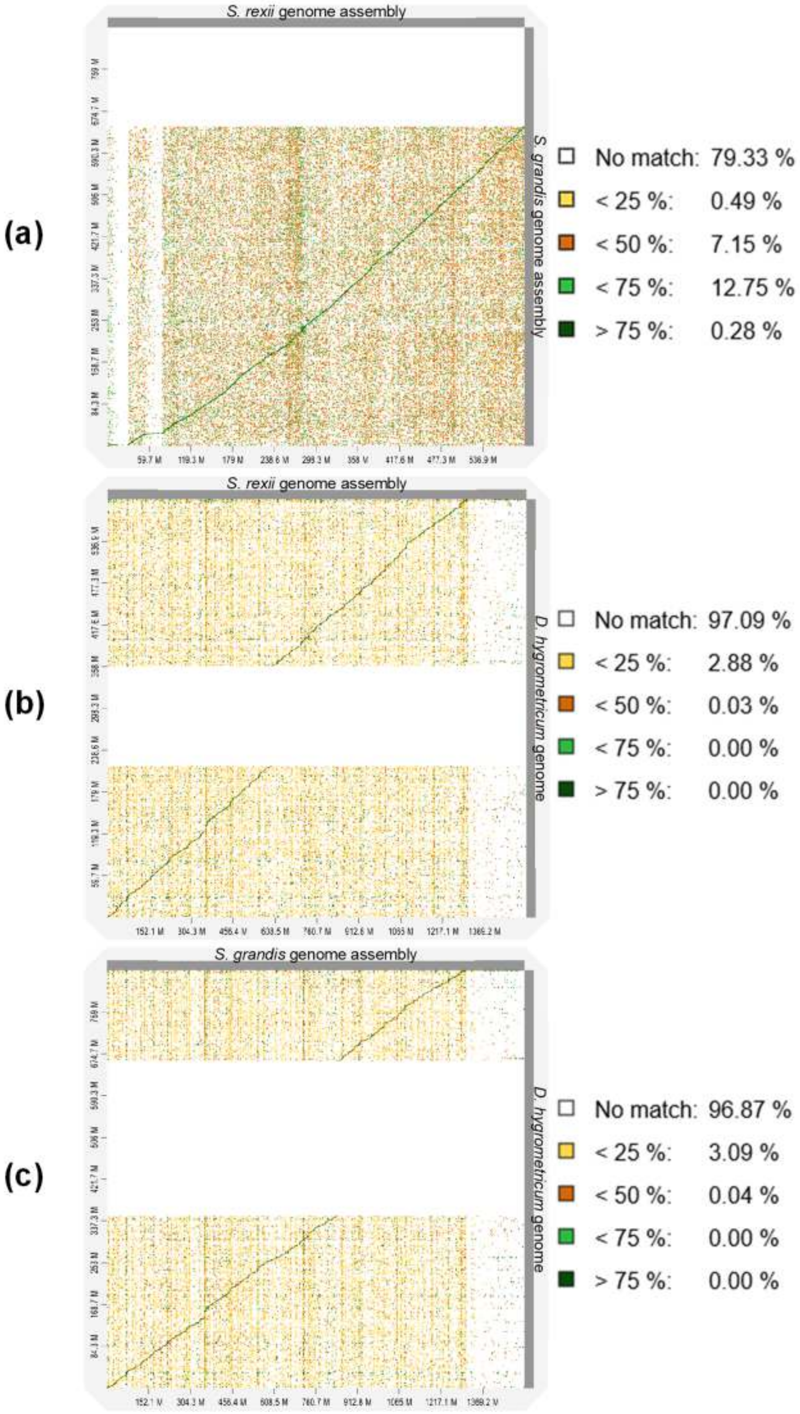


Figure 3.20 Dot plot of the comparison between two *Streptocarpus* genome assemblies and the *Dorcorceras* genome assembly (NCBI accession: GCA_001598015.1). (a) *S. rexii* v.s. *S. grandis*. (b) *S. rexii* v.s. *D. hygrometricum*. (c) *S. grandis* v.s. *D. hygrometricum*.

3.4 Discussion

3.4.1 Assembly of the *Streptocarpus* organellar genomes

The *S. rexii* and *S. grandis* plastid genomes share similar assembly metrics (Table 3.14). The two assemblies also share a high sequence identity of 99.65%, and can be aligned and identified as a single synteny block, indicating their identical genome structure (Figure 3.4). The differences found concerned deletions and SNPs between *S. rexii* and *S. grandis* plastid assemblies. For instance, a 65 bp deletion in the *trn*L-F region is present in the *S. rexii* plastid but not in *S. grandis*, which is congruent with previous observations (Möller et al., 2004).

It is known that the plastid structure and sequences are conserved among currently sequenced Lamiales species including Gesneriaceae, and their total plastid genome size is around 153 Kbp (Kyalo et al., 2018). This was also observed in our results (Table 3.14). In terms of gene content, the plastid genomes of the three *Streptocarpus* species and the *D. hygrometricum* have 107 and 103 protein coding genes annotated, respectively, roughly 20 proteins more compared to *Haberlea rhodopensis* Friv. and *Lysionotus pauciflorus* Maxim. (Table 3.14; Ren et al., 2016; Ivanova et al., 2017). Interestingly, all species had eight rRNAs except for *S. teitensis* and *H. rhodopensis* with four rRNAs. It is possible that this difference was due to the different annotation pipelines used. As shown in Appendix 3.5, when annotating the *S. teitensis* and *H. rhodopensis* plastids using GeSeq tool (method described in section 3.2.3), eight rRNA genes were identified in these assemblies. A more comprehensive comparison is required, such as annotating all genomes using the same annotation pipeline, to confirm whether the differences observed is due to pipeline or actual genetic differences.

S. teitensis was placed in subgenus *Streptocarpella* in *Streptocarpus*, while *S. rexii* and *S. grandis* reside in subgenus *Streptocarpus* (Nishii et al., 2015). This classification was reflected in the high similarities of the plastid genomes of *S. rexii* and *S. grandis*, and wider distance to *S. teitensis* whose plastid genome was about 500 bp longer than in the two *Streptocarpus* species in this study (Table 3.14). Comparative study of these three plastids sequences also in relation to other Gesneriaceae genomes may provide interesting insights into the evolution of the chloroplast genome in Gesneriaceae (Kyalo et al., 2018).

Table 3.14 Plastid assembly metrics across species gathered from the present study and previously published studies

	Streptocarpus rexii	***Streptocarpus grandis***	***Streptocarpus teitensis***
Total span (bp)	152,724	152,770	153,207
No. protein coding genes	107	107	116
No. rRNA annotated	8	8	4
No. tRNA annotated	33	33	32
Reference	This study	This study	Kyalo et al., 2018

	Dorcoceras hygrometricum	***Lysionotus pauciflorus***	***Haberlea rhodopensis***
Total span (bp)	153,493	153,856	153,099
No. protein coding genes	103	88	86
No. rRNA annotated	8	8	4
No. tRNA annotated	36	37	36
Reference	Zhang et al., 2012	Ren et al., 2016	Ivanova et al., 2017

While the chloroplast genome organisation was highly conserved between the two species studied here, the mitochondrial assemblies on the other hand were highly variable between *S. rexii* and *S. grandis* and differed in their statistics and annotation results (Table 3.15). The *S. rexii* mitochondrial genome was about 40,000 bp shorter than that of the *S. grandis*, and had 7 fewer protein coding genes and missed 1 rRNA gene compared to *S. grandis* (Table 3.15). The sequence alignment between the two assemblies revealed large gaps and 1,080 SNPs. The synteny analysis between the two assemblies indicated that the two genomes greatly differed in their structure. One large proportion of the *S. rexii* assembly was not even identified in *S. grandis*, and *vice versa* (Figure 3.6). In particular, the mitochondrial complex III was not found in the *S. rexii* mitochondrial genome (Figure 3.5). This complex consisted of cytochrome c reductase, involved in the electron transport chain reaction, which is a key component for mitochondrion functioning (Siedow and Umbach, 1995). It is possible that the absence of complex III in *S. rexii* is related to the 40,000 bp difference between the *S. rexii* and *S. grandis* mitochondrial assemblies. A detailed examination should be made of the sequence alignment between the two assemblies to ascertain whether the *S. rexii* assembly has failed to recover the sequence of complex III due to misassembly.

When comparing the *Streptocarpus* mitochondrial genome assemblies with other Lamiales species, both *Streptocarpus* assemblies were found to be much smaller (about 200 Kbp shorter) than those of *D. hygrometricum* (Zhang et al., 2012) and *E. guttata* (Mower et al., 2012) (Table 3.15). Despite the difference in assembly size, the two *Streptocarpus* assemblies are still within the typical range of angiosperm mitochondrial genomes (200 Kbp to 750 Kbp; Gualberto et al., 2014). However, both *Streptocarpus* assemblies had almost twice as many protein coding genes identified (Table 3.15). This is again likely to be due to the different annotation pipelines used. As shown in Appendix 3.6, when annotating the *D. hygrometricum* and *E. guttata* mitochondrial genomes using the method described in this study, 152 and 149 protein coding genes were identified in the two assemblies, respectively.

Table 3.15 Mitochondria assembly metrics across species gathered from the present study and previously published studies

	Streptocarpus rexii	***Streptocarpus grandis***	***Dorcoceras hygrometricum***	***Erythrante guttata***
Total span (bp)	314,134	352,540	510,519	525,671
No. protein coding genes	72	79	33	35
No. rRNA annotated	2	3	4	3
No. tRNA annotated	17	17	28	24
Reference	This study	This study	Zhang et al., 2012	Mower et al., 2012

Since the main objective of this study was not to analyse the organellar genomes themselves, only the basic assemblies and annotation metrics were analysed and compared here. A more thorough analysis and characterisation of the *Streptocarpus* plastid and mitochondrial genomes would be desirable but is beyond the scope of the present study.

3.4.2 Assembly of the *Streptocarpus* nuclear genome

The *S. rexii* and *S. grandis* were sequenced, assembled, and compared. The ABySS2 and SOAPdenovo2 assemblers were first tested on the *S. rexii* dataset, which the ABySS2 assembler was chosen for further optimisation. The ABySS2 assembler was finally used to reconstruct the *S. rexii* genome at k = 137 and the *S. grandis* genome at k = 127. For both assemblies their contaminants were filtered out, and the quality and BUSCO completeness assessed. The metrics of the final assembly for both species are summarised in Table 3.16.

Table 3.16 Assembly metrics of the *S. rexii* and *S. grandis* genomes. A total of 1,440 BUSCOs were searched.

	S. rexii **ABySS2** ***k* = 137** **filtered**	***S. grandis*** **ABySS2** ***k* = 127** **filtered**
Assembly metrics		
Total no. scaffolds	95,845	127,951
Total span (bp)	596,583,869	843,329,708
No. scaffolds (≥500 bp)	95,845	127,951
Total span (≥500 bp) (bp)	596,583,869	843,329,708
Largest scaffold (bp)	421,987	738,916
N50 (bp)	35,609	31,638
L50	4,571	7,221
GC (%)	37.75	38.31
N base count (bp)	907,432	874,532
No. N bases per 100 kbp	152.10	100.37
Genome completeness		
BUSCO completeness (%)	88.8	88.6
No. complete BUSCOs	1,279	1,276
No. fragmented BUSCOs	43	36
No. missing BUSCOs	118	128

The total span of the *S. rexii* assembly was about 596 Mbp, which was 247 Mbp smaller than the *S. grandis* assembly (843 Mbp). The *S. rexii* assembly had fewer scaffolds assembled; the N50 value was about 4,000 bp longer than the *S. grandis* assembly, and the L50 value lower, suggesting an overall better contiguity of the *S. rexii* assembly. However, *S. rexii* also had more N bases in the assembly. This is possibly due to the usage of a library with longer insert size (550 bp), which was not used for the *S. grandis* (only with insert size of 350 bp), thus more read pair information was utilised for scaffolding and more gaps were created.

There was a discrepancy between the estimated genome size and the assembly total spans. The *S. rexii* and *S. grandis* genome size estimation by flow cytometry gave C-values of 929 Mbp and 1,260 Mbp respectively (Möller, 2018); the estimations obtained from the *k*-mer histograms were significantly lower, 542 Mbp to 710 Mbp for *S. rexii*, and 990 Mbp to 1,003 Mbp for *S. grandis* (Tables 3.6, 3.10). Both C-value and *k*-mer estimations were larger than the final genome assemblies (Table 3.16). It is known that repeat content of genomes are difficult to be assembled (Claros et al., 2012; Compeau et al., 2017). Plants with smaller genomes such as *A. thaliana* (~135 Mbp; Rhee et al., 2003) tend to have fewer but longer repeat sequences (2 Kbp to 6 Kbp) interspersed among longer non-repetitive regions, while plants with relatively large genome sizes such as maize (~2.1 Gbp; Hirsch et al., 2016) tend to have many shorter repeated sequences (50 bp to 2 Kbp) interspersed among shorter non-repetitive sequences, and are more difficult to assemble (Lapitanz, 1992). Both *Streptocarpus* species have genome sizes around 1 Gbp, and the repeat content estimated from the *k*-mer histograms was about 12% to 16% (Table 3.6 and 3.10). It is possible that the repeat content of the *Streptocarpus* genomed were not reconstructed, thus the final assembly size is smaller than the estimated genome size. Alternatively, it is possible that the flow cytometry results are inconsistent and the genome size was estimated incorrectly. For example, two very different 1C value for *Streptocarpus cyaneus* were reported, 0.875 pg (Möller, 2018) and 0.675 pg (Hansen et al., 2001), which correspond to 855 Mbp and 660 Mbp respectively. Another example is *Streptocarpus ionantha*, which its 1C value was independently calculated as 0.87 pg (Möller, 2018) and 0.75 pg (Loureiro et al., 2007), corresponding to 850 Mbp and 733 Mbp respectively. This suggests that the flow cytometry results can have 100 Mbp to 200 Mbp variations, and maybe the flow cytometry has overestimated the *Streptocarpus* genome size while the assemblies presented here underestimated it.

Both *Streptocarpus* assemblies showed similar BUSCO completeness: the *S. rexii* had 1,279 BUSCO identified (88.8%) and *S. grandis* 1,276 BUSCO identified (88.6%) among the 1,440 BUSCOs searched. The *S. rexii* assembly also had more fragmented BUSCOs (43) and less missing BUSCOs (118) than the *S. grandis* assembly (36 fragmented and 128 missing), implying that the *S. rexii* assembly was able to reconstruct more core genes even if they were fragmented. Interestingly, in both assemblies the BUSCO completeness improved after filtering out short scaffolds and cross-species contaminants. This suggests that our filtering strategy was able to improve the assembly quality, but at the same time was not too stringent as to remove key information from the genome.

Genome-to-genome alignment between *S. rexii* and *S. grandis* assemblies suggested a low similarity, where only 12.18% of the *S. rexii* assembly matching the *S. grandis* assembly with an identity higher than 50%. Likewise, 79.33% of the *S. rexii* assembly failed to match *S. grandis* (Figure 3.20 a). The dot plots suggested that there was a large proportion of the *S. grandis* genome that cannot be identified in *S. rexii* genome, and *vice versa* (Figure 3.20 a). This difficulty encountered in aligning the two genomes may be related to the phylogenetic distance between the two species. As previously described, the two species differ in 45 nucleotides and 4 insertions/deletions in their ribosomal Internal Transcribed Spacer (ITS) region (Chen et al., 2018). Using the average substitution rate for ITS for herbaceous plants (Kay et al., 2006), this would result in a divergence time of c. 9.8 (±1.4SE) million years (*c.f.* Puglisi et al., 2011). This is presumably long enough for the two genomes to diverge considerably. Nevertheless, the genome-to-genome alignment analysis may need to be optimised, such as changing the parameters and comparisons with other alignment tool (Marçais et al., 2018). It is also possible that the repeat content of the genomes were not assembled as previously discussed, and reanalysis of genome-to-genome alignment with improved genome assemblies (such as inclusion of PacBio or Nanopore data) may be required.

3.4.3 Identification of contaminant species in the nuclear genome assemblies

Cross-species contamination was observed in both *Streptocarpus* genome assemblies. FastQC results for the *S. rexii* sequencing data indicated the presence of contaminants that had a high GC content (~65%) (Figure 3.2), as well as the '2 heaped' GC distribution observed in the unfiltered assemblies (Figure 3.11 c and 3.13 c). Analysis using Blobtools identified the major sources of contamination in the *S. rexii* assembly as stemming from Actinobacteria and Proteobacteria (Figure 3.12 and Table 3.8). In addition, there were several other contaminants that shared a similar GC content to that of the plant genomes, including fungi and arthropods (Figure 3.12 and 3.17). The *S. grandis* assembly showed much less cross-species contamination than that of *S. rexii* (Figure 3.17 and Table 3.12). One possibility is that the plant material used for DNA extraction were maintained under different conditions: the *S. rexii* plants were grown in a glasshouse, where the plants are exposed a wide range of organisms, including to insects, animals, fungi, bacteria and other microbes present in and on other plants kept in the glasshouses. On the other hand, the *S. grandis* material had always been kept isolated in a growth chamber under fixed environmental conditions prior to DNA extraction, and was thus exposed to fewer contaminants. Interestingly, the Proteobacterial scaffolds identified in *S. rexii* and *S. grandis* have different GC percentage and may have came from different species. In *S. rexii* the scaffolds have around 60% to 70% GC percentage (Figure 3.12), while in *S. grandis* they have approximately 40% to 50% (Figure 3.17). Further examination revealed that the Proteobacteria in *S. rexii* is mainly *Methylobacterium extorquens*, with ~65% GC percentage, whereas in *S. grandis* they are mainly an unidentified *Methylophilus* sp. TWE2, with ~45% GC percentage.

The list of contaminants identified could represent a potential resource for horticultural pest control purposes (Appendix 3.3 and 3.4). In both *Streptocarpus* assemblies, many plant pathogenic microbes were identified. In terms of bacteria, the list includes *Janthinobacterium agaricidamnosum* which causes soft rot disease in common mushrooms (Lincoln et al., 1999); *Pectobacterium polaris* causes soft-rot disease in potato (Dees et al., 2017); *Pseudomonas cichorii* a non-host specific pathogen causing water-soaked lesions on leaves (Li et al., 2014); *Acidovorax citrulli* causing fruit blotch in melons (Eckshtain-Levi et al., 2016); *Serratia marcescens* the causative agent for yellow vine disease in melons

(Rascoe et al., 2003), and the *Xanthomonas* sp. that produce spots and blights on different plant organs of a wide range of hosts (Da Silva et al., 2002). As for the potential fungal pathogen identified, *Peronospora tabacina* is known for causing the blue mold disease in tobacco (Ristaino et al., 2007), and other Peronospora sp. are known to cause Downy mildew (Slusarenko and Schlaich, 2003). *Pythium ultimum* causes rot disease on a wide range of crop and ornamental plants (Lévesque et al., 2010). The Dahlia mosaic virus causes mosaic patterns on leaves (Brunt, 1971). Finally, the *Aphelenchoides fragariae* is a parasitic nematode found in strawberries and several ornamental plants (Sánchez-Monge et al., 2015). Besides the pathogenic organisms, sequences of plant growth stimulating-microbes and possible symbionts were also found in the assemblies. For example, *Methylobacterium* sp., *Azorhizobium* sp., *Rhizobium* sp., *Sinorhizobium* sp., *Bradyrhizobium* sp., and *Mesorhizobium* sp. that are known symbionts of legumes involved in nitrogen fixation (Masson-Boivin et al., 2009).

However, it should be noted that the accuracy of the identification is limited to the availability of reference genomes, i.e. if the reference genome of the actual species does not exist in the NCBI nucleotide database, the sequence will be identified as the most closely related species. A problem with the current contamination-filtering strategy is that many assembled scaffolds remained unidentified or undefined (no-hit). This may possibly be improved by BLAST searching of the genome assemblies against other available genomic databases, such as UniProt (Apweiler et al., 2004). The BLAST search results of multiple databases can be integrated using the 'bestsumorder' option in Blobtools and may increase the proportion of taxonomical rank-assigned scaffolds (Laetsch and Blaxter, 2017). Another possible improvement for the overall genome assembly strategy, is to remove the contaminant reads instead of the scaffolds, and repeat the genome assembly with the cleaned reads which may increase the assembly contiguity (Koutsovoulos et al., 2016).

3.4.4 Comparisons with other closely related genomes and possible future improvement for genome assembly

When comparing the metrics of the *Streptocarpus* genome assemblies to those of the closely related *D. hygrometricum* and to *E. guttata* (Table 3.17), the former assemblies were found to be much more fragmented. The longest scaffold was less than half the length of those in *Dorcoceras* and *Erythrante*, and the N50 value was about one third of the other two assemblies. This was likely due to the limitation of the data availability in our project. Differences in the library construction and sequencing strategy of the other two genomes may provide us with guidance for our future sequencing experiments. For the *Dorcoceras* assembly, both short-insert paired-end library and long-insert mate-pair libraries for Illumina sequencing were used, as well as 454 Pyrosequencing to produce data with longer read length of up to 1 Kbp (Xiao et al., 2015). For the *Erythrante* genome, even though next generation sequencing technologies were not used, libraries of long insert size (3.3 Kbp insert size to 105 Kbp insert size) were used (Hellsten et al., 2013). In both cases, long insert size libraries were used to improve the scaffold assembly (Ekblom and Wolf, 2014). Thus, in addition to the usage of third generation sequencing data to improve our genome assembly, long insert size libraries such as mate-pair might be another option.

In terms of sequence similarity, genome-to-genome alignments between the *Streptocarpus* sp. and the *D. hygrometricum* indicated that the two *Streptocarpus* genome assemblies were very different from the *D. hygrometricum* genome. Especially with a large proportion of the *D. hygrometricum* genome that cannot be found in the *Streptocarpus* assemblies (Figure 3.20 b and c), though again the alignment procedure may also require optimisation.

Table 3.17 Comparison of the *Streptocarpus* assemblies with other Lamiales genomes

	Streptocarpus rexii	***Streptocarpus grandis***	***Dorcoceras hygrometricum***	***Erythrante guttata***
Sequencing approach	350 bp and 550 bp insert libraries Sequenced on HiSeq X/2500	350 bp insert library Sequenced on HiSeq X	170 – 800 bp insert library Sequenced on HiSeq 2000 600 and 1,000 bp shotgun libraries Sequenced by 454 GF FLX	Three 3.3 Kbp insert libraries Two 6.6 Kbp insert libraries One 7.9 Kbp insert library Six 33 – 36 Kbp insert fosmid libraries Two BAC 64 – 105 Kbp insert libraries Sanger sequencing
Assembly metrics				
Total span (Mbp)	596	843	1,548	322
Total no. scaffolds	95,845	127,951	520,969	2,212
Largest scaffold (bp)	421,987	738,916	1,434,191	4,921,564
N50 (bp)	35,609	31,638	110,988	1,123,783
GC (%)	37.75	38.31	42.30	36.31

To summarise this chapter, here we generated the first draft genome assemblies for the genus *Streptocarpus*. These represent only the second and third Gesneriaceae genome to be sequenced and assembled alongside that of *D. hygrometricum*. The assemblies will serve as an important reference resource for the analysis of the RNA-Seq and RAD-Seq data in the following chapters. The organellar genomes could be assembled and circularised as a by-product, making them invaluable resources for future studies.

Chapter 4 Building transcriptome resources – RNA sequencing transcriptome analysis of *Streptocarpus rexii* and *Streptocarpus grandis*

4.1 Introduction

4.1.1 RNA sequencing and its applications

RNA sequencing (RNA-Seq) uses NGS technologies to sequence and quantify transcripts (Wang et al., 2009). Compared to traditional transcriptomic study approaches, such as microarrays or genomic tiling arrays, RNA-Seq does not require a reference genome or design of hybridisation probes, thus is suitable for studies in non-model organisms (Wang et al., 2009; Korpelainen et al., 2014). RNA-Seq reads can be assembled into a transcriptome providing transcribed mRNA sequence information, and can potentially be used for identifying differentially expressed genes or investigating isoforms and alternative splicing events (Korpelainen et al., 2014). For example, RNA-Seq derived transcriptomes from a developmental series have been used for comparative transcriptomics to identify candidate regulators of phyllotaxy in *Antirrhinum* (Wang et al., 2017a) and shoot apical meristem and floral development in legumes (Singh and Jain, 2014).

A *Streptocarpus rexii* transcriptome database has previously been generated to identify candidate gene sequences for hormone biosynthesis and cotyledonary development (Chiara et al., 2013; Chen et al., 2017). However, so far no transcriptome resource is available for the comparison between rosulate and unifoliate growth forms in the genus. The transcriptomic data of *S. rexii* and *S. grandis* would thus be invaluable in this aspect, as it would be the first transcriptome of a unifoliate species. In this chapter the transcriptome of both *S. rexii* and *S. grandis* are generated using RNA-seq performed on pooled RNA from different tissue types (vegetative+reproductive).

4.1.2 Currently available *Streptocarpus* and Gesneriaceae transcriptomes

For twenty-three Gesneriaceae species across 6 genera RNA-Seq derived transcriptomes are published (Figure 4.1), including *S. rexii* and *Streptocarpus ionanthus* (formerly *Saintpaulia ionantha*; see also Nishii et al., 2015) (Table 4.1). However, the RNA used for both these *Streptocarpus* transcriptomes were extracted soley from vegetative tissues (Chiara et al., 2013; Matasci et al., 2014). Other Gesneriaceae transcriptomes include *Damrongia clarkeana* (as *Boea clarkeana*, see also Puglisi et al., 2016) and *Dorcoceras hygrometricum* (as *Boea hygrometrica*, see also Puglisi et al., 2016), were used for studying drought tolerance and rehydration processes (Xiao et al., 2015; Wang et al., 2017b). Transcriptomes of *Achimenes* were used for the study of floral development (Roberts and

Roalson, 2017). *Sinningia eumorpha*, *Sinningia magnifica*, and *Primulina* transcriptomes were derived from a mixture of tissue types, e.g. leaf and flower or leaf and root (Ai et al., 2014; Serrano-Serrano et al., 2017). The *Primulina* transcriptomes were used for studying species and population genetic diversity (Ai et al., 2014).

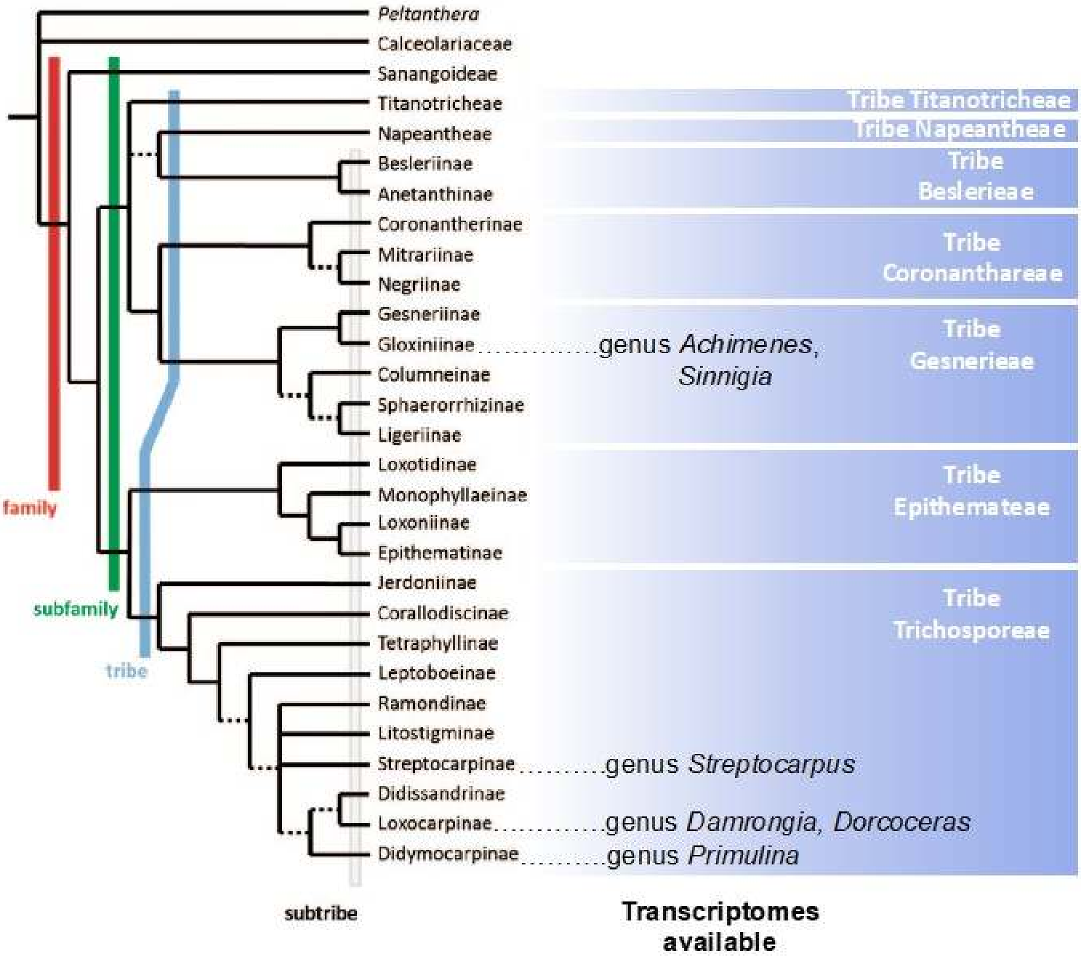


Figure 4.1 Summary phylogeny of the Gesneriaceae family with indication of the genera for which transcriptome resources are available (Figure modified from Weber et al., 2013)

Table 4.1 List of available Gesneriaceae transcriptomes

Species	Tissues	Reference
Achimenes cettoana	Flower	Roberts and Roalson, 2017
Achimenes erecta	Flower	Roberts and Roalson, 2017
Achimenes misera	Flower	Roberts and Roalson, 2017
Achimenes patens	Flower	Roberts and Roalson, 2017
Damrongia clarkeana	Leaf	Wang et al., 2017
Dorcoceras hygrometricum	Leaf	Xiao et al., 2015
Sinningia tuberosa	Leaf	Matasci et al., 2014
Sinningia eumorpha	Leaf and flower	Serrano-Serrano et al., 2017
Sinningia magnifica	Leaf and flower	Serrano-Serrano et al., 2017
Streptocarpus rexii	Leaf and cotyledon	Chiara et al., 2013
Streptocarpus ionantha	Leaf	Matasci et al., 2014
Primulina eburnea	Leaf and root	Ai et al., 2014
Primulina fimbrisepala	Leaf and root	Ai et al., 2014
Primulina heterotricha	Leaf and root	Ai et al., 2014
Primulina huaijiensis	Leaf and root	Ai et al., 2014
Primulina lobulata	Leaf and root	Ai et al., 2014
Primulina lutea	Leaf and root	Ai et al., 2014
Primulina pteropoda	Leaf and root	Ai et al., 2014
Primulina sinensis	Leaf and root	Ai et al., 2014
Primulina swinglei	Leaf and root	Ai et al., 2014
Primulina tabacum	Leaf and root	Ai et al., 2014
Primulina villosissima	Leaf and root	Ai et al., 2014

4.1.3 RNA-seq analysis strategy

In this chapter, analysis of the RNA-seq data involves four steps: read preprocessing (i.e., trimming), assemblying, isolation of open reading frames of genes (ORFs), and fuctional annotation. Read preprocessing step focuses on removing low quality reads (to remove potential sequencing error) and Illumina adaptor sequences, which if not filtered can causes misassembly (Andrews, 2010). The assembly step is done through *de novo* assembly and reference guided assembly: *de novo* transcriptome assembly is performed in a similar fashion to that of *de novo* genome assembly, which RNA-seq reads are break down into *K*-mers and transcripts are reconstructed using de Bruijn graph approach (see section 3.1.4; Haas et al., 2013); reference guided assembly is carried out in the presence of a reference genome sequence. First RNA-seq reads are aligned to the genome sequence, and transcripts can be reconstructed from overlapped-mapped reads (Korpelainen et al., 2014). An essential consideration for reference guided assembly is the aligner's capability in creating gaps during alignment process. As the reference genome sequence usually contains intron-exon gene structures, which is not presented in the RNA-seq reads, the aligners must be able to create gaps at the intron position, thus separates a read into 2 or potentially more exons regions in order to achieve correct read mapping (Korpelainen et al., 2014). Software such as HISAT2 has such function and is designed for aligning RNA-seq data to a reference genome (Kim et al., 2015). Assemblers such as Trinity (Haas et al., 2013) is designed for both *de novo* and reference-guided assembly.

ORF identification step aims to isolate translated proportion of the assembled transcripts, which the isolated sequences (as amino acids) can later be used for protein functional annotation. The software TransDecoder is designed for this purpose and can effectively remove untranslated regions (UTRs; Haas et al., 2013). Finally, the functional annotation step aims to assign (annotate) a possible biological function to the protein sequences by comparing them to existing protein sequence database (Bolger et al., 2018). For example, the webtool Mercator assign 'MapMan Bin ontology' to assembled sequences, a plant specific collection of protein functions catagorised by biological processes (Loshe et al., 2014); the Kyoto Encyclopedia of Genes and Genomes ontology, KEGG, is a database of genes and related biological pathways, and can be used to assign 'KO terms' to the assembled proteins (Kanehisa et al., 2016).

As the scope of this study is to generate reference gene transcripts dataset for *S. rexii* and *S. grandis*. For both species, total RNA was extracted from seedlings, roots, shoots, floral buds, flowers and developing fruits to cover as wide an expression profile as possible. Since the intention was not to compare gene expression differences between different tissue types, no biological repeat was made for gene differential expression analysis, and all RNA were pooled and sequenced together. The RNA-Seq reads were assembled *de novo* and through a reference-guided approach (i.e. mapping the RNA-Seq reads to a reference genome). The assemblies were filtered for ORFs and cross-species contaminants. The resulting transcriptomes were compared to those of other Gesneriaceae and to model species and will serve as a fundamental genomic resource for the genus *Streptocarpus*.

4.2 Materials and methods

4.2.1 Plant materials

All plant materials were harvested from plants grown in the research glasshouses and growth chambers at the Royal Botanic Garden Edinburgh. For *S. rexii*, the accession 20150819 was used, and for *S. grandis* accession 20020577. The sample collection was performed as described in Chapter 2, section 2.2.1. In brief, the seedlings, roots, shoots, floral buds, flowers, and developing fruit tissues of *S. rexii* and *S. grandis* were collected from young and mature plants. The RNA extraction and the sequencing of *S. grandis* materials was done and the data kindly provided by K. Nishii.

4.2.2 RNA extraction, library preparation and RNA-Seq

The RNA extraction and quality assessment of the RNA samples were described in Chapter 2. In brief, RNA were extracted using TRIzol reagents followed by acid phenol:chloroform purification and PureLink Kit clean up. The extracted total RNA were pooled and delivered to Edinburgh Genomics (University of Edinburgh, Edinburgh, UK) for library preparation and sequencing. The library was prepared using TruSeq Stranded mRNA Prep Kit (Illumina, San Diego, CA, USA). The *S. rexii* library was first sequenced in one lane of MiSeq (Illumina) together with two other libraries. However, the read yield was not sufficient to generate a good transcriptome. Thus, an additional sequencing run was carried out, which was performed in one lane of HiSeq 4000 (Illumina) together with 13 other libraries outside the scope of this thesis. For the *S. grandis*, the sequencing was performed using one lane of MiSeq (Illumina). The sequencing results was returned in Fastq format and the read quality was accessed by FastQC v0.11.5 (Andrews, 2010).

4.2.3 Sequence data pre-processing

Trimmomatic-v0.35 (Bolger et al., 2014) was used for adopter and quality trimming with the following settings: ILLUMINACLIP:TruSeq3-PE2.fa:2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:50. The pre-processed reads were stored in fastq.gz format (Box 4.1).

Box 4.1 Script for Trimmomatic preprocessing of the RNAseq reads. Text in bold with brackets **[Text]** indicate the input files and output file names to be specified.

```
java –jar trimmomatic-0.35.jar PE -threads [NO_CPU] –phred33 \
    [READ1fastq] [READ2.fastq] \
    [TRIMMED_READ1.fastq] [UNPAIR_TRIMMED_READ1.fastq] \
    [TRIMMED_READ2.fastq] [UNPAIR_TRIMMED_READ2.fastq] \
    ILLUMINACLIP:/TruSeq3-PE-2.fa:2:30:10 \
    LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:50
```

4.2.4 Transcriptome *de novo* and reference-guided assembly

Trinity v2.4.0 (Haas et al., 2013) was used for both *de novo* and reference-guided assembly of the transcriptomes. For *de novo* assembly, the preprocessed reads were used and the assembly carried out under default parameter settings (Box 4.2).

For the reference-guided assembly, the software STAR (Dobin et al., 2013) was used to map the RNA-Seq reads onto the corresponding genome assemblies (i.e. *S. rexii* RNA-Seq reads mapped to *S. rexii* genome, and *S. grandis* reads to *S. grandis* genome). The assembly was carried out in two stages. First, mapping and assembly was performed under default parameter settings. Second, three STAR mapping parameters were tested to improve the mapping percentage; this includes the 'minimum and maximum intron size', 'maximum mismatch allowed during alignment' and the '2nd pass mode'. The detail parameter values tested for the assembly optimisation are summarised in Table 4.2. Optimal parameter values were chosen to generate a bam file of the assembly to the genome, which was then used for transcriptome assembly via Trinity v2.4.0 (Haas et al., 2013). The detailed commands used are listed in Box 4.2 (*de novo* assembly) and Box 4.3 (reference-guided assembly).

Table 4.2 STAR parameters tested for the mapping of RNA-Seq reads to reference genome

Parameter	Parameter definition	Values tested
--alignIntronMin / --alignIntronMax	The minimum / maximum intron length allowed when mapping RNA-Seq reads	21 / 288 (default), 10 / 10,000, 10 / 20,000
--outFilterMismatchNmax	The maximum number of mismatch allowed when writing a read to the output file (i.e. the mapping procedure was not affected, but by altering this parameter one can decide whether to consider a reads with more mismatches to be considered as mapped or not)	5, 10 (default), 20, 30, 40, 50, 60, 70, 80
--twopassMode	2nd pass mode. Enabled to improve the mapping of reads on splice junctions. When enabled, STAR will carry out a normal mapping procedure first (1st mapping) and identify the splice junctions. The splicing information was then used for the 2nd mapping, which will not identify new junction but will align the spliced reads with short overhangs across the previously detected junctions	Disabled (default), Enabled

Box 4.2 Script for transcriptome *de novo* assembly. Text in bold with brackets **[Text]** indicate the input files and output file names to be specified.

```
## For Trinity de novo assembly
Trinity --seqType fq --CPU [NO_CPU] \
    --left [READ1.fastq] --right [READ2.fastq]
```

Box 4.3 Script for transcriptome reference-guided assembly. Text in bold with brackets **[Text]** indicate the input files and output file names to be specified.

```
# Build STAR genome indices
STAR --runThreadN [NO_THREADS] --runMode genomeGenerate \
    --genomeDir [OUTPUT_DIRECTORY] \
    --genomeFastaFiles [GENOME_ASSEMBLY.fasta] \
    --limitGenomeGenerateRAM 300000000000 \
    --genomeSAindexNbases 13

# STAR mapping of RNA-Seq reads
STAR --runThreadN [NO_THREADS] \
    --genomeDir [GENOME_INDEX_DIRECTORY] \
    --readFilesIn [READ1.fq] [READ2.fq] \
    --outFilterMismatchNmax [MISMATCH_ALLOWED]
    --outFileNamePrefix [OUTPUT_NAME]

# Convert the STAR output sam file into bam format
samtools view -Sb -@ [NO_THREADS] [STAR_OUTPUT.sam] | \
    samtools sort -O bam -@ [NO_THREADS] -o [OUTPUT_NAME.bam]

# Trinity genome-guided assembly
Trinity --seqType fq --left [READ1.fq.gz] --right [READ2.fq.gz] \
    --genome_guided_bam [OUTPUT_NAME.bam] \
    --genome_guided_max_intron 20000 \
    --max_memory 200G –CPU [NO_THREADS]
```

4.2.5 Preliminary assembly quality evaluation

For both *de novo* and reference-guided assemblies, tools from the Trinity package v2.4.0 (Haas et al. 2013) were used to evaluate the metrics of the assemblies (1) the 'analyze_blastPlus_topHit_coverage.pl' was used to calculate the gene completeness. (2) 'TrinityStats.pl' was used to calculate the basic metrics, e.g. N50, total base pairs assembled, and number of predicted genes. In addition, Bowtie2 v2.2.8 (Langmead and Salzberg, 2012) was used to check the remapping rate of RNA-seq reads on the assembly under default settings. BUSCO v3 (Simão et al., 2015) was used to check the completeness of universal single copy orthologous genes. This involved BLASTing the assembled transcripts against the Embryophyta odb9 database (Last update date 13/02/2017) and assessing transcriptome completeness by the proportion of BUSCO genes found. The detailed commands are listed in Box 4.4.

Box 4.4 Script for preliminary quality assessment of the transcriptome assembly. Text in bold with brackets **[Text]** indicate the input files and output file names to be specified.

```
## 1. Calculating gene completeness using Trinity script
# Build the blastdb from uniprot dataset fasta file
makeblastdb -in [uniprot_sprot.fasta] -dbtype prot
# Blast the transcriptome against the database
blastx -query [TRANSCRIPTOME.fasta] -db [uniprot_sprot_DB] \
    -out [BLAST_OUTPUT_NAME] -evalue 1e-20 \
    -max_target_seqs 1 -outfmt 6
# Analyse the output file
Trinity/util/analyze_blastPlus_topHit_coverage.pl \
    [BLAST_OUTPUT] \
    [TRANSCRIPTOME.fasta] \
    [uniprot_sprot.fasta]
# Group BLAST hits to improve the coverage
Trinity/util/misc/blast_outfmt6_group_segments.pl \
    [BLAST_OUTPUT] [TRANSCRIPTOME.fasta] \
    [uniprot_sprot.fasta] > [GROUPED_BLAST_OUTPUT]
# Re-analyse the output file
Trinity/util/misc/ blast_outfmt6_group_segments.tophit_coverage.pl \
    [GROUPED_BLAST_OUTPUT] \
    [TRANSCRIPTOME.fasta] \
    [uniprot_sprot.fasta]

## 2. Basic assembly metrics calculation using Trinity script
Trinity/util/TrinityStats.pl [TRANSCRIPTOME.fasta]

## 3. Checking remapping rate
bowtie2-build [TRANSCRIPTOME.fasta] [OUTPUT_INDEX_NAME]
bowtie2 -q -x [BOWTIE2_INDEX] -1 [READ1.fastq] -2 [READ2.fastq]

## 4. Checking gene completeness using BUSCO
python BUSCO.py -i [TRANSCRIPTOME.fasta] -o [OUTPUT_NAME] -m tran \
    --cpu [NO_CORES] -l embryophyta_odb9/
```

4.2.6 Post-assembly filtering and functional annotation

For both *de novo* and reference-guided approaches, the assemblies were filtered for sequences containing open reading frames (ORFs) via TransDecoder v5.1.0 (Haas et al., 2013) with the '--single_best_orf' option, which keeps the longest isoform among all the identified isoforms. The identified ORFs were then filtered for potential contaminant sequences using the KEGG Orthology And Link Annotation (KOALA) function on the KEGG server (Kyoto Encyclopedia of Genes and Genomes). The tool GhostKOALA v2.0 (Kanehisa et al., 2016) assigned taxonomical information to all the ORFs, and those ORFs labelled non-eudicot or non-monocot-originated were subsequently removed (i.e. assigned as basal plants, animals, fungi, protists, bacteria, archaea, viruses, or unknown). The assembly procedure used in this study is and filtering commands are listed in Box 4.5.

Box 4.5 CDS identification and contaminant removal for the transcriptome. Text in bold with brackets **`[Text]`** indicate the input files and output file names to be specified.

```
## Using TransDecoder for CDS identification
./TransDecoder.LongOrfs -t [TRANSCRIPTOME.fasta]
./TransDecoder.Predict -t [TRANSCRIPTOME.fasta] --single_best_only

## Transcriptome contaminant filtering based on GhostKOALA results
# Identify the angiosperm originated transcripts
grep -i "monocot\|eudicot" [KOALA.TAXANOMY.OUTPUT] | \
    awk '{print $1}' | sed 's/user://' > transcript.to.keep
# Isolate the corresponding transcripts
perl -ne 'if(/^>(\S+)/){$c=$i{$1}}$c?print:chomp;$i{$_}=1 if @ARGV' \
    transcript.to.keep [TRANSDECODER_PROCESSED_CDS.fasta] >
    filtered.fasta
```

The basic statistics of the assemblies and BUSCO completeness were evaluated as described in Box 4.4, and their functional annotation carried out using KEGG GhostKOALA v2.0 (Kanehisa et al., 2016a; 2016b) and the Mercator web tool (Lohse et al., 2013) under default settings. The transcriptome analysis flowchart is summarised in Figure 4.2.

4.2.7 Orthogroup identification

Orthofinder v1.1.8 (Emms and Kelly, 2015) was used to identify conserved orthogroups in the four final assemblies (i.e. *de novo* and reference-guided *S. rexii* transcriptomes; *de novo* and reference-guided *S. grandis* transcriptomes). The TransDecoder output file, which consisted of the peptide sequences of the identified ORFs, was used as input. An OrthoFinder analysis was performed under default settings (Box 4.6). The output file 'Orthogroups.csv' was used for the visualisation as a Venn diagram using an online tool (http://bioinformatics.psb.ugent.be/webtools/Venn/).

Box 4.6 Identification of orthologous transcripts using OrthoFinder. Text in bold with brackets **`[Text]`** indicate the input files and output file names to be specified.

```
# Execute OrthoFinder, where all transcriptome fasta files are in the
# same directory
Orthofinder -f [FASTA_FILE_DIRECTORY] -t [NO._CPU]
```

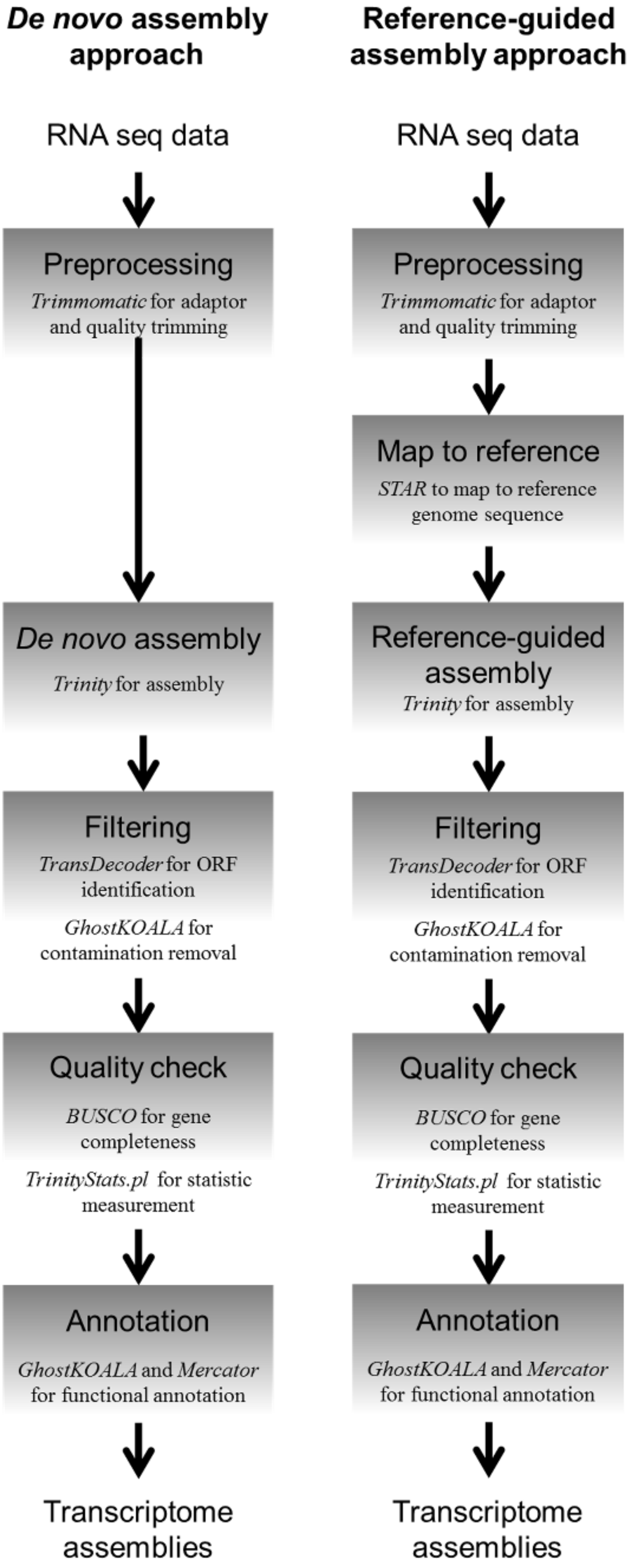

De novo assembly approach
Reference-guided assembly approach
RNA seq data
RNA seq data
Preprocessing
Trimmomatic for adaptor and quality trimming
Preprocessing
Trimmomatic for adaptor and quality trimming
Map to reference
STAR to map to reference genome sequence
De novo assembly
Trinity for assembly
Reference-guided assembly
Trinity for assembly
Filtering
TransDecoder for ORF identification
GhostKOALA for contamination removal
Filtering
TransDecoder for ORF identification
GhostKOALA for contamination removal
Quality check
BUSCO for gene completeness
TrinityStats.pl for statistic measurement
Quality check
BUSCO for gene completeness
TrinityStats.pl for statistic measurement
Annotation
GhostKOALA and *Mercator* for functional annotation
Annotation
GhostKOALA and *Mercator* for functional annotation
Transcriptome assemblies
Transcriptome assemblies

Figure 4.2 Flowchart of *Streptocarpus* transcriptome analysis

4.3 Results

4.3.1 RNA-Seq data preprocessing

The MiSeq run of *S. rexii* generated about 4.3 million read pairs; the HiSeq 4000 run of *S. rexii* generated the highest read counts among all three sequencing experiments with approximately 24.7 million read pairs. The MiSeq *S. grandis* run gave about 16.5 million read pairs (Table 4.3).

Prior to preprocessing, all three datasets showed good sequence quality with an average quality score above Q30 (Figure 4.3 to 4.5 a). Some biases in ‘per base sequenced content’ were observed, where the proportion of C bases did not corresponded to the proportion of G bases in the data, neither did the proportion of A to T bases (Figure 4.3 to 4.5 b and c). The ‘GC distribution’ graphs failed to fit a normal distribution, and showed skewed pattern towards the 55% to 65% GC content (Figure 4.3 to 4.5 d). Adopter contamination was found at the 3’ end of the raw reads (Figure 4.3 to 4.5 e).

After preprocessing, only about half of the sequencing data remained for each dataset: this was approximately 2.3, 12.7, and 10.2 million read pairs for the *S. rexii* MiSeq, *S. rexii* HiSeq 4000, and *S. grandis* MiSeq sequencing datasets respectively (Table 4.3). The biased ‘per base sequence content’ and the skewed GC distributions persisted after preprocessing (Figure 4.3 to 4.5 b, c, d).

Table 4.3 *S. rexii* and *S. grandis* RNA-Seq experiment summary

	S. rexii	***S. rexii***	***S. grandis***
Sequencer	MiSeq	HiSeq 4000	MiSeq
No. read pairs obtained	4,364,215	24,739,455	16,501,920
Total base pairs (bp)	1,309,264,000	7,313,836,500	4,950,576,000
No. read pairs after trimming	2,357,477	12,725,705	10,266,774
Total base pairs after trimming (bp)	703,323,621	3,773,167,877	3,064,274,789

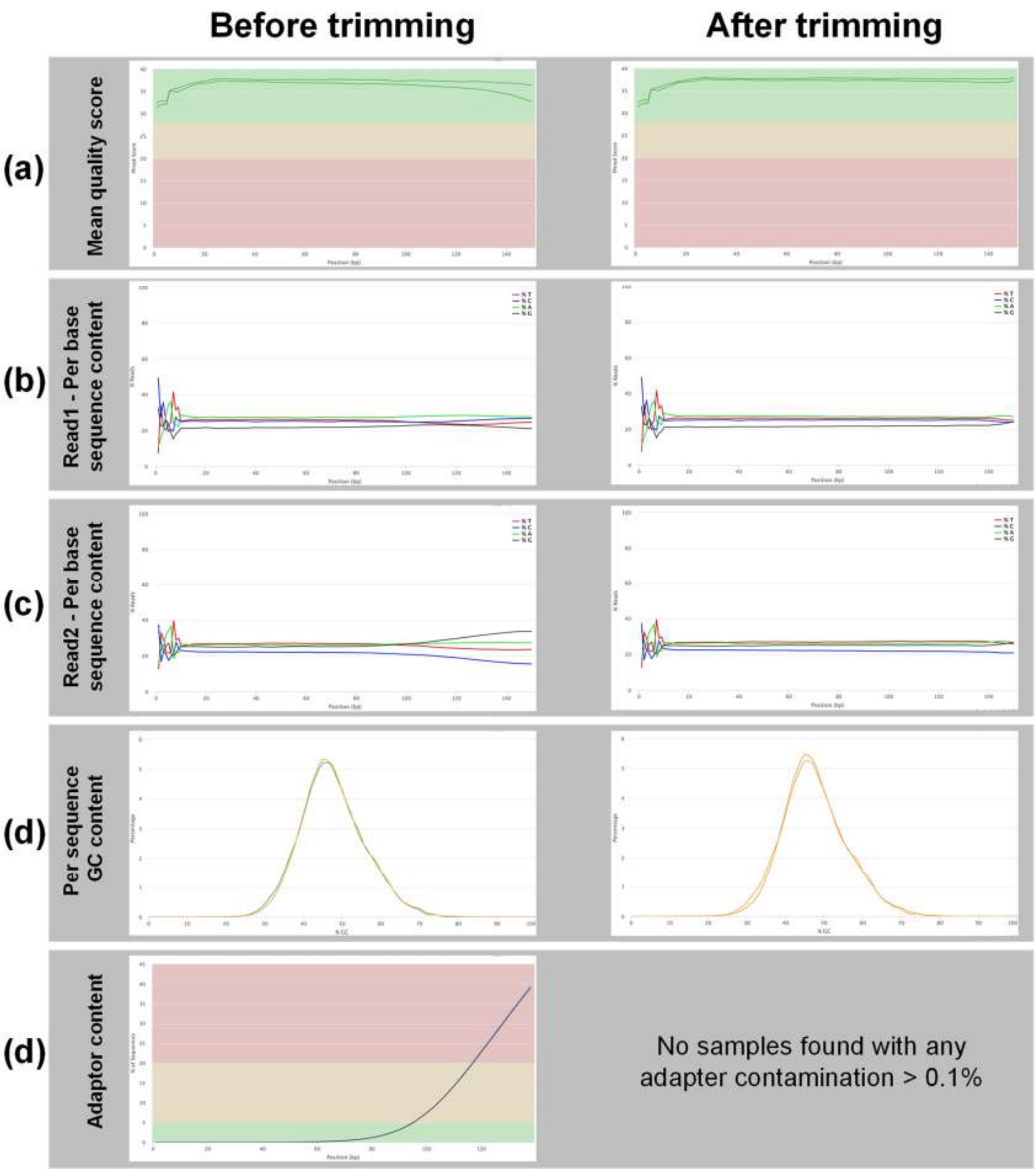


Figure 4.3 RNA-Seq read quality check results of the *S. rexii* MiSeq experiment

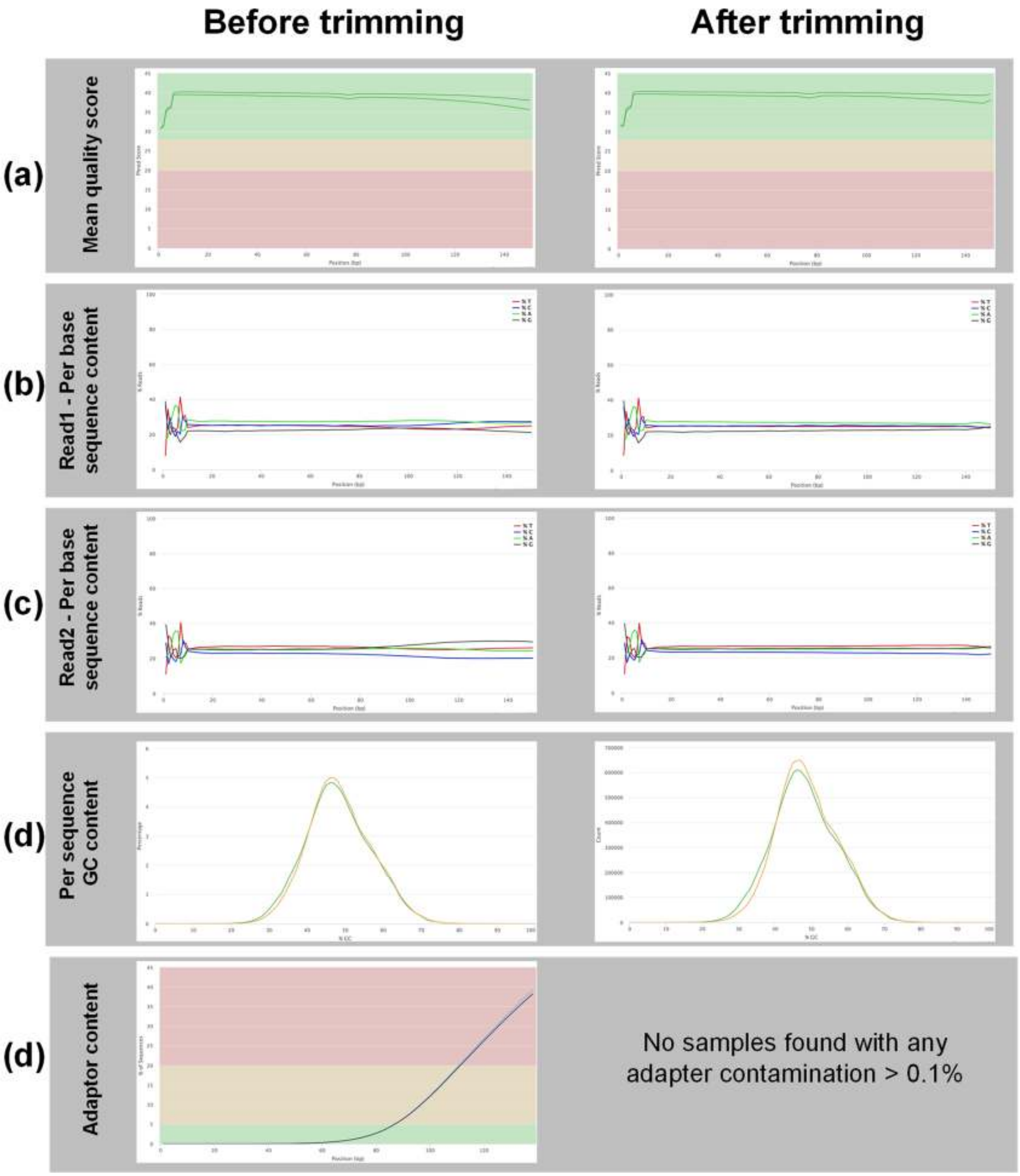


Figure 4.4 RNA-Seq read quality check results of the *S. rexii* HiSeq 4000 experiment

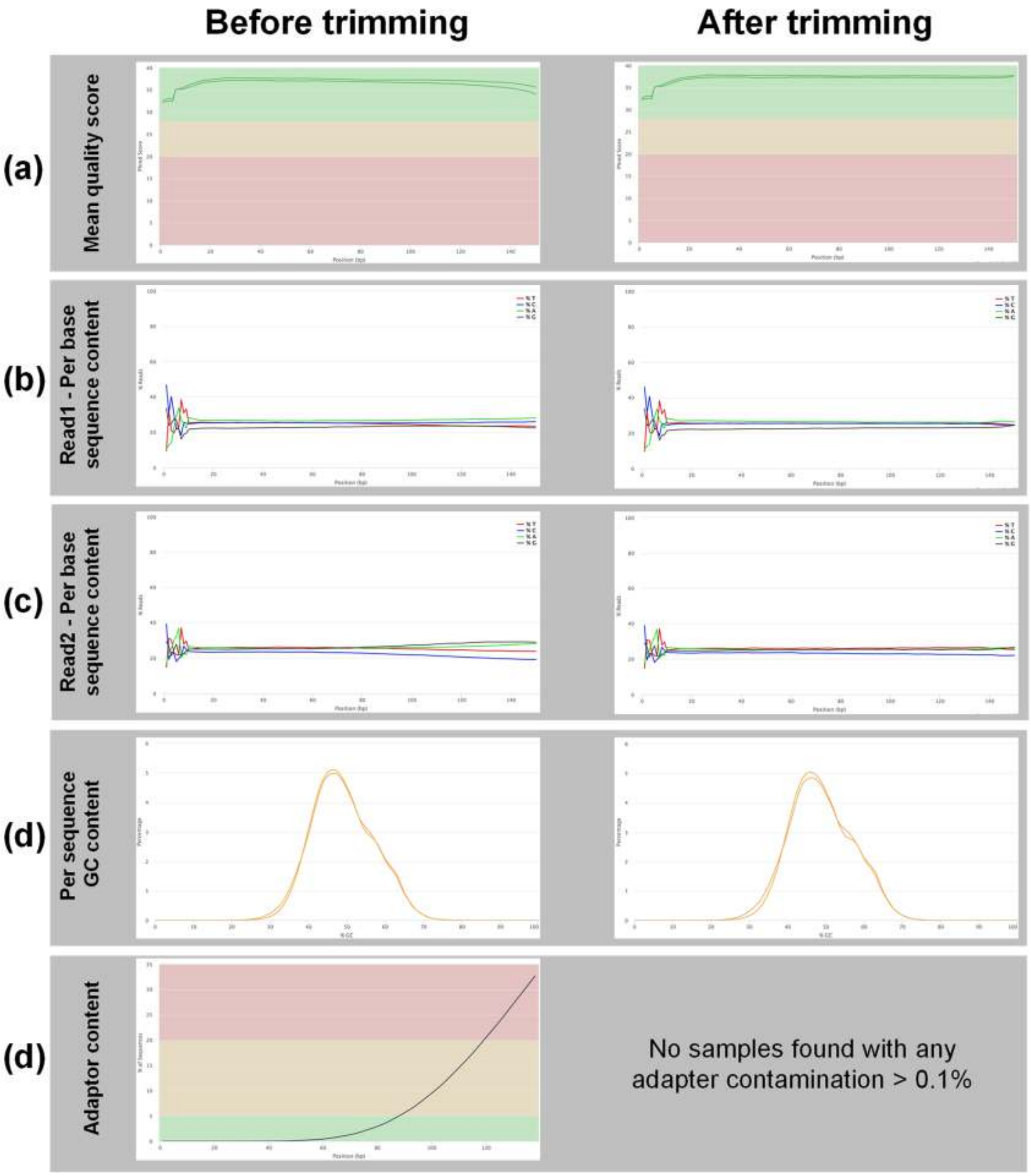


Figure 4.5 RNA-Seq read quality check results of the *S. grandis* MiSeq experiment

4.3.2 *De novo* assembly of the *S. rexii* and *S. grandis* transcriptomes

The *de novo* assembling of the *S. rexii* data using either MiSeq data alone gave the lowest number of reads, the HiSeq 4000 almost 6x more data, while the combined MiSeq + HiSeq 4000 data had the additive read numbers of the first two experiments (Table 4.4). The assembly metrics of total number of contigs and contig N50 were positively correlated with the number of input reads (Table 4.4; Figure 4.6). The analysis with MiSeq data alone recovered the lowest number of contigs (59,042), with an N50 value of 1,098 bp. Assembly using HiSeq 4000 data alone gave 101,640 contigs, and a longer N50 value of 1,580 bp. Assembly using MiSeq + HiSeq 4000 data combined generated the highest number of contigs and the highest contig N50 values (Table 4.4), and a total of 123,213,415 bp assembled with a BUSCO completeness of 79% (1,137 BUSCOs found). From the assembly including both MiSeq and HiSeq 4000 data, 64,516 ORFs were identified, with 4,016 of these labelled as possible cross-species contamination. After removing these ORFs, 60,500 ORFs remained (Table 4.4).

For the *S. grandis* MiSeq dataset, 87,665 contigs with a total of 102,299,541 bp were assembled. The contig N50 value of about 1,500 was very similar to that of the *S. rexii* analysis using HiSeq 4000 data alone, while the BUSCO completeness was with 79% identical to the *S. rexii* MiSeq + HiSeq data analysis (Table 4.4). For *S. grandis*, 53,132 ORFs were identified of which 1,855 were found to be possible cross-species contaminations and were subsequently removed. A final 51,267 ORFs were retained for the *de novo S. grandis* assembly (Table 4.4).

Table 4.4 Metrics of the *S. rexii* and *S. grandis de novo* assembled transcriptomes

	S. rexii	***S. rexii***	***S. rexii***	***S. grandis***
Dataset used	MiSeq	HiSeq 4000	MiSeq + HiSeq 4000	MiSeq
No. read pairs used	2,357,477	12,725,705	15,083,182	10,266,774
Assembly metrics				
No. contigs	59,042	101,640	110,955	87,665
Total base pairs (bp)	53,293,812	103,654,725	123,213,415	102,299,541
Average contig length (bp)	902	1,019	1,110	1,166
Contigs N50 (bp)	1,098	1,580	1,716	1,541
No. gene with > 80% length	5,724	7,777	8,343	7,798
GC (%)	42.9	42.3	42.1	42.5
Reads remapping (%)	94.9	99.2	99.3	97.1
Transcriptome completeness				
BUSCO completeness (%)	54.8	72.0	79.0	79.0
No. completed BUSCOs	789	1,036	1,137	1,138
No. missing BUSCOs	651	404	303	302
ORF identification				
No. ORFs identified	35,277	58,383	64,516	53,132
Contaminant identification				
No. contaminant contigs	763	3,928	4,016	1,855
Archaea	21	38	38	23
Bacteria	264	851	888	656
Protist	94	244	232	193
Fungi	97	483	491	204
Animal	225	2,152	2,196	458
Other plants	49	119	108	282
Virus	9	37	42	32
Unidentified	12	27	21	17
Possible contaminant (%)	2.2	6.7	6.2	3.5
Final number of contigs kept after filtering	34,506	54,432	60,500	51,267

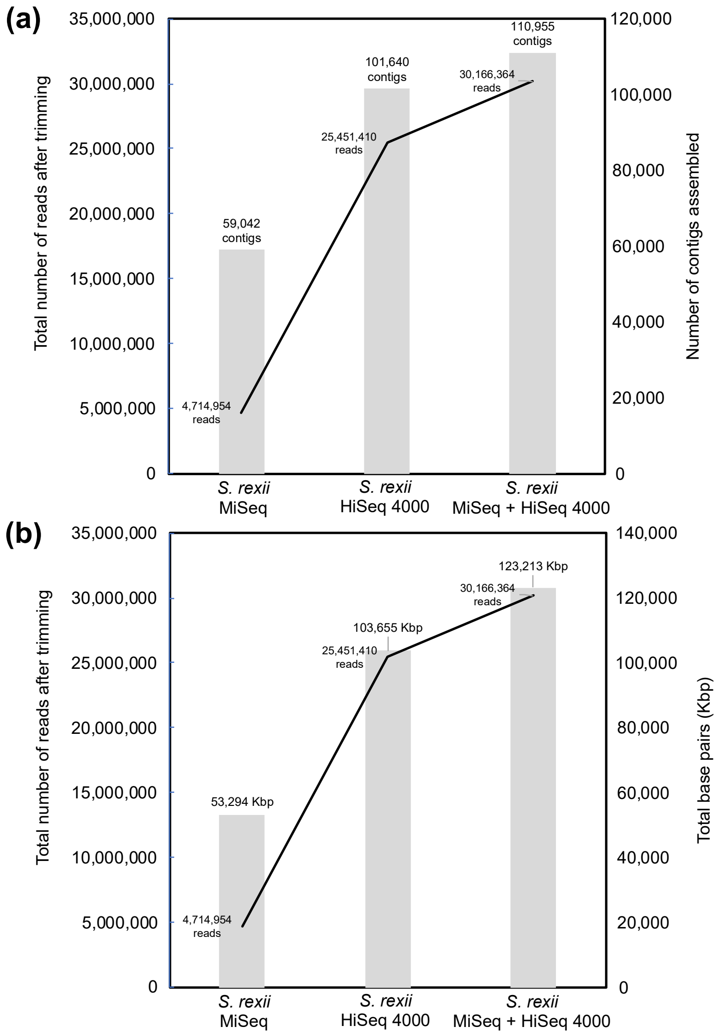


Figure 4.6 Relationships between the number of input reads and number of base pairs assembled in the *de novo* assembly of the *S. rexii* transcriptomes. (a) Number of contigs assembled. (b) Number of base pairs assembled.

The assembly metrics of the post-filtering transcriptomes were recalculated (Table 4.5). The filtered *S. rexii* transcriptome had 60,500 transcripts with a total of 64,548,015 bp assembled. The average contig length was 1,066 bp, and the contig N50 1,323 bp. The average GC% was 44.6%. The number of core genes identified decreased, with the BUSCO completeness lowered to 76.4% compared to the unfiltered transcriptome (Table 4.4).

The filtered *S. grandis* transcriptome had 51,267 transcripts and a total of 58,554,237 assembled base pairs (Table 4.5). The average contig length was 1,142 bp and the N50 1,410 bp. The average GC% was 44.3%. The transcriptome had a 79% BUSCO completeness, identical to the unfiltered transcriptome (Table 4.4).

Functional annotation was performed for the filtered *S. rexii* and *S. grandis* transcriptome assemblies (Table 4.5; Transcriptome annotation). The *S. rexii* transcriptome had 27,811 contigs annotated (45.9% of all contigs) using the KEGG pipeline, and 41,002 contigs annotated (67.7%) using the Mercator pipeline (Figure 4.7). For the *S. grandis* transcriptome, these numbers were 23,520 (45.8%) and 34,898 contigs (68.0%) by KEGG and Mercator pipelines respectively (Figure 4.8).

Table 4.5 Metrics of the *de novo* transcriptomes assembled after filtering

	S. rexii	***S. grandis***
Dataset used	MiSeq + HiSeq 4000	MiSeq
Assembly metrics		
No. transcripts	60,500	51,267
Total base pairs (bp)	64,548,015	58,554,237
Average transcript length (bp)	1,066	1,142
Longest transcript (bp)	10,947	11,934
Transcript N50 (bp)	1,323	1,410
GC (%)	44.6	44.3
Transcriptome completeness		
BUSCO completeness (%)	76.4%	79.0%
No. completed BUSCOs	1,100	1,138
No. missing BUSCOs	340	302
Transcriptome annotation		
No. gene annotated by KEGG	27,811	23,520
No. gene annotated by Mercator	41,002	34,898

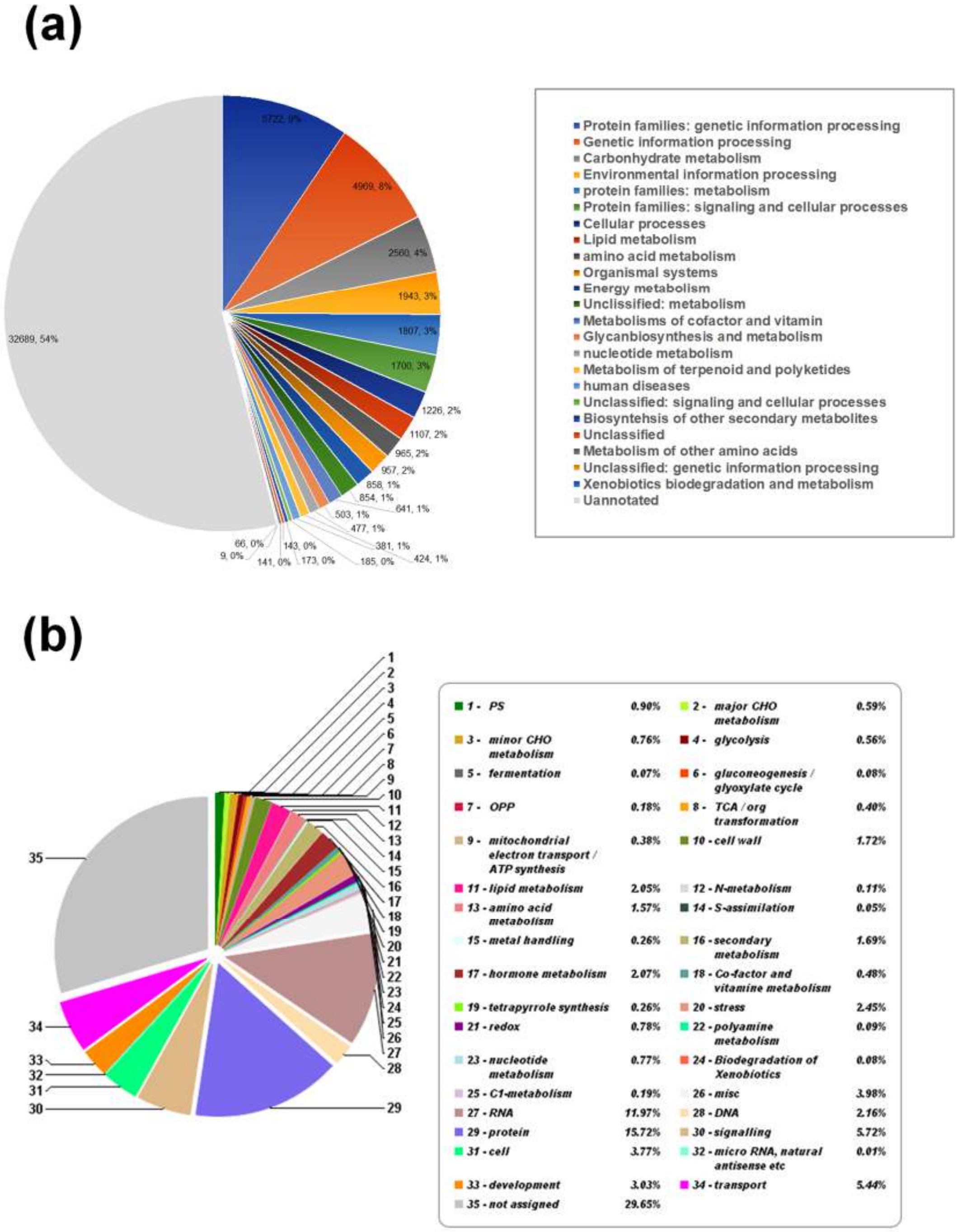


Figure 4.7 Functional annotation results of the *S. rexii de novo* transcriptome assembly (a) KEGG annotation. Data labels show the number of transcripts of the given category (right-hand side), followed by the percentage of that category in the whole transcriptome. (b) Mercator annotation. The labels show functions annotated.

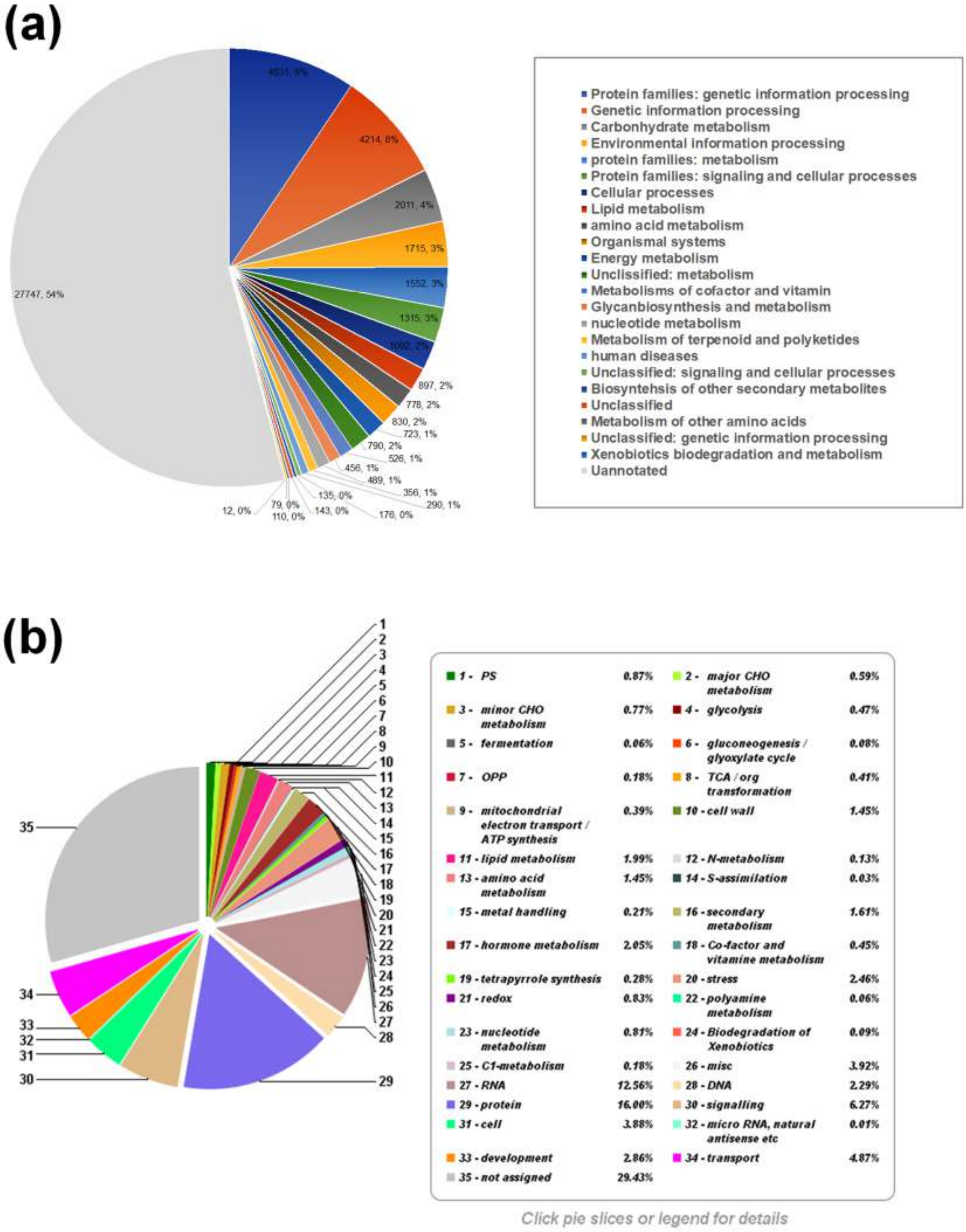


Figure 4.8 Functional annotation result of the *S. grandis de novo* transcriptome assembly (a) KEGG annotation. Data labels show the number of transcripts of the given category (right-hand side), followed by the percentage of that category in the whole transcriptome. (b) Mercator annotation. The labels show functions annotated.

4.3.3 Reference-guided transcriptome assembly of *S. rexii* and *S. grandis*

RNA-Seq reads of *S. rexii* and *S. grandis* were first mapped to the assembled draft genome under default parameters (Table 4.6). For *S. rexii*, the combined data of MiSeq and HiSeq 4000 runs were used (15,083,182 read pairs after preprocessing). For *S. grandis*, the MiSeq dataset was used (10,266,774 read pairs after preprocessing). 88.6% of the *S. rexii* reads (13,362,520 from 15,083,182 read pairs) were mapped to the *S. rexii* genome, and 93.5% *S. grandis* reads (9,600,119 from 10,266,774 read pairs) were mapped to the *S. grandis* genome. From these, 90,977 and 73,957 contigs were assembled for *S. rexii* and *S. grandis* respectively. In total, about 101 Mbp and 85 Mbp were assembled for *S. rexii* and *S. grandis* respectively (Table 4.6). The BUSCO completeness was 78.5% for *S. rexii* and 80.7% for *S. grandis*.

Table 4.6 Metrics of the reference-guided assemblies under default mapping parameters

	S. rexii	***S. grandis***
Sequencer	MiSeq + HiSeq4000	MiSeq
No. reads (after trimming)	15,083,182 × 2	10,266,774 × 2
No. reads mapped	13,362,520 × 2	9,600,119 × 2
Read mapped (%)	88.6	93.5
No. contigs	90,977	73,957
Total base pairs (bp)	101,157,269	85,843,471
Contigs N50 (bp)	1,644	1,662
GC (%)	41.8	42.2
Transcriptome completeness		
BUSCO completeness (%)	78.5	80.7
No. completed BUSCOs	1,131	1,163
No. missing BUSCOs	309	277

The STAR mapping parameters were optimised using the *S. rexii* dataset (MiSeq + HiSeq 4000 data). First, the minimum and maximum intron size was tested, with the minimum intron size decreased and maximum intron size increased (Table 4.7). The number of mapped reads increased by 1,000 to 1,500 read pairs after the adjustment, and the total numbers of contigs and base pairs assembled decreased compared to the default parameter settings (Table 4.7). On the other hand, the contig N50 and GC% remained nearly identical (Table 4.7). The BUSCO completeness was unaffected. Since no actual improvement was found in the assembly metrics, the default parameter values (minimum intron length = 21, maximum intron = 288) were kept with the optimisation moved on to the next parameter.

Table 4.7 Effect of 'intron size' mapping parameters on the assembly of *S. rexii* RNA-Seq data

	S. rexii	***S. rexii***	***S. rexii***
Parameter (Min intron length / Max intron length)	(Default) 21 / 288	10 / 10,000	10 / 20,000
No. reads mapped	13,362,520 × 2	13,364,087 × 2	13,363,238 × 2
Read mapped (%)	88.6	88.6	88.6
No. contigs	90,977	90,539	90,644
Total base pairs (bp)	101,157,269	100,960,080	101,027,579
Contigs N50 (bp)	1,644	1,645	1,645
GC (%)	41.8	41.8	41.8
Transcriptome completeness			
BUSCO completeness (%)	78.5	78.5	78.5
No. completed BUSCOs	1,131	1,131	1,131
No. missing BUSCOs	309	309	309

The number of mismatches allowed for STAR alignment was optimised. Under the default parameter (maximum 10 bp mismatches per read), 13,362,520 read pairs were mapped and 90,977 contigs assembled (Table 4.8). When the setting for mismatches allowed was decreased to 5 bp, fewer reads were mapped and the number of contigs reconstructed decreased by 431. When the value of maximum mismatch allowed was increased to 20 bp, about 7000 more read pairs were mapped (totalling 13,369,423 read pairs) and 481 more contigs were assembled. The total number of base pairs assembled also increased by around 150,000 bp (Table 4.8). The number of mapped read pair increased with high mismatches allowed, and the number was saturated once the value reached 50 bp, with 13,372,943 out of 15,083,182 read pairs mapped (Table 4.8). However, the number of total contigs assembled varied and the highest number of contigs assembled was achieved with 91,668 when the maximum mismatch allowance was set to 60 bp. The number of assembled contigs and the total base pairs decreased once the parameter was set above 70 (Table 4.8). The BUSCO completeness was not affected by any change of the mismatch parameter. Thus, the parameter value of 60 for the mismatch allowance was used as the optimal parameter, with the optimisation moving on to the next parameter.

Table 4.8 Effect of changing 'mismatch allowed' mapping parameter on the assembly of *S. rexii* RNA-Seq data

	S. rexii	***S. rexii***	***S. rexii***
Parameter (Max mismatch allowed to be written into output file)	5	(Default) 10	20
No. reads mapped	13,368,275 × 2	13,362,520 × 2	13,369,423 × 2
Read mapped (%)	88.6	88.6	88.6
No. contigs	90,546	90,977	91,458
Total base pairs (bp)	100,883,359	101,157,269	101,334,267
Contigs N50 (bp)	1,645	1,644	1,641
GC (%)	41.8	41.8	41.8

Table 4.8 continued

	S. rexii	*S. rexii*	*S. rexii*
Parameter (Maximum mismatch allowed to be written into output file)	30	40	50
No. reads mapped	13,371,845 × 2	13,372,671 × 2	13,372,943 × 2
Reads mapped (%)	88.7	88.7	88.7
No. contigs	91,578	91,612	91,646
Total base pairs (bp)	101,430,383	101,370,348	101,421,998
Contigs N50 (bp)	1,641	1,640	1,640
GC (%)	41.8	41.8	41.8

Table 4.8 continued	*S. rexii*	*S. rexii*	*S. rexii*
Parameter (Maximum mismatch allowed to be written into output file)	60	70	80
No. reads mapped	13,372,943 × 2	13,372,943 × 2	13,372,943 × 2
Reads mapped (%)	88.7	88.7	88.7
No. contigs	91,668	91,652	91,651
Total base pairs (bp)	101,460,759	101,433,335	101,426,394
Contigs N50 (bp)	1,640	1,640	1,640
GC (%)	41.8	41.8	41.8

The last parameter optimised was the 2nd pass mapping mode. When 2nd mapping pass mode is enabled alone, the number of read pairs mapped increased from 13,362,520 to 13,410,025 and the mapping percentage improved from 88.7 to 88.9% (Table 4.9). The total number of contigs assembled was 91,036, 632 lower than that of the best settings with a maximum mismatch allowance of 60 (Table 4.9, 91,668 contigs) but higher than the default setting (Table 4.8; 90,977 contigs). When combining both ‘2nd pass mapping mode’ and 'maximum mismatch allowed = 60', 13,416,581 read pairs were mapped (89.0%) which was the highest among all the tested parameter settings (Table 4.9). However, the total number of contig assembled decreased to 91,491 and is lower than using the 'maximum mismatch allowed = 60' parameter alone (Table 4.9). The BUSCO completeness of all three assemblies were identical (78.5%). Since the aim of the optimisation was to reconstruct and rescue as many contigs as possible, and the 2nd pass mapping optimisation did not really improve the BUSCO completeness, the parameter settings that generates the highest number of contigs (i.e. maximum mismatch allowed = 60) was chosen for the final assembly of both *S. rexii* and *S. grandis* reference-guided transcriptomes.

Table 4.9 Effect of enabling '2nd mapping pass' mapping parameter on the assembly of *S. rexii* RNA-Seq data

	S. rexii	***S. rexii***	***S. rexii***
Parameters	2nd mapping pass	Maximum mismatch allowed = 60	2nd mapping pass + Maximum mismatch allowed = 60
No. reads mapped	13,410,025 × 2	13,372,943 × 2	13,416,581 × 2
Read mapped (%)	88.9	88.7	89.0
No. contigs	91,036	91,668	91,491
Total base pairs (bp)	101,340,620	101,460,759	101,420,359
Contigs N50 (bp)	1,645	1,640	1,641
GC (%)	41.8	41.8	41.8
Transcriptome completeness			
BUSCO completeness (%)	78.5	78.5	78.5
No. completed BUSCOs	1,131	1,131	1,131
No. missing BUSCOs	309	309	309

Using the optimised parameter setting for reads mapping (i.e. maximum mismatch allowed = 60), reference-guided assembly was carried out for both *S. rexii* and *S. grandis* (Table 4.10). In *S. rexii* 88.7% of the total reads were mapped, and in *S. grandis* this figure was 93.5% (Table 4.10). 91,668 contigs were assembled for *S. rexii* and 73,962 contigs for *S.*

grandis. The N50 value was 1,640 bp and 1,661 bp, and the average GC percentage 41.8% and 42.2% for *S. rexii* and *S. grandis* respectively. In total, 1,131 and 1,163 BUSCO genes were found to be completed in the two assemblies, corresponding to 78.5% and 80.7% BUSCO completeness. 55,013 and 47,709 ORFs identified in the *S. rexii* and *S. grandis* assemblies respectively. Among these, 1,691 ORFs from *S. rexii* and 1,266 ORFs from *S. grandis* were classified as possible contamination and were removed (Table 4.10). Finally, 53,322 ORFs remained in the *S. rexii* transcriptome and 46,429 ORFs in the *S. grandis* transcriptome.

Table 4.10 Filtering of the optimised reference-guided transcriptome assemblies of *S. rexii* and *S. grandis*

	S. rexii	***S. grandis***
Sequencing platform	MiSeq + HiSeq4000	MiSeq
No. reads (after trimming)	15,083,182 × 2	10,266,774 × 2
Parameters (Max mismatch allowed)	60	60
No. of reads mapped	13,372,943 × 2	9,600,119 × 2
Read mapped (%)	88.7	93.5
No. contigs	91,668	73,962
Total base pairs (bp)	101,460,759	85,831,430
Contig N50 (bp)	1,640	1,661
GC (%)	41.8	42.2
Transcriptome completeness		
BUSCO completeness (%)	78.5	80.7
No. completed BUSCOs	1,131	1,163
No. missing BUSCOs	309	277
ORF identification		
No. ORFs identified	55,013	47,709
Contaminant identification		
Total no. of contaminant contigs	1,691	1,266
Archaea	26	16
Bacteria	617	440
Protists	138	121
Fungis	211	144
Animal	588	476
Other plants	75	54
Viruses	28	13
Unidentified	8	2
Possible contaminant (%)	3.0	2.6
Final number of contigs kept after filtering	53,322	46,429

For the filtered *S. rexii* transcriptome assembly, 53,322 contigs were retained with a total of 53,617,050 bp assembled. The average transcript length and transcript N50 was 1,005 bp and 1,242 bp, respectively. The GC percentage was 44.5% and the BUSCO completeness 76.7% (Table 4.11). For the *S. grandis* assembly, 46,429 contigs were kept with a total of 48,891,699 bp assembled (Table 4.11). The average transcript length was 1,053 bp, with the longest transcript of 10,758 bp, and the N50 value was 1,302 bp. The average GC percentage was 44.3%, and the BUSCO completeness 78.4%. For *S. rexii* 24,171 transcripts and 36,883 transcripts were annotated by the KEGG and Mercator pipelines respectively (Figure 4.9). These figures *S. grandis* were 20,971 transcripts and 31,026 transcripts respectively (Figure 4.10).

Table 4.11 Metrics of the filtered reference-guided assembled transcriptomes for *S. rexii* and *S. grandis*

	S. rexii	***S. grandis***
Dataset used	MiSeq + HiSeq 4000	MiSeq
Assembly metrics		
No. transcripts	53,322	46,429
Total base pairs (bp)	53,617,050	48,891,699
Average transcript length (bp)	1,005	1,053
Longest transcript (bp)	10,944	10,758
Transcript N50 (bp)	1,242	1,302
GC (%)	44.5	44.3
Transcriptome completeness		
BUSCO completeness (%)	76.7	78.4
No. completed BUSCOs	1,104	1,129
No. missing BUSCOs	336	311
Transcriptome annotation		
No. gene annotated by KEGG	24,171	20,971
No. gene annotated by Mercator	36,883	31,026

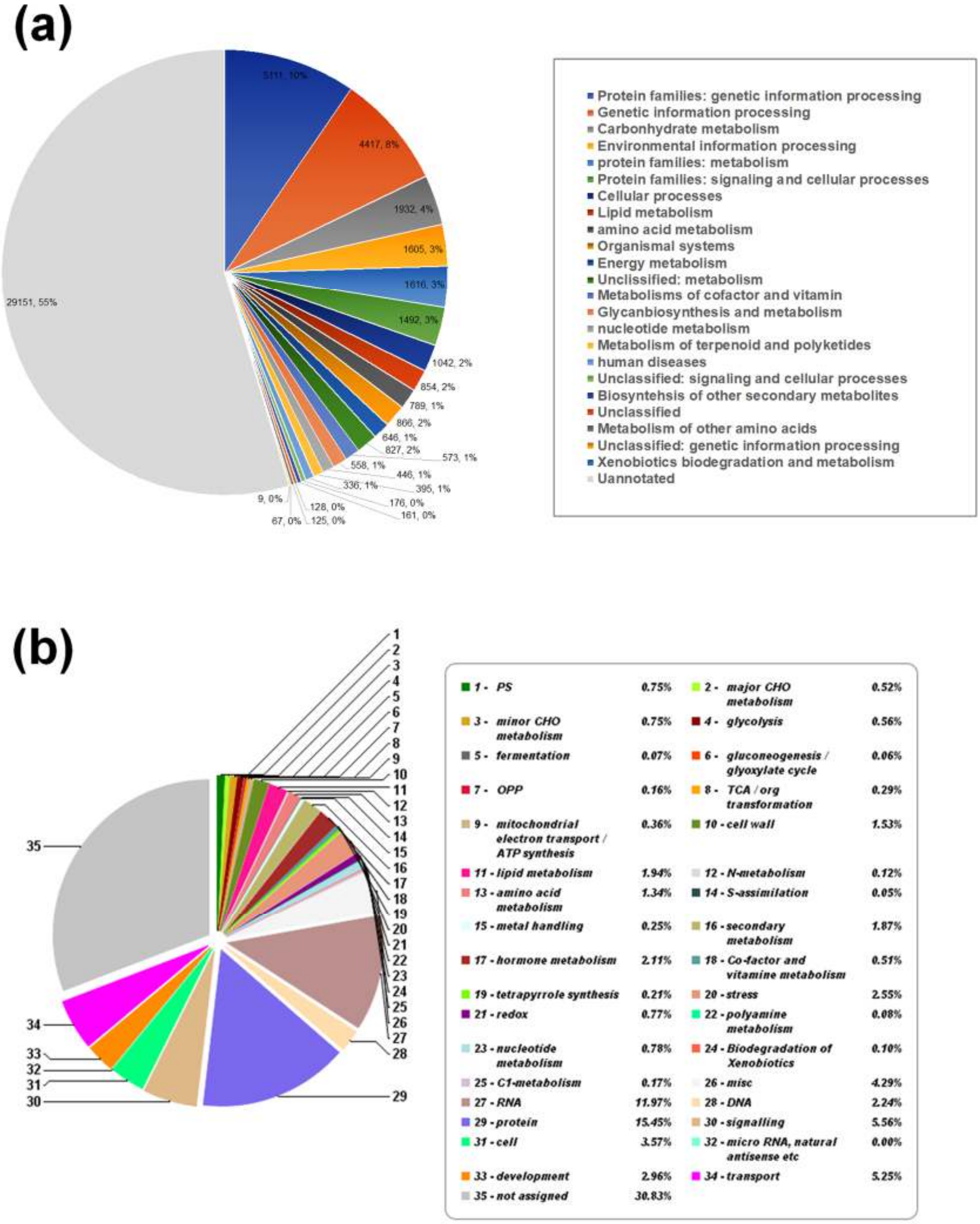


Figure 4.9 Functional annotation results of the *S. rexii* reference-guided transcriptome assembly (a) KEGG annotation. The data labels show the number of transcripts of the given category (right-hand side), followed by the percentage of that category in the whole transcriptome. (b) Mercator annotation. The labels show functions annotated.

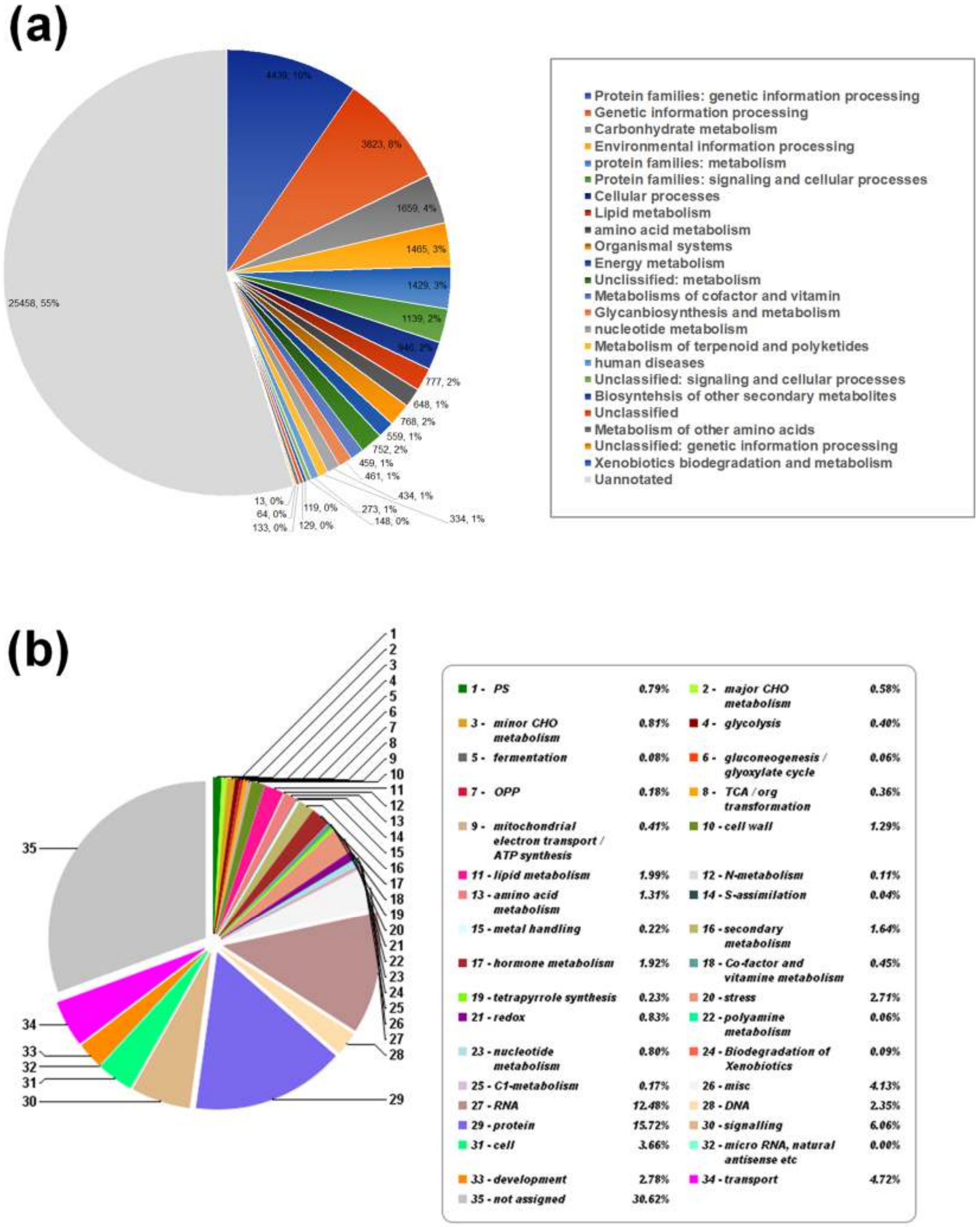


Figure 4.10 Functional annotation results of the *S. grandis* reference-guided transcriptome assembly (a) KEGG annotation. The data labels show the number of transcripts of the given category (right-hand side), followed by the percentage of that category in the whole transcriptome. (b) Mercator annotation. The labels show functions annotated.

A total number of 39,921 orthogroups were identified where each orthogroup contained at least one transcript (Figure 4.11; Appendix 4.1). Among these, 18,947 orthogroups were found in all four transcriptomes (Figure 4.11). Unique orthogroups were found in each transcriptome (i.e. unique within the assemblies), with seven, three, one, and five transcriptome-specific orthogroups found in *S. rexii de novo*, *S. rexii* reference-guided, *S. grandis de novo*, *S. grandis* reference-guided transcriptome respectively. When comparing the two assemblies within species (i.e. *de novo* and reference-guided), very different gene sets were found: between *S. rexii de novo* and reference-guided transcriptomes, only 30,205 orthogroups were found conserved (Figure 4.11, overlap between blue and purple; 8,453 + 1,150 + 18,947 + 1,655); between *S. grandis de novo* and reference-giuded transcriptomes, only 26,721 groups were found to be conserved (Figure 4.11, red and green; 5,443 + 1,182 + 18,947 + 1,149). When comparing between species (*S. rexii* to *S. grandis*), 8,463 unique orthogroups were found for *S. rexii* (Figure 4.11; 3 + 8,453 + 7) and 5,449 were found in *S. grandis* (Figure 4.11; 1 + 5,443 + 5).

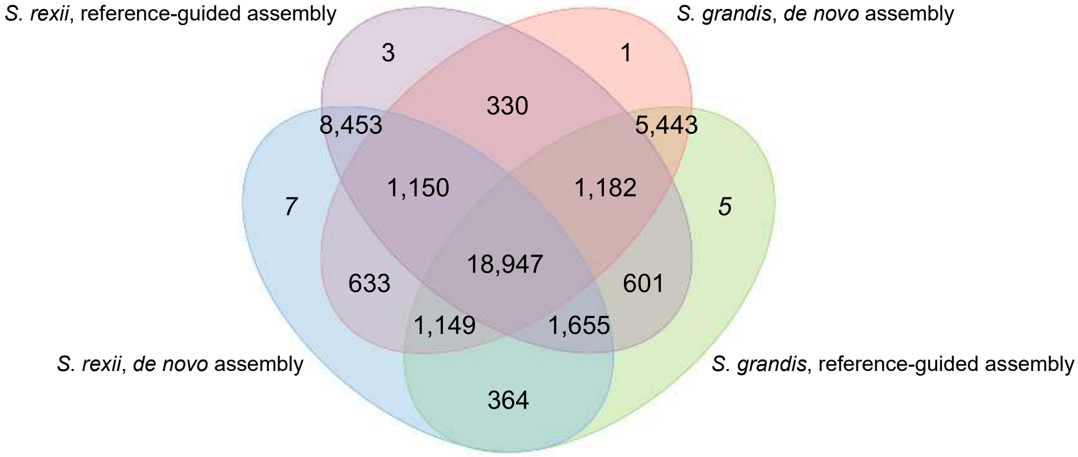

S. rexii, reference-guided assembly
S. grandis, *de novo* assembly
3
1
330
8,453
5,443
1,150
1,182
7
5
18,947
633
601
1,149
1,655
S. rexii, *de novo* assembly
S. grandis, reference-guided assembly
364

Figure 4.11 Venn diagram showing the number of shared and unique orthogroups identified in all four finalised transcriptome assemblies

4.4 Discussion

4.4.1 Comparison of the transcriptome assemblies of *S. rexii* and *S. grandis*

In this chapter, RNA-Seq was performed for generating transcriptome resources for *S. rexii* and *S. grandis*. The RNA-Seq included transcripts from both vegetative and reproductive tissues, and *de novo* and reference-guided assembly approaches were pursued. The assembly metrics of all four finalised assemblies compared and summarised in Table 4.12. Overall, the *de novo* approach produced more transcripts and incorporated more base pairs compared to the reference-guided assemblies. The average length and N50 values were similar between the *de novo* and reference-guided assemblies, with the *de novo* assemblies showing slightly better contiguity (about 50 to 100 bp longer in both average length and N50 value). The length of the longest transcript assembled was similar among all four assemblies, as was the average GC content. A possible explaination for the higher transcript count observed in the *de novo* assembly was that *de novo* approaches tend to recover more isoforms per locus, and detect more short transcribed fragments that are possibly ignored by reference-guided assembly (Lu et al., 2013).

In terms of gene completeness, the *S. rexii* the reference-guided assembly had four more BUSCO genes identified compared to the *de novo* assembly (Table 4.12). However, for *S. grandis* the reference-guided assembly had 9 fewer BUSCOs identified compared to the *de novo* assembly (Table 4.12). One possible explanation could be over-stringent filtering, as the BUSCO completeness of the *S. grandis* assembly prior to the filtering had the highest BUSCO completeness of 80.7% (Table 4.12). A similar trend was observed in the *S. rexii* assemblies, with the BUSCO completeness decreasing after ORF and contaminant filtering (Table 4.12). Interestingly, additional BUSCO analyses of the transcriptomes at different filtering stages suggested that the drop of the value occurred mainly at the ORF identification stage. This implies that the ORF identification process carried out by the TransDecoder software failed to identify the ORFs of some core genes. Possible improvements could be made for this step, is by decreasing the ORF length cut off value, and by incorporating homology searches of the transcripts to a known protein database to increase the sensitivity of ORF detection (Haas et al., 2013). Still, both prior- and post-filtering assemblies should be retained, in the case that the target gene of interest cannot be found in the post-filtering transcriptome one can still try to find it in prior-filtering assemblies.

The orthogroup identification results suggested that *de novo* and reference-guided approaches recovered different sets of genes. For example, between the two *S. rexii* assemblies, the *de novo* and reference-guided assemblies consisted of 2,153 and 2,116 non-overlapping orthogroups respectively. On the other hand, the *de novo* and reference-guided *S. grandis* assemblies had 2,114 and 2,625 non-overlapping orthogroups respectively (Figure 4.11). One possibility is that in *de novo* assemblies, the reconstruction of the transcript was limited by the depth of sequencing coverage of some expressed genes, and by providing high-coverage data it may even out-perform the reference-guided assembly (Lu et al., 2013). On the other hand, the reference-guided assembly was good at recovering transcripts with low sequencing coverage. However, here the assembly quality was limited by the quality of the genome assembly, i.e. in the case that the genomic region was not assembled, the transcript cannot be reconstructed. The *de novo* and reference-guided assemblies were suggested to be complementary to each other (Lu et al., 2013; Visser et al., 2015). By keeping the results of both assembly approaches, a more complete transcriptome profile of *S. rexii* and *S. grandis* can be captured.

Table 4.12 Statistics summary of the transcriptome assemblies in this study

	S. rexii* *De novo	***S. rexii* Reference-guided**	***S. grandis* *De novo***	***S. grandis* Reference-guided**
Dataset used	MiSeq + HiSeq 4000	MiSeq + HiSeq 4000	MiSeq	MiSeq
Assembly metrics (before filtering)				
No. transcripts	110,955	91,668	87,665	73,962
Total base pairs (bp)	123,213,415	101,460,759	102,299,541	85,831,430
Transcript N50 (bp)	1,716	1,640	1,541	1,661
GC (%)	42.1	41.8	42.5	42.2
BUSCO completeness (%)	79.0	78.5	79.0	80.7
Assembly metrics (after filtering)				
No. transcripts	60,500	53,322	51,267	46,429
Total base pairs (bp)	64,548,015	53,617,050	58,554,237	48,891,699
Average transcript length (bp)	1,066	1,005	1,142	1,053
Longest transcript (bp)	10,947	10,944	11,934	10,758
Transcript N50 (bp)	1,323	1,242	1,410	1,302
GC (%)	44.6	44.5	44.3	44.3
BUSCO completeness (%)	76.4	76.7	79.0	78.4
No. annotated gene (KEGG)	27,811	24,171	23,520	20,971
No. annotated gene (Mercator)	41,002	36,883	34,898	31,026

4.4.2 Comparison with other transcriptome resources

When compared to other Gesneriaceae transcriptomes, our assemblies showed reasonable assembly metrics (Table 4.13). The *S. rexii* assemblies recovered about twice the amount of transcripts compared to the previous *S. rexii* transcriptome, although with less contiguity (Chiara et al., 2013). On the other hand, our *S. rexii* and *S. grandis* assemblies had a more reasonable number of transcripts (containing 46,429-60,500 transcripts) compared to the *S. ionanthus* transcriptome (120,278), but with a much better contiguity (1,005-1,142 *versus* 326 average length). When compared to other Gesneriaceae species, our assemblies had a medium number of transcript counts (Table 4.13; *S. ionanthus*, *Primulina eburnea* and *Primulina pteropoda* had over 100,000 transcripts; *Achimenes cettoana* about 29,000 transcripts) and a roughly similar transcript length and N50 value. When compared to the transcriptome of the model species *Arabidopsis thaliana* (Zhang et al., 2017), our assemblies had lower transcript counts and less continuity (Table 4.13). However, the *A. thaliana* transcriptome was based on multiple sequencing libraries of different strains of plants exposed to various growth conditions, which would have allowed for the retrieval of more genes (Zhang et al., 2017).

A main feature of the *Streptocarpus* transcriptomes generated here was that the RNA samples were derived from various vegetative and reproductive tissues, compared to the other Gesneriaceae transcriptomes where the sample collections were limited to either vegetative or reproductive organs (Table 4.13). It is known that gene expression patterns vary widely between different cells and developmental stages (Yanofsky, 1995; Fletcher, 2002; Huijser and Schmid, 2011; Banks, 2015). Our inclusion of seedlings, roots, shoots, floral buds, flowers and developing fruits covered most of the cell types, and several of the main developmental stages of *Streptocarpus*. Thus, our transcriptome is likely largely complete and suitable for the purpose of genome annotation (Yandell and Ence, 2012; Hoff et al., 2016). However, a disadvantage is that since the extracted RNAs were pooled and sequenced together, the actual expression level of the genes in each tissue is no longer retrievable. Thus, more RNA-Seq experiments of specific tissues would be needed to examine the overall changes of gene expression levels at different developmental stages.

Table 4.13 Statistical summary of the transcriptomes of Gesneriaceae and *A. thaliana*

Species	Tissues	No. transcript	Avg. length / N50 (bp)	Reference
Genus *Streptocarpus*				
S. rexii (*de novo*)	All*	60,500	1,066 / 1,323	This study
S. rexii (*ref-guided*)	All	53,322	1,005 / 1,242	This study
S. grandis (*de novo*)	All	51,267	1,142 / 1,410	This study
S. grandis (*ref-guided*)	All	46,429	1,053 / 1,302	This study
S. rexii	Leaf+ cotyledon	33,113	2,064 / 2,556	Chiara et al., 2013
S. ionantha	Leaf	120,278	326 / 488	Matasci et al., 2014
Genus *Achimenes*				
A. cettoana	Flower	29,065	1,417 / 2,113	Roberts and Roalson, 2017
A. erecta	Flower	41,381	1,268 / 2,061	Roberts and Roalson, 2017
A. misera	Flower	41,285	1,260 / 1,990	Roberts and Roalson, 2017
A. patens	Flower	37,898	1,304 / 2,109	Roberts and Roalson, 2017
Genus *Damrongia*				
D. clarkeana	Leaf	94,546	487 / 1,075	Wang et al., 2017
Genus *Dorcoceras*				
D. hygrometricum	Leaf	49,374	2,535 / -	Xiao et al., 2015
Genus *Sinningia*				
S. tuberosa	Leaf	56,809	691 / 1,573	Matasci et al., 2014
S. eumorpha	Leaf + flower	87,053	1,687 / 2,597	Serrano-Serrano et al., 2017
S. magnifica	Leaf + flower	97,023	1,545 / 2,794	Serrano-Serrano et al., 2017
Genus *Primulina*				
P. eburnea	Leaf + root	106,665	1,086 / 1,823	Ai et al., 2014
P. fimbrisepala	Leaf + root	94,033	989 / 1,607	Ai et al., 2014
P. heterotricha	Leaf + root	92,255	1,201 / 1,915	Ai et al., 2014
P. huaijiensis	Leaf + root	76,495	962 / 1,582	Ai et al., 2014
P. lobulata	Leaf + root	81,271	1,144 / 1,847	Ai et al., 2014
P. lutea	Leaf + root	70,426	903 / 1,506	Ai et al., 2014
P. pteropoda	Leaf + root	108,947	1,036 / 1,709	Ai et al., 2014
P. sinensis	Leaf + root	75,523	965 / 1,609	Ai et al., 2014
P. swinglei	Leaf + root	91,113	921 / 1,538	Ai et al., 2014
P. tabacum	Leaf + root	82,357	1,113 / 1,785	Ai et al., 2014
P. villosissima	Leaf + root	75,614	926 / 1,470	Ai et al., 2014
Genus *Arabidopsis*				
A. thaliana	All	82,190	1,858 / 2,163	Zhang et al., 2017

*: All includes seedlings, leaves, roots, flower bud, flowers and developing fruits

-: Information was not provided

4.4.3 Conclusions

Here we present four finalised transcriptome assemblies of *S. rexii* and *S. grandis*. The different gene sets recovered by different assembly approaches suggested that the assemblies were complimentary to each other, and together they provided a largely complete gene expression profile for both *Streptocarpus* species. The focus of the study here was to produce a preliminary resource for *Streptocarpus*. Further characterisation of the transcriptomes could be made to improve the datasets, such as transcript length and number of transcript / isoform per gene, identification of alternative splicing, gene ontology annotation, molecular pathway reconstruction, or orthogroup identifications with other species. Still, the RNA-Seq data and transcriptomes generated here represent invaluable resources for future studies on *Streptocarpus* and the Gesneriaceae family, either for candidate gene approaches or for the structural and functional annotation of the whole genome assembly.

Chapter 5 Building a genetic map – Linkage analysis for constructing a genetic map from a *S. rexii* × *S. grandis* backcross population

5.1 Introduction

5.1.1 Applications of genetic mapping in candidate gene identification

A genetic map (or genetic linkage map) shows the relative positions of genes or genetic markers on the chromosomes in the genome (Van Ooijen and Jansen, 2013). The genetic distance is determined by the mean number of recombination events (crossovers) occuring per meiosis. In theory, the more frequently recombination between the markers occurs, the longer the relative physical distance (Semagn et al., 2006; Van Ooijen and Jansen, 2013). The first genetic map was calculated in 1913 for fruit fly by Alfred H. Sturtevant, who showed that the frequency of recombination events could be an index of the distance between two genes (Sturtevant, 1913). The 'map unit' for the genetic distance was termed centiMorgan (cM), in honour of Thomas H. Morgan, Sturtevant's supervisor. The distance of 1 cM corresponds to an average of one crossover in every 100 gametes, or 1% recombination frequency (Kole and Abbott, 2008).

Genetic mapping is widely used in plant research. To date, genetic maps have been constructed for hundreds of plant species, including key model plants such as *Arabidopsis*, rice, maize, and tomato (Koornneef et al., 1983; Beavis and Grant, 1991; Harushima et al., 1998; Meinke et al., 2003; Frary et al., 2005). Genetic maps serve five major purposes: (1) They allow the genetic analysis of quantitative traits and the mapping of the quantitative trait loci (QTL mapping). (2) They facilitate marker-assisted selection for the introgression of genes or QTLs for plant breeding. (3) They allow comparative mapping between species, to identify similarity and differences in gene order and distance. (4) They can be used to anchor the physical map of DNA scaffolds. (5) Building a genetic map is also the first step for positional or map-based cloning of genes (Lewis, 2002; Semagn et al., 2006; Broman and Sen, 2009). Genetic maps are widely used in molecular breeding. For example, the wheat genetic map was used to locate drought resistant alleles in wild wheat species, which were later introgressed into the bread wheat genome and successfully improved bread wheat stress resistance (Peleg et al., 2008; 2009; Merchuk-Ovnat et al., 2016). A rice genetic map constructed to identify QTLs related to biomass yield allowed QTL-based selection of rice lines to produce grains with higher biomass (Matsubara et al., 2016). Genetic maps have also been used for evolutionary studies. For instance, the map of *Petunia* was used to identified two floral scent related QTLs, which are associated with the pollinator preference (Klahre et al., 2011). The genetic map of *Rhytidophyllum* identified QTLs regulating pollination syndromes including floral dimensions, nectar volumes and flower colour, potentially linked to speciation (Alexandre et al., 2015).

Prior to this study, no genetic map was available for *Streptocarpus* (Chen et al., 2018). Genetic maps are available for other Gesneriaceae genera: *Rhytidophyllum* (Alexandre et al., 2015) and *Primulina* (Feng et al., 2016), however, these maps focused on floral characters, and are not informative for vegetative development. *Streptocarpus* material at RBGE provides the opportunity to identify the genetic basis of differences between rosulate and unifoliate growth forms by QTL mapping in a hybrid population between rosulate and unifoliate *Streptocarpus* species.

5.1.2 RAD-Seq for linkage analysis in non-model organisms

The development of NGS technologies allow high-throughput-genotyping of hundreds to thousands of markers with reasonable costs and time, without the need of preliminary knowledge of genetic marker and reference sequences (Davey et al., 2011). Such high-throughput genotyping methods also enable the construction of ultra-dense genetic maps that greatly improve the precision of QTL localisation and help distinguish closely located QTLs (Stange et al., 2013). Restriction site-Associated DNA sequencing (RAD-Seq) is one of these high-throughput genotyping methods, which has been widely applied for genetic mapping in non-model organisms (Baird et al., 2008; Davey and Blaxter, 2010). For example, the recently published genetic map of oil palm (*Elaeis guineensis*) was constructed using a combination of RAD-Seq and microsatellite markers, which resulted in 10,023 mapped markers (of which 9,712 markers were derived from RAD-Seq data) across 16 linkage groups with a total span of 2,938.2 cM (Bai et al., 2018).

RAD-Seq is a reduced representation sequencing approach, which sequences the subset of the genome flanking restriction sites of a selected restriction enzyme (Baird et al., 2008). This is achieved by the combination of restriction-enzyme reaction, ligation of a molecular identifier (MID), and Illumina sequencing technology (Illumina Inc, San Diego, CA, USA) (Baird et al., 2008). In brief, the genomic DNA of each individual of the mapping population is first digested by restriction enzymes, and the overhanging site tagged with an Illumina sequence primer site attached with an individual-specific barcode. The DNA samples are then pooled, and the standard Illumina library preparation protocol taken to prepare an Illumina sequencing library, which includes mechanical shearing, size-selection, ligation of adapters and PCR (Figure 5.1; Davey and Blaxter, 2010). The prepared library is then sequenced, and the reads demultiplexed into individuals based on the individual-specific barcodes, thus the genotype of each individual can be recovered (Davey and Blaxter, 2010).

RAD-Seq allows the genotyping of hundreds to thousands of loci of multiple individuals within a single Illumina sequencing lane (Baird et al., 2008; Davey and Blaxter, 2010). The number of markers genotyped can be regulated by the choice of restriction enzymes (i.e. 8-base cutter produces less markers than 6-bases cutters; Catchen et al., 2017). A RAD-Seq protocol was first used for genotyping of bulk segregant analysis in sticklebacks (Baird et al., 2008). Alternative protocols such as double digest RAD (ddRAD) were later developed. This method omits the random shearing step during library preparation to reduce the bias caused by randomisation, and incorporates digestion by two restriction enzymes to produce the DNA fragments allowing fine tuning of marker numbers by different enzyme combinations (Peterson et al., 2012).

RAD-Seq data can be analysed to reconstruct the genetic marker by *de novo* or reference-guided approaches (Baird et al., 2008; Davey et al., 2011; Davey and Blaxter, 2010; Catchen et al., 2011). In *de novo* approach, RAD reads can be *de novo* assembled to form stacks of sequences, in which each stack represents a genetic locus. Single-nucleotide polymorphism (SNP) genotypes can be determined from the assembled sequences. For the reference-guided approach, the RAD reads are aligned to a reference sequence. The aligned reads form stacks, which can be used to retrieve the SNP genotypes from the corresponding genome location. The advantage of the *de novo* approach is it does not require a reference genome sequence, which most often is not available for non-model species. On the other hand, the reference-guided approach has been shown to recover more markers from the RAD-Seq data (Shafer et al., 2016; Fountain et al., 2016), and requires less sequencing depth for correct genotyping (Fountain et al., 2016).

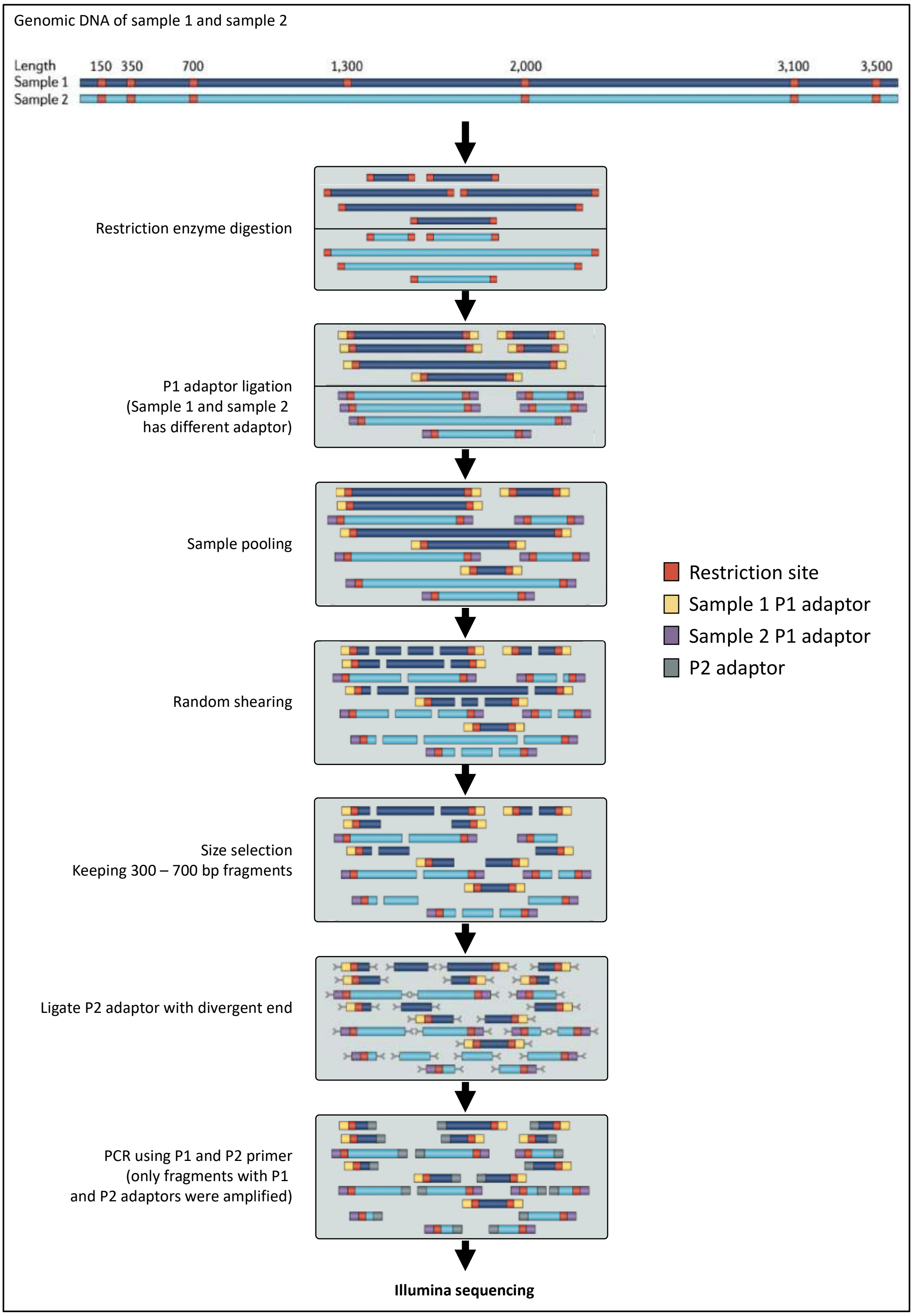


Figure 5.1 Schematic illustration of the procedure of RAD-Seq library preparation, using two samples as example (Modified from Davey et al., 2011)

5.1.3 Objectives

In this chapter, a backcross (BC) population ((*S. grandis* × *S. rexii*) × *S. grandis*) was used for genetic linkage mapping. RAD-Seq was performed on the BC population and the parental materials for genotyping, and d*e novo* and reference-based approaches were carried out with the obtained data (Figure 5.2). For the reference-based approaches, the *S. rexii* genome assemblies presented in Chapter 3 were used as references; *S. rexii* genome was chosen over the *S. grandis* one because the former shows better assembly contiguity, which is important for downstream analysis (e.g., designing new markers or gene mining). The alignment of the RAD reads was carried out using the software Burrow-Wheeler Aligner (BWA, Figure 5.2 a; Li and Durbin, 2009) and Stampy (Figure 5.2 a; Lunter and Goodson, 2011): The former uses Burrow-Wheeler Transform to perform string matching for alignment (Li and Durbin, 2009); the later uses a hash-based method, by building the hash table of the reads and reference, and alignment done by comparing the hashes (Lunter and Goodson, 2011).

The software Stacks (Catchen et al., 2011, 2013) is used for genotype calling and filtering of both *de novo* and reference-based analyses (Figure 5.2). In brief, in *de novo* approach the software first (1) attempts to cluster RAD-Seq reads that share highly similar sequences to form 'stacks' (parameter M). These stacks can be interpreted as presented haplotypes. Also, filtering of the read depth of stacks is applied to remove those without sufficient read depth (parameter m). For example, a haplotype derived from a read depth of 8× stack can be more reliable than that derived from a read depth of 1× (which can possibly be a sequencing error). (2) The unaligned reads, named 'secondary reads', are attempted to be aligned again to 'stacks' formed in previous step using a user-defined base pair mismatch allowance (parameter N). This is meant to recue reads from being wasted. (3) The software then tries to merge similar 'stacks' (again using a user-defined number of mismatch base pair; parameter n) to determine which haplotypes are originated from the same locus/marker. (4) Finally, by comparing the genotype of each locus with the parental genotypes, the genotype of each indivudal at each locus is determined. In reference-based approach, as all the read alignment and adjustment is done by a chosen aligner, only the stacks read depth filter (i.e., parameter m) is used. After the genotypes of all individuals are called, one can improve the reliability of the result by filtering out markers that have too many missing data. Typically, a threshold of <20% missing data is recommended (Hackett and Broadfoot, 2003).

The linkage analysis of the genotyping results is performed using software JoinMap (Van Ooijen, 2006; Van Ooijen and Jansen, 2013). The software performs further genetic marker filtering, such as removing markers showing segregation distortion or similar segregation patterns. The linkage group is then calculated using a maximum likelihood method, which gives LOD values (i.e., log value of the likelihood ratio) for each group defined; a LOD threshold ≥3 and ≤7 is generally recommended when selecting linkage groups (Van Ooijen and Jansen, 2013). Later, the genetic linkage map is calculated for each group using either regression mapping algorithm or maximum likelihood mapping algorithm. The former calculates genetic map by adding one marker at a time, and test the goodness-of-fit at each possible position the marker is added; the later is a quicker approach for linkage groups with > 100 loci, and works by smart search algorithm, expectation-maximisation (EM), and spatial sampling to approach the global optimal loci order within the linkage group (Van Ooijen and Jansen, 2013).

To maximise the number of markers recovered and the resolution of the final map, a series of mapping optimisations were carried out and the results compared. The first series of genetic map calculations (MapA) was done when only the preliminary genome assembly was available (*S. rexii* preliminary genome assembly using SOAPdenovo2; Table 3.4). The *de novo* approach, reference-based approach using BWA, and reference-based approach using Stampy were used to analyse the data. Later, the combined approach (i.e. combining markers generated from the *de novo* and both reference-based approach) was taken to calculate the final map (MapB), where the improved and filtered *S. rexii* genome assembly was used as reference sequence (*S. rexii* genome assembly using ABySS2; Table 3.16).

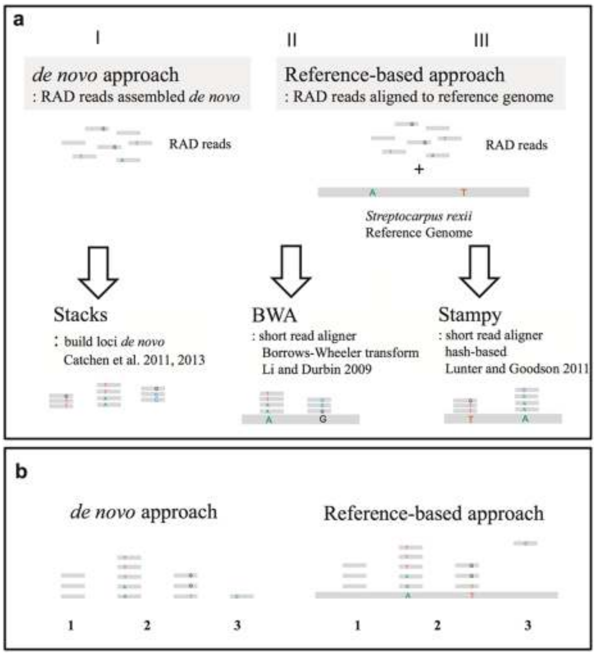


Figure 5.2 Schematic illustration of the RAD-Seq data analysis. **(a)** The three different approaches used to generate genetic markers from RAD-Seq data. **(b)** Type and quality of the recovered markers. 1 - Non-informative markers that do not contain SNP sites. 2 - Informative markers that have sufficient sequencing depth and contain SNP sites. 3 - Non-informative markers that do not have sufficient sequencing depth, or are unmapped reads. (Original figure from Chen et al., 2018)

5.2 Materials and methods

5.2.1 Plant materials

The backcross (BC) population ((*S. grandis* × *S. rexii*) × *S. grandis*) generated using the rosulate *Streptocarpus rexii* and unifoliate *Streptocarpus grandis* was used for genetic mapping (Figure 5.3). First, *S. rexii* (accession 19990270*I) pollen was transferred to the stigma of *S. grandis* flowers (denoted lineage *S. grandis*^{F1}; accession 20020577*Q) to produce an F1 hybrid. Later, pollen of another *S. grandis* lineage (denoted *S. grandis*^{BC}; accession 20130764*B) was used to pollinate the F1 plant (accession 20071108*J), and 233 backcross individuals were cultivated in this study (accession 20150825). A second accession of *S. grandis* had to be used for backcrossing as the species is monocarpic and the first parent died after flowering and fruiting. The full list of plant materials used in this study are summarised in Table 5.1.

All plant materials were cultivated in the Royal Botanic Garden Edinburgh living research collection throughout the experiments. The F1 hybrid *S. grandis* × *S. rexii* used for backcrossing to *S. grandis* and the BC seeds used in this study were generated in a previous study in 2007 and 2014 respectively (M. Möller, unpublished). The seeds of the BC population used for this study were sown in January 2015, and the DNA extractions of the BC individuals were carried out throughout 2015 until early 2016. For the parental lineages, *S. rexii* 19990270, *S. grandis*^{F1} 20020577, and *S. grandis 20130764* ^{BC}, were only available as silica-dried-leaf tissue at the time when this study was carried out.

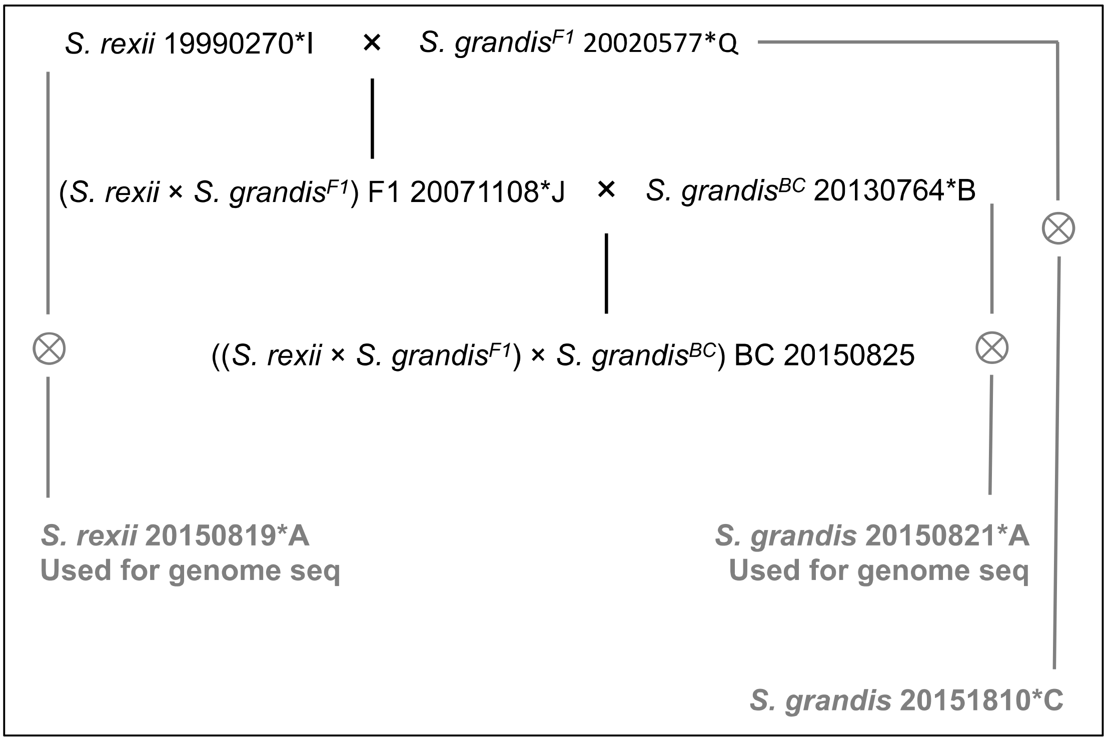

S. rexii 19990270*I × S. grandisF1 20020577*Q
(S. rexii × S. grandisF1) F1 20071108*J × S. grandisBC 20130764*B
((S. rexii × S. grandisF1) × S. grandisBC) BC 20150825
⊗
⊗
⊗
S. rexii 20150819*A
Used for genome seq
S. grandis 20150821*A
Used for genome seq
S. grandis 20151810*C

Figure 5.3 Pedigree of the *Streptocarpus* genetic mapping population

Table 5.1 List of plant materials used in the study

Taxon	Accession*Qualifier	Note
S. rexii	19990270*I	Parent of F1. F1 descendant of 19870333, collected by K Jong Collection number: JNG s.n. From Grahamstown; 'Faraway' Estate, Cape Prov., SE, ZA
S. rexii	20150819*A	Used for genome sequencing (Chapter 3); Descendent of 19990270*I
S. grandis^{F1}	20020577*Q	Parent of F1. Collected by M Möller. Collection number: MMO 2000-21. From Nkandla forest, KwaZulu-Natal Prov., ZA
S. grandis	20151810*C	Descendant of 20020577*Q
S. grandis^{BC}	20130764*B	Parent of BC. F3 descendant of 19771210, collected by OM Hilliard and BL Burtt Collection number: HBT 5923. From Ngome forest, KwaZulu-Natal Prov., ZA
S. grandis	20150821*A	Used for genome sequencing (Chapter 3); Descendent of 20130764*B
S. grandis^{F1} × *rexii*	20071108*J	Parent of BC
(*S. grandis*^{F1} × *rexii*) × *grandis*^{BC}	20150825*A-IS	Genetic mapping

5.2.2 DNA extraction, library preparation and sequencing

The DNA of the *S. rexii*, *S. grandis*, F1 hybrids and backcross individuals were extracted using the modified CTAB extraction method described in Chapter 2 (Appendix 2.5). In brief, the proximal regions of young-growing phyllomorphs were freshly collected and extracted using 4% CTAB solution with 2% PVPP. For the parental lineages, *S. rexii* 20020577, silica-dried-leaf tissues were used. The extracted DNA was further treated with RNase A and cleaned up using a phenol:chloroform:isoamyl alcohol 25:24:1 solution. The DNA of each sample was quantified using the Qubit 2.0 fluorometer with the Qubit dsDNA HS assay kit (Thermo Fisher Scientific). The samples were then diluted into 20 ng/µl based on the Qubit measurements to achieve the requested concentration for the RAD-Seq library preparation. The final concentration was confirmed again by Qubit fluorometry, and the results should fall between 15 – 21 ng/µl or the dilution step was repeated.

The diluted DNA samples were sent to our collaborator Dr Atsushi Nagano's group (Ryokoku University, Kyoto, Japan), where the library preparation and RAD-Seq was performed. The double digest RAD-Seq library was prepared following the protocols described previously (Peterson et al., 2012; Sakaguchi et al., 2015), in which the restriction enzymes *Eco*RI and *BglI*I were used. The libraries were sequenced on 3 lanes of HiSeq 2500 (Illumina), with 50 to 100 libraries per lane and 51 bp single-end sequencing of the *BglI*I restriction-end. In total 237 libraries were sequenced (233 BC individuals + 2 *S. rexii* + *S. grandis*^{F1} + *S. grandis*^{BC}). The data was demultiplexed and returned in fastq.gz format. The materials sequenced in each lane are listed in Appendix 5.1 a.

5.2.3 Quality control and preprocessing of the RAD-Seq data

The quality and adapter trimming of the RAD-Seq data was carried out using Trimmomatic v0.35 (Bolger et al., 2014). The LEADING, TRAILING, SLIDINGWINDOW, and AVGQUAL parameters were used for quality trimming, and reads shorter than 51 bp were discarded (MINLEN:51). The output reads were then processed by PRINSEQ-lite v0.20.4 (Schmieder and Edwards, 2011) to remove reads containing any unidentified nucleotide 'N' as suggested in Catchen et al., 2011. Finally, Bowtie v2-2.2.8 (Langmead and Salzberg, 2012) was used to filter out reads generated by organellar genomes, by mapping RAD-Seq reads to the assembled *S. rexii* chloroplast and mitochondrial genomes presented in Chapter 3 and keeping the unmapped reads (i.e. hence the reads) were kept thus removing the reads that was originated from the organellar genomes. The quality of the prior- and post-prepressed datasets was checked by FastQC v0.11.7 (Andrews, 2010}, and the results were summarised using MultiQC v1.5 (Ewels et al., 2016). The detailed commands and parameters used are listed in Box 5.1.

Box 5.1 Script for RAD-Seq data preprocessing. Text in bold with brackets **[Text]** indicate the input files and output file names to be specified.

```
# Pipe for Trimmomatic, prinseq-lite, and Bowtie2-preprocessing
java -jar trimmomatic-0.35.jar SE -phred33 \
    [RADseq.fastq] [Output.fastq] \
    ILLUMINACLIP: TruSeq3-SE.fa:2:30:10 \
    LEADING:20 TRAILING:20 \
    SLIDINGWINDOW:30:20 AVGQUAL:20 MINLEN:51 | \
    prinseq-lite -ns_max_n 0 \
    -fastq [Output.fastq] -out_format 3 -out_good [Output2.fastq] | \
    bowtie2 -x [index of Organelle genome] -q [Output2.fastq] -N 1 \
    --un-gz [RADseq_cleaned.fq.gz]
```

For *S. grandis*, the RAD-Seq data from the accession 20150821 (YYD17, Appendix 5.1 a) were used as the *S. grandis* parental lineage for all following RAD-Seq genotyping analysis via Stacks v1.47 (Catchen et al., 2011; 2013). For *S. rexii*, due to the low read count of both libraries, a 'superparent' file was constructed according to Catchen et al. (2011) by combining preprocessed RAD-Seq reads of accession 20150819 (YYD16, Appendix 5.1 a) and 19990270 (YYD19, Appendix 5.1 a). This 'superparent' file was used as the *S. rexii* parental lineage data for all the genotyping analysis described in the following sections. The 200 BC individuals with the highest read counts were identified and used for all the following genetic map calculations (Appendix 5.1 b).

5.2.4 Genotyping of RAD-Seq data - *de novo* approach

Stacks v1.47 was used for the *de novo* analysis of the RAD-Seq data. More specifically, the script *denovo_map.pl* was used, which clusters the reads based on sequence alignment. Each cluster (i.e. stack) represents a genetic locus, and the genotype can be determined using the script *genotypes* (Catchen et al., 2011; 2013).

For the genotyping of MapA (the preliminary map), three major parameters of *denovo_map.pl* were chosen for optimisation (Table 5.2; Catchen et al., 2011). These were 'm' (minimum stack depth), 'M' (mismatch allowed within an individual), and 'N' (mismatch allowed for merging secondary reads to the primary stacks). The 50 BC individuals with the highest read counts after preprocessing were chosen for the optimisation (Appendix 5.1 b). This involved testing different values of each parameter on the data of these 50 BC individuals and the two parents. The settings that generated the highest marker numbers in the optimisation tests were selected, and then applied for the analysis the data of the 200 BC individuals with the highest read counts (Appendix 5.1 b). The Stacks commands used are summarised in Box 5.2.

Table 5.2 Parameters tested for the *de novo* approach analysis for MapA

Parameters	Parameter definition	Values tested
m	Minimum stack depth	1,2,3,4,5,9,10
M	Mismatch allowed within individuals	1,2,3,4,5
N	Mismatch allowed for aligning 2nd reads to primary stacks	M, M+1, M+2

Box 5.2 Script for the genotyping using the *de novo* approach. Text in bold with brackets **[Text]** indicate the input files and output file names to be specified.

```
# denovo_map.pl analysis using the trimmed RADseq files as input
denovo_map.pl \
-p [S. grandis parent reads.fq.gz] -p [S. rexii parent reads.fq.gz] \
-r [BC individual 1 reads.fq.gz] \
-r [BC individual 2 reads.fq.gz] \
-r [BC individual 3 reads.fq.gz] \
…
-r [BC individual 199 reads.fq.gz] \
-r [BC individual 200 reads.fq.gz] \
-o [Output directory] \
-T [no. threads] –m [m value] –M [M value] –N [N value] -A BC1 -S -b 1

# Genotype calling. Only retain markers with less than 20% missing data
# (-r 160)
genotypes -b 1 -P . -r 160 -t BC1 -o joinmap
```

For the genotyping of MapB, an additional parameter *n* (mismatch allowed when building catalog) of *denovo_map.pl* was optimised in addition to the three parameters tested for MapA (Table 5.3; m, M, N). The optimisation was carried out using the data of the 50 BC individuals and the two parents as previously described (Appendix 5.1). Parameter settings showing the highest marker number were selected and applied for the analysis of the 200 backcross population selected as above. The commands used are described in Box 5.2.

Table 5.3 Parameters tested for the *de novo* approach analysis for MapB

Parameters	Parameter definition	Values tested
m	Minimum stack depth	1,2,3,4,5,6,8,10
M	Mismatch allowed within individuals	1,2,3,4,5
N	Mismatch allowed for aligning 2nd reads to primary stacks	M, M+1, M+2
n	Mismatch allowed when building catalog	1,2,3,4,5

5.2.5 Genotyping of RAD-Seq data – Reference-based approach using BWA aligner

BWA v0.7.15 (Li and Durbin, 2009) was used to align the RAD-Seq reads to the *S. rexii* draft genome produced in Chapter 3. SAMtools v1.7 (Li et al., 2009) was used to convert the alignment into BAM files. The BAM files were used as the input files for the Stacks v1.47 script *ref_map.pl*, which reconstructed the genetic loci information based on the alignment. Finally, the Stacks script *genotypes* was used for the genotyping and the generation of the locus genotype file.

For the genotyping of the MapA, the preliminary *S. rexii* genome assembly was used as the reference for the alignment (Table 3.4). The alignment was carried out using BWA *aln* under default parameters. The Stacks script *ref_map.pl* was used to reconstruct the loci from the BAM files with a minimum stack depth of 3 (-m 3). The detailed commands used are summarised in Box 5.4.

For the genotyping of MapB, the finalized *S. rexii* genome assembly was used as reference for mapping the RAD-Seq reads (Table 3.16, *S. rexii*). Furthermore, optimisation of two of the BWA alignment parameters was carried out. These were the maximum edit distance (n) and maximum edit distance in seed (k; Table 5.4). The data of 50 BC plants and the parents were used for their optimisation (Appendix 5.1 b). The optimal parameter settings that gave the highest number of markers were chosen, and subsequently applied for the alignment of the RAD-Seq data of 200 BC individuals (Appendix 5.1 b). In addition, the statistics of all the BAM files were accessed using SAMStats v1.5.1 (Lassmann et al., 2011). Data for the 200 BC individuals were combined and the number of mismatches per reads calculated and plotted. The Stacks script *ref_map.pl* was used to reconstruct the loci from the BAM files with a minimum stack depth of 3 (-m 3). The detailed commands used are summarised in Box 5.4.

Table 5.4 Parameters of the BWA aligner tested for the calculation of MapB

Parameter	Parameter definition	Value tested
n	Maximum edit distance	3, 4, 6, 12
k	Maximum edit distance in seed	1, 2 (default), 3

Box 5.4 Script for the genotyping using the reference-based approach with the BWA aligner. Text in bold with brackets **[Text]** indicate the input files and output file names to be specified.

```
# Generate BWA index file from S. rexii genome assembly
bwa index [genome assembly.fa]

# Align the RADseq reads using BWA, then use SAMtools to generate
# the BAM file
```

```
bwa aln –t [no. threads] –n [n value] –k [k value] \
    [genome index] [RADseq reads.fq.gz] > temp.sai
bwa samse [genome index] temp.sai [RADseq reads.fq.gz] \
    | samtools view -Sb \
    | samtools sort -O bam -o [Aligned RADseq reads.bam]

# ref_map.pl analysis to analyse the generated BAM files
ref_map.pl \
    -p [S. grandis parent reads.bam] -p [S. rexii parent reads.bam] \
    -r [BC individual 1 reads.bam] \
    -r [BC individual 2 reads.bam] \
    -r [BC individual 3 reads.bam] \
    …
    -r [BC individual 199 reads.bam] \
    -r [BC individual 200 reads.bam] \
    -o [Output directory] \
    –T [no. threads] –m 3 -A BC1 -S -b 1

# Genotype calling. Only retain markers with less than 20% missing data
# (-r 160)
genotypes -b 1 -P . -r 160 -t BC1 -o joinmap
```

5.2.6 Genotyping of RAD-Seq data – Reference-based approach using Stampy aligner

The overall analysis process was the same as with the BWA approach described in the previous section, except for that the aligner used here was Stampy v1.0.29 (Lunter and Goodson, 2011). Stampy aligner was chosen due to its different alignment algorithm used (hash-based) comparing to that of BWA (Burrows-Wheeler transform), and the alignment results generated from the two aligners were suggested to be complimentary to each others in terms of alignment speed and sensitivity (Lunter and Goodson, 2011). In brief, Stampy was used to align the preprocessed RAD-Seq reads to the *S. rexii* draft genome. SAMtools v1.7 was used to convert the alignment into BAM files. The Stacks v1.47 script *ref_map.pl* was used to reconstruct the genetic loci information, and script *genotypes* for the genotyping and the generation of the locus genotype files.

For the calculation of MapA, the reads were mapped to the preliminary *S. rexii* genome assembly under default parameters (Table 3.4). For the calculation of MapB, the reads were aligned to the filtered assembly, also under default parameters (Table 3.16, *S. rexii*). The alignment files were converted into BAM format using SAMtools v1.7 (Li et al., 2009), followed by the Stacks script *ref_map.pl* and *genotypes* analysis as described in the previous BWA mapping section. In addition, the mapping statistics were accessed using SAMStats v1.5.1 (Lassmann et al., 2011), which calculates the number of mismatches per read across the 200 BC individuals (Appendix 5.1 b). The detailed commands used are given in Box 5.5.

Box 5.5 Script for the genotyping using the reference-based approach with the Stampy aligner. Text in bold with brackets **[Text]** indicate the input files and output file names to be specified.

```
# Build the genome index file (.stidx)
stampy.py -G [output genome file name] [genome assembly.fasta]
# Build the genome hash table (.sthash)
stampy.py -g [output genome file name] -H [output hash table name]

# Perform Stampy alignment and generate the BAM file
stampy.py -g [genome file] -h [hash table] -M [RADseq reads.fq.gz] \
    | samtools view -Sb \
    | samtools sort -O bam -T tmp -o [Aligned RAD reads.bam]

# ref_map.pl analysis to analyse the generated BAM files
ref_map.pl \
    -p [S. grandis parent reads.bam] -p [S. rexii parent reads.bam] \
    -r [BC individual 1 reads.bam] \
    -r [BC individual 2 reads.bam] \
    -r [BC individual 3 reads.bam] \
    …
    -r [BC individual 199 reads.bam] \
    -r [BC individual 200 reads.bam] \
    -o [Output directory] \
    –T [no. threads] –m 3 -A BC1 -S -b 1

# Genotype calling. Only retain markers with less than 20% missing data
# (-r 160)
genotypes -b 1 -P . -r 160 -t BC1 -o joinmap
```

5.2.7 Combining the genotype locus file and calculation of the genetic map

For both MapA and MapB, the markers recovered from all three different approaches (i.e. *de novo* approach, reference-based approach using BWA, and reference-based approach using Stampy) were combined to test the effect on the number of markers when combining the approaches. To do so, a prefix with capital letters was first added to the genotype locus files to distinguish the markers recovered in the different approaches: "DN" represented markers recovered from the *de novo* approach, "BW" for the BWA approach, and "ST" for the Stampy approach. The genotype locus files from all three approaches were then concatenated, and filtered by keeping only the markers with parental genotypes of [aa×bb], which indicates that the genotype of the parents at these markers were homozygous and thus appropriate to be used for genetic mapping. The concatenated genotype locus file was then had its' header and footer information updated to reflect the actual property of the combined dataset (Box 5.6), i.e. file name, cross type, number of individuals, number of markers information, and the list of all individuals. The updated genotype locus file was used for genetic map calculation.

Box 5.6 Script for filtering and combining the markers generated from three approaches, i.e. *de novo*, BWA and Stampy. Text in bold with brackets **[Text]** indicate the input files and output file names to be specified.

```
# Add prefix to the genotype locus files generated from each approach
sed ‘s/^/DN/’ [original_denovo_loc.loc] > [denovo_loc.loc]
sed ‘s/^/BW/’ [original_BWA_loc.loc] > [BWA_loc.loc]
sed ‘s/^/ST/’ [original_Stampy_loc.loc] > [Stampy_loc.loc]

# Filter the markers and keeps the one with a parental genotype of aa×bb.
# Then remove the <aaxbb> string from the file as they will not be
# recognized by the software JoinMap 4.1
grep “aaxbb” [denovo_loc.loc] [BWA_loc.loc] [Stampy_loc.loc] |
    sed ‘s/<aaxbb>//’ > [Combined_file.loc]

# Update the header information (add following lines to Combined_file.loc)
# The number of markers can be calculated using wc command
name = Combined_file.loc
popt = BC1
nloc = [no. markers]
nind = [no. individuals]

# Append the footer information
tail –n 202 [any of the three loc file] >> Combined_file.loc
```

All genetic map calculations were performed in JoinMap 4.1 (Van Ooijen, 2006; Van Ooijen and Jansen, 2013). The Stacks output file (genotype locus file; .loc file) was loaded into the program and the markers were first filtered. Markers showing missing data in more than 20% of the BC individuals (i.e. less than 160 out of 200 BC genotyped individuals) were removed. Markers that showed severe segregation distortion ($P < 0.0005$) were also removed (the related information is indicated in the ‘Locus Genotype Frequency’ tab). The JoinMap diagnostic *Similarity of Loci*, which checks the pairs of markers showing identical genotypes, was then used to remove loci with similar segregation patterns under default settings.

For the calculation of MapA, the linkage groups were first identified based on the LOD scores calculated using the *Grouping Tree* function in JoinMap, with a minimum LOD threshold of 4 and a maximum of 7. The regression mapping algorithm was selected, and the Haldane’s mapping function was used for the calculation of the map distance (Haldane, 1919). The quality of the map was checked by Chi-square values (values should be <5) and nearest neighbour fit values (N.N.fit; no outstanding values) of each marker. If markers with outlying values were observed, those markers were excluded and the map recalculated. This process was iterated until no more markers showed outlying Chi-square or N.N.fit values. The final linkage map was visualized using MapChart v2.2 (Voorrips, 2002). The numbering of the linkage groups was then determined by the length of the calculated maps, which the longest linkage group denoted LG1, followed by the second longest linkage group as LG2, and so on. The flow chart of the analysis and calculation of MapA is summarised in Figure 5.4.

For the calculation of MapB, the same settings as described above were used. However, additional settings were tested to construct a map with a 'less-stringent' marker filtering strategy (Figure 5.5). As listed in Table 5.5, MapB-1 was calculated with exactly the same filtering strategy as described above for MapA. For MapB-2, the filtering strategy allowed markers with higher proportions of missing data (<30%) to be kept, and also includes markers with slightly more significant segregation distortion (markers with Chi-square value ≤ 0.0001 removed). MapB-3 had the exactly the same filtering strategy as MapB-2, but markers that were mapped in MapB-1 but not in MapB-2 (excluded due to Chi-square contribution > 5) were forcibly added in (Figure 5.5). The order of the linkage groups of MapB-2 and MapB-3 followed that of MapB-1, rather than being assigned a new linkage group order based on the map length. The remaining steps of the calculation were the same as described above for MapA. The final maps were visualised using MapChart v2.30 (Voorrips, 2002).

Table 5.5 Different marker filtering strategies for the calculation of the diverse MapBs

	MapB-1	**MapB-2**	**MapB-3**
Marker removal threshold for missing genotype%	Remove marker with >20% missing data	Remove marker with >30% missing data	Remove marker with >30% missing data
Threshold for removing segregation distorted markers*	Remove ≤ 0.0005	Remove ≤ 0.0001	Remove ≤ 0.0001
Additional modification	N/A	N/A	Forcibly include markers mapped in MapB-1 but not in MapB-2

* The value indicates the Chi-square test value of the deviation of segregation ratio from the expected 1:1 ratio. The lower the value the more significant the distortion.

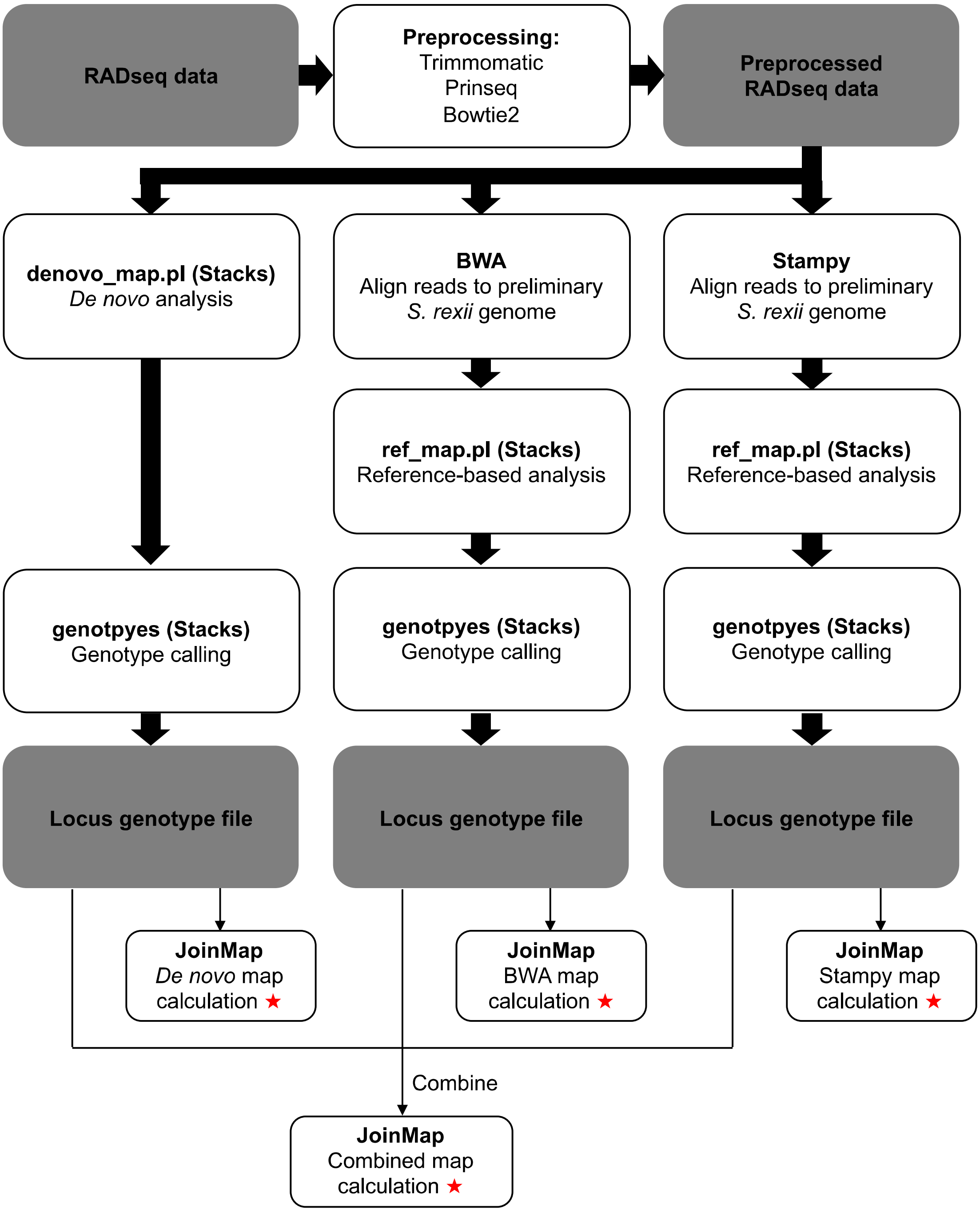


Figure 5.4 Flow chart of data analysis and the calculation of the MapA series. Grey rectangles – data file, White rectangles – analysis steps, Red stars – map calculation.

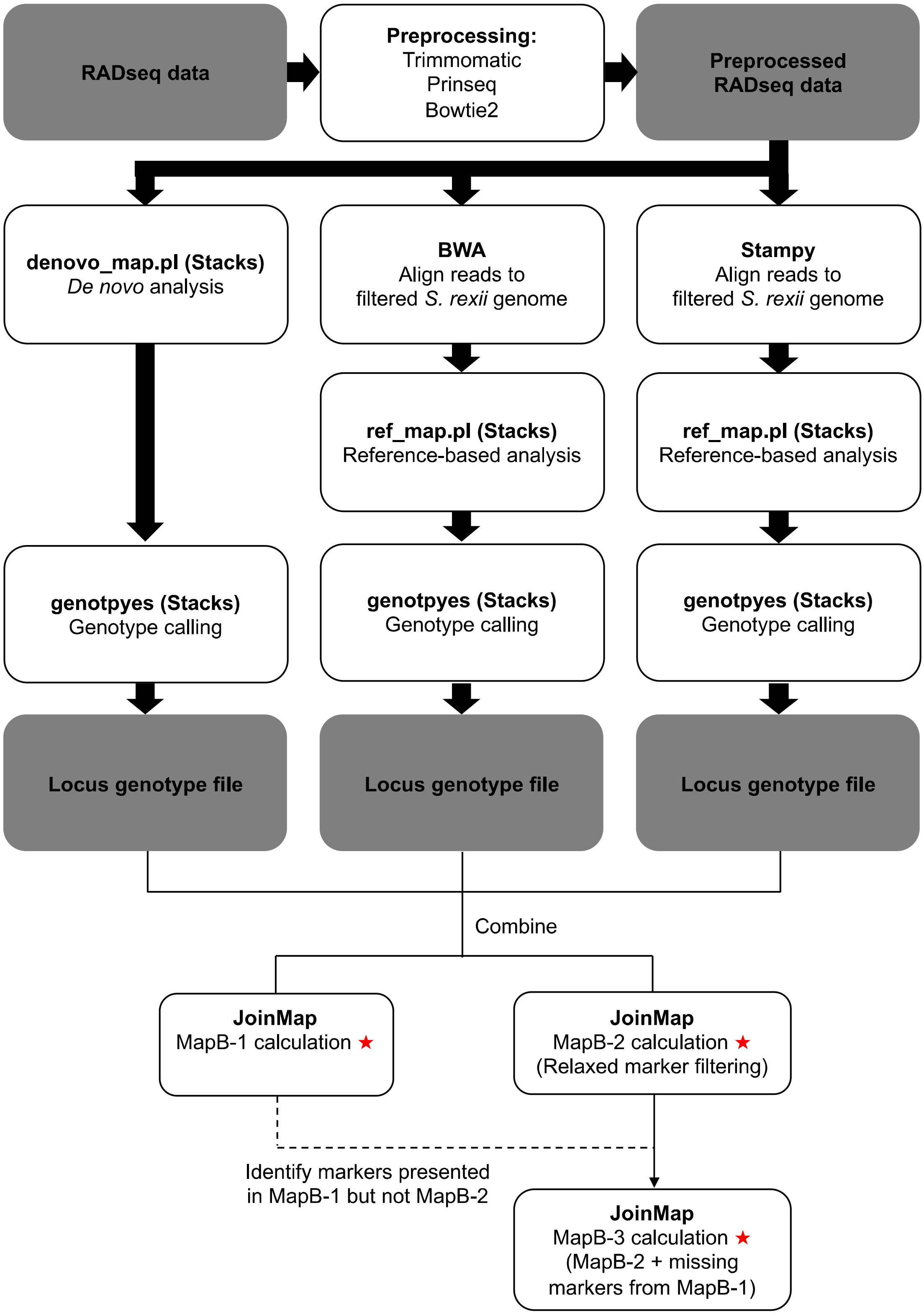


Figure 5.5 Flow chart of data analysis and the calculation of the MapB series. Grey rectangles – data file, White rectangles – analysis steps, Red stars – map calculation.

5.2.8 Synteny analysis of the genetic maps

Synteny analyses were carried out between MapB-1 and MapB-2, MapB-3 and the combined approach-MapA. For the synteny analyses between MapB-1, MapB-2 and MapB-3, the synteny relationships were visualised using MapChart v2.30 via the function 'Show homologs' (Voorrips, 2002). By enabling this function, the software draws lines among markers that share the same name, thus allowing a visual inspection of the differences in marker order and distances between two genetic maps.

For the synteny analysis between MapB-1 and MapA, since the markers from the two maps were generated from different Stacks analyses, the marker names assigned by Stacks were different. Thus, the marker synteny had to be compared based on the marker sequences. To do so, the information of the mapped markers was first extracted from the 'catalog.tags.tsv' file, an intermediate file produced during Stacks analyses described previously. The information extracted included marker name and locality (Genome scaffold name, strand, and position), and sequence (Catchen et al., 2011). For markers recovered by the *de novo* approach, no marker locality information was available.

The recovered information of the marker sequences and names was then used to produce a marker sequence-fasta file, with each fasta entry representing one genetic marker with the name of the entry identical to the name of the marker. *blastn* of the BLAST+ package v2.7.1+ (Camacho et al., 2009) was used to search for identical markers that was shared between two maps. This was done by using the 'marker sequence-fasta file' of the MapB-1 as BLAST query, and the 'marker sequence-fasta file' of the combined approach-MapA as BLAST database. Hits identified by the *blastn* search indicated the markers that were shared between the two genetic maps. The names of these shared markers were homogenised in the corresponding genetic map file (.map file; JoinMap4.1 manual). For example, if marker 'A' mapped in MapA, and marker 'B' mapped in MapB, were found to have the same sequences, they were renamed to 'Shared_marker_1' in the .map files of both MapA and MapB. Finally, the synteny relationship between two maps was visualised using MapChart v2.30 as described above. The detail commands used are summarised in Box5.7, including the commands for retrieving the marker sequencing information, transforming of the file to fasta format, blast search, and for the modification of marker names in the .map files.

Box 5.7 Script used for the synteny analysis between genetic maps. Text in bold with brackets **[Text]** indicate the input files and output file names to be specified.

```
# Retrieve marker information from Stacks output files
awk ‘$3==[MARKER_NAME]’ batch_1.catalog.tags.tsv >> [TAGS.INFO.OUTPUT]

# Transform the output file into fasta format
awk ‘{print “>”$3“,”$10}’ [TAGS.INFO.OUTPUT] | tr , ‘\n’ > [MARKER.fasta]

# Example of blastn search, using markers from MapB as query and MapA as
# database
makeblastdb -in [MAPA_MARKER.fasta] -dbtype nucl
blastn -task megablast \
    -query [MAPB_MARKER.fasta] -db [MAPA_MARKER.fasta] \
    -outfmt “6 qseqid sseqid qseq sseq” -max_target_seqs 1 \
    -out [OUTPUT_FILE.tsv]

# Example of changing marker name into ‘Shared_marker_N’ in the
# map1.map file. The final output was written into map1.synteny.map file
sed 's/[MARKER_NAME_1] /Shared_marker_1/' map1.map |
sed ‘s/[MARKER_NAME_2] /Shared_marker_2/’ |
sed ‘s/[MARKER_NAME_3] /Shared_marker_3/’ |
…
sed ‘s/[MARKER_NAME_N] /Shared_marker_N/’ > map1.synteny.map
```

5.3 Results

5.3.1 Quality check and preprocessing of the RAD-Seq reads

In total, 386,334,623 reads (19,703,065,773 bp) were obtained from all three lanes of the RAD-Sequencing (Table 5.6). On average, 1,630,104 reads per library were obtained (containing 386,334,623 bp). However, in reality the read count of each individual BC plant varied greatly, with the highest read count of 14,685,141 (Figure 5.6; BC individual qualifier HV), and the lowest of 10,706 reads (Figure 5.6; BC individual qualifier DL). The read quality was good in general, with most positions having average quality scores above 35 except for position 30 that showed a drop in average quality to 25 (Figure 5.7 a). The per sequence GC content graph showed a severe GC content bias in all sequenced libraries, with most libraries showing a pattern of multiple peaks of GC distribution (Figure 5.7 b). Further examination of the reads giving rise to these peaks indicated that these were highly represented reads derived from organellar genomes. The per base sequence N content graph showed N bases were called between position 40 to 50 bp in 29 libraries (Figure 5.7 c). The sequence duplication level showed a large proportion of overrepresented reads with over 5,000 or 10,000 duplicates in the data (Figure 5.7 d).

After preprocessing, 147,913,800 reads were kept (38.2% of the original reads), with a total of 7,395,690,000 base pairs (Table 5.6). The sample with the highest read count had 7,261,353 reads (BC individual, qualifier HV), and that with the fewest had 3,457 reads (BC individual, qualifier CO). The mean quality score of all libraries was good, with all positions showing a quality score above 30 (Figure 5.7 a). The per base GC content graph had improved greatly, with the peaks contributed by the organellar reads disappearing after reads derived from these non-nuclear-genomes had been removed (Figure 5.7 b). Some bias and peaks were still observed, but further examination of the identity of these peaks revealed that they were mostly transposable elements in the nuclear genome, and were thus kept as they represent a part of the target genome to be mapped. The per base N content graph indicated that the N bases were no longer detected in the data (Figure 5.7 c). The sequence duplication level graph indicated that the highly represented sequences (possibly derived from the organellar genomes) had been removed by the preprocessing procedure, with some reads showing mostly >10 to < 500 duplicates being kept (Figure 5.7 d). The metrics of all libraries before and after preprocessing are summarised in Appendix 5.1 a.

Table 5.6 Number of reads before and after preprocessing

	Before preprocessing	After preprocessing	% retained
Total read count	386,334,623	147,913,800	38.3
Total base pairs (bp)	19,703,065,773	7,395,690,000	37.5

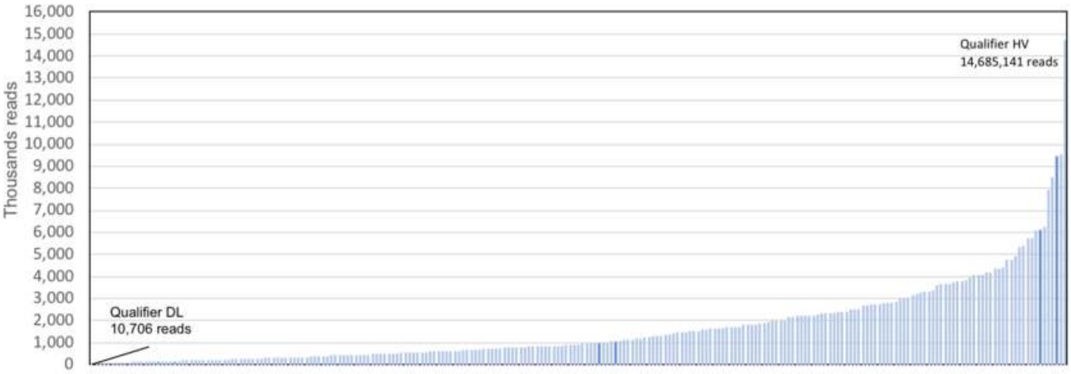


Figure 5.6 Read count obtained from all libraries sequenced, ordered from the lowest to the highest.

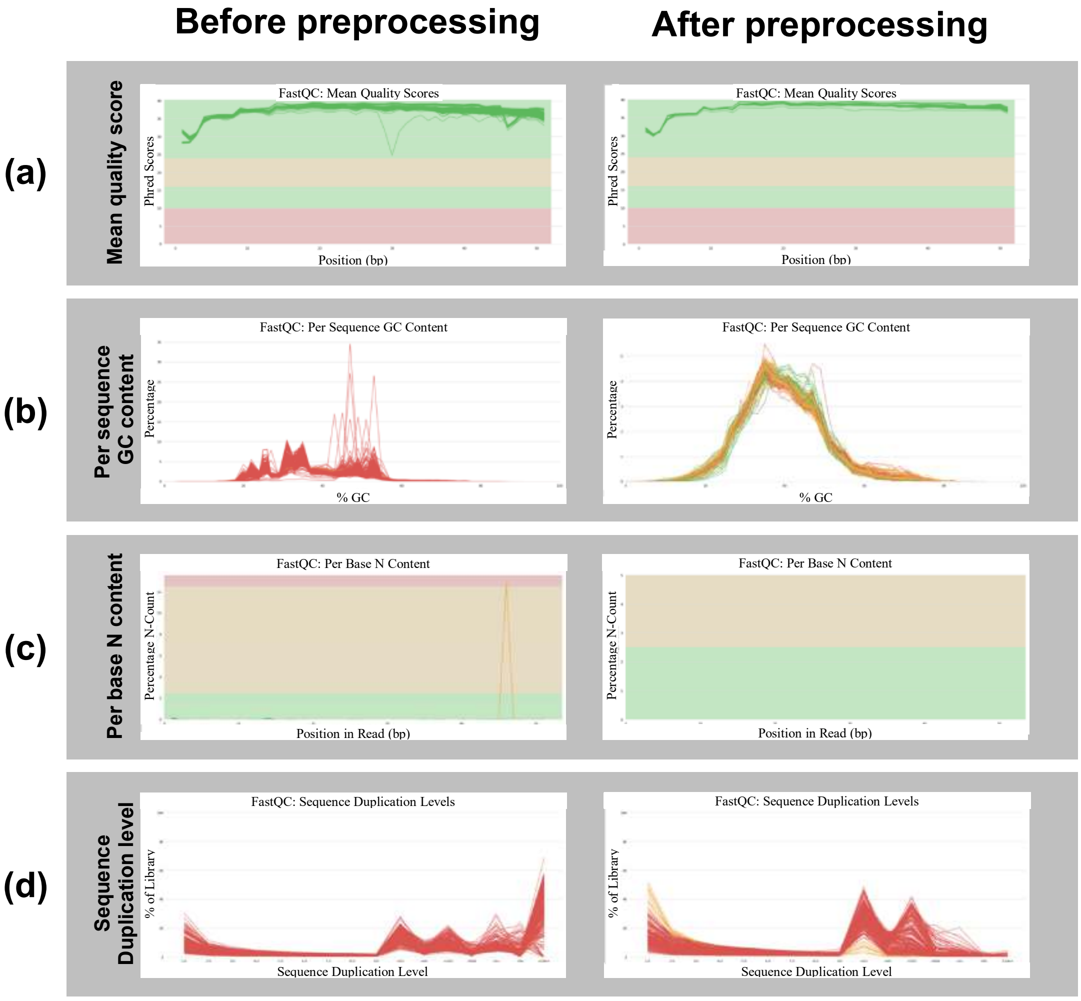


Figure 5.7 FastQC quality check result of all 237 libraries of RAD-Seq data, before and after preprocessing, with each line representing one sequencing library (one BC individual). **(a)** Mean quality score. **(b)** Per sequence GC content. **(c)** Per base N content. **(d)** Sequence duplication level.

5.3.2 MapA calculation – *de novo* approach

Three major parameters (m, M, and N) were optimised (Table 5.7). m = 3 and m = 4 gave highest number of usable markers (664 and 585 markers respectively). These two parameter values for m were chosen for the next step of optimisation (optimisation of within individual distance, M), where M = 1 and M = 2 gave the best results. These M values were chosen again for the optimisation of the last parameter (N), in which the final parameters were decided as (m=3, M=1, N=1) and the number of markers recovered (705 markers) is more than the default setting (Table 5.7; 702 markers).

A correlation was observed between the read count per individual and the number of marker recovered. With lower read count in an individual, the fewer markers can be recovered (Figure 5.8). Libraries with over one million reads usually had missing genotypes (i.e. undetermined genotype) in less than 20% of all the markers recovered (Figure 5.8). On the other hand, with lower than one million reads, the percentage of missing genotype increased almost exponentially. In individuals with lower than 500,000 reads, the percentage of missing genotype can increase to 60% to 70% (Figure 5.8). This indicated that including individuals with low read counts increased the proportion of markers with high levels of missing data that had to be filtered out, thus leading to a decrease in the number of mappable markers. Thus, as a trade-off between the number of individuals used for genetic map building and the number of recoverable markers, 200 BC individuals with the highest read counts were chosen for the calculation of the genetic maps for all the following analyses (i.e. MapA and MapB); i.e. the remaining 33 BC individuals with low read counts were excluded from the map calculation (see Appendix 5.1 b for the list selected individuals).

Table 5.7 Parameter optimisation for the *de novo* approach calculation of the MapA (data based on 50 BC plants)

Stacks parameters			Usable markers recovered*
Minimum stack depth (-m)	Within individual distance (-M)	Mismatch allowed for 2^{ndry} reads (-N)	
default (2)	default (2)	default (M+2=4)	702
Optimisation of minimum stack depth (-m)			
1	2	M+2	281
3	**2**	**M+2**	**664**
4	**2**	**M+2**	**585**
5	2	M+2	507
9	2	M+2	314
10	2	M+2	298
Optimisation of within individual distance (-M)			
3	**1**	**M+2**	**698**
3	**2**	**M+2**	**664**
3	3	M+2	639
3	4	M+2	614
3	5	M+2	597
4	1	M+2	623
4	2	M+2	585
4	3	M+2	563
4	4	M+2	536
4	5	M+2	518
Optimisation of mismatch allowed for merging secondary reads (-N)			
3	1	M+2	698
3	1	M+1	702
3	**1**	**M**	**705**
3	1	0	673
3	2	M+1	666
3	2	M	668
4	1	M+2	623
4	1	M+1	626
4	1	M	637
4	1	0	633
4	2	M+1	588
4	2	M	592

* Markers genotyped in at least 40 out of the 50 BC individuals used for optimisation, i.e. 80% of the individual. Bold text indicates the best results achieved in each step of the optimisation. The parameter showing the overall highest marker counts is highlighted in grey

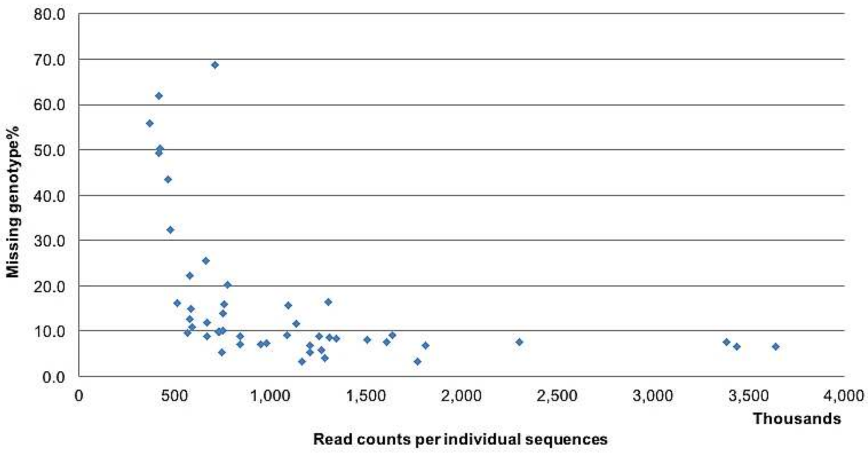


Figure 5.8 Correlation between read count per individual and proportion of missing genotypes among markers. Each dot represents the library of one BC individual of the 50 chosen for the initial mapping experiments.

The *de novo* approach MapA was calculated under the optimised parameters using the data of 200 BC individuals. In total 1,361 markers were recovered, but only 62 markers remained after removing markers with more than 20% missing data and markers showing strong segregation distortion, with most of the markers removed due to a high proportion of missing data (Table 5.8; >20% missing data). Eventually, 10 linkage groups were identified, with a total span of 716 cM with 55 mapped markers (Figure 5.9 a). The longest linkage group (LG1) was 120.1 cM and the shortest (LG10) 16.5 cM. The average marker interval was 13 cM (Table 5.11).

To increase the number of markers recovered and improve the final map density, more individuals with lower read counts were removed, leaving 150 BC individuals with higher read counts. The recalculated linkage map (Table 5.8; see Appendix 5.1 b for the removed individuals) recovered 1,359 markers. Of these, 198 markers were kept after filtering (Table 5.8), and finally 16 linkage groups were identified, which is identical to the haploid chromosome number of the *S. rexii* and *S. grandis* (n=16). The map had a total span of 1,119.8 cM with 183 mapped markers (Figure 5.9 b). The longest linkage group (LG1) was 120.1 cM, and the shortest (LG16) 8.5 cM. The average marker interval was 6.1 cM (Table 5.9).

Table 5.8 Statistics of the *de novo* approach MapA calculation

	***De novo* MapA 200 BC plants**	***De novo* MapA 150 BC plants**
No. of BC plant used	200	150
Total number of reads (after preprocessing)	144,479,900	136,102,282
No. reads used for analysis	130,186,134	122,384,682
No. read used (%)	90.1	89.9
Stacks analysis		
Total number of marker recovered	1,361	1,359
Mean coverage of marker (×)	13.6	14.8
Marker filtering		
No. marker remained after missing genotype filtering*	121	311
No. marker remained after segregation distortion filtering	62	198
No. markers remained after identical marker filtering	62	198
Final map statistics		
No. of linkage group recovered	10	16
No. of mapped markers	55	183
Total map distance (cM)	716.0	1,119.8
Average distance between markers (cM)	13.0	6.1

* More than 20% missing genotype

Table 5.9 Linkage group statistic summary of *de novo* approach MapA

***De novo* MapA – 200 BC individuals**			
Linkage group	**No. marker**	**Total distance (cM)**	**Average marker interval (cM)**
LG1	10	120.1	13.3
LG2	9	91.5	11.4
LG3	3	84.6	42.3
LG4	8	84.0	12.0
LG5	5	75.0	18.8
LG6	5	68.0	17.0
LG7	4	60.5	20.2
LG8	4	58.8	19.6
LG9	4	56.4	18.8
LG10	3	16.5	8.3
TOTAL	55	715.4	13.2
***De novo* MapA – 150 BC individuals**			
Linkage group	**No. marker**	**Total distance (cM)**	**Average marker interval (cM)**
LG1	32	120.1	3.9
LG2	14	116.0	8.9
LG3	17	104.3	6.5
LG4	9	95.9	12.0
LG5	18	85.8	5.0
LG6	15	85.4	6.1
LG7	12	73.2	6.7
LG8	10	72.3	8.0
LG9	10	72.2	8.0
LG10	13	70.8	5.9
LG11	9	59.4	7.4
LG12	7	58.3	9.7
LG13	8	52.5	7.5
LG14	3	27.0	13.5
LG15	4	17.0	5.7
LG16	2	8.4	8.4
TOTAL	183	1,118.6	6.1

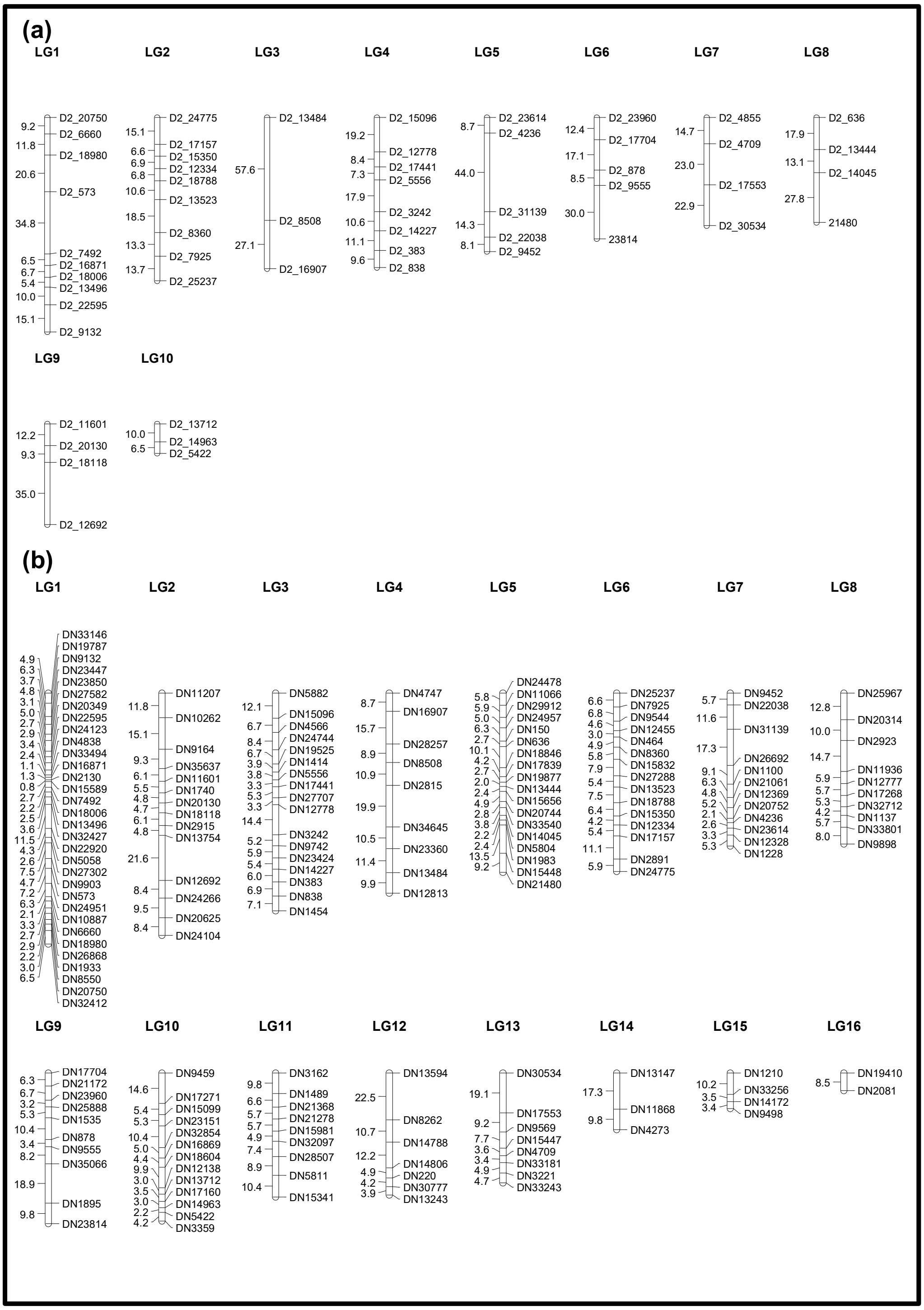


Figure 5.9 *De novo* approach MapA. **(a)** Map calculated based on 200 BC individuals. The marker name prefix ‘D2’ stands for *de novo* map calculated based on 200 individuals. **(b)** Map calculated based on 150 BC individuals. The marker interval distance is shown on the left (cM), and the marker name is shown on the right of each linkage.

5.3.3 MapA calculation – reference-based approach using BWA and Stampy aligners

The preprocessed RAD-Seq reads were mapped to the *S. rexii* SOAPdenovo2 assembly using BWA. In total, about 90 million reads originating from the 200 BC individuals plus parent plants were mapped, representing about 62.5% of the total input reads (Table 5.10). The Stacks analysis of the BAM files recovered a total of 3,751 markers. Amongst these, 414 markers were kept after filtering out markers with a high proportion of missing data or showing strong segregation distortion. Finally, 317 markers were mapped across 16 linkage groups, with a total span of 1,468.6 cM (Table 5.10, Figure 5.10). The longest linkage group (LG1) was 129.1 cM, and the (LG16) 15.0 cM. The average marker interval was 4.6 cM (Table 5.11).

Table 5.10 Statistics of the reference-based BWA approach MapA calculation (data based on 200 BC plants)

	BWA-MapA
No. of BC plant used	200 plants
Total number of reads (after preprocessing)	144,479,900
No. mapped reads	90,233,330
Mapped read (%)	62.5
Stacks analysis	
Total number of marker recovered	3,751
Mean coverage of marker (×)	23.3
Marker filtering	
No. marker remained after missing genotype filtering	699
No. marker remained after segregation distortion filtering	414
No. markers remained after identical marker filtering	414
Statistics of the final map	
No. of linkage group recovered	16
No. of mapped markers	317
Total map distance (cM)	1,468.6
Average distance between markers (cM)	4.6

Table 5.11 Linkage group statistic summary of reference-based BWA approach MapA (data based on 200 BC plants)

Linkage group	No. marker	Total distance (cM)	Average marker interval (cM)
LG1	45	129.1	2.9
LG2	21	128.9	6.4
LG3	17	115.6	7.2
LG4	29	115.0	4.1
LG5	22	112.4	5.4
LG6	21	111.3	5.6
LG7	23	96.8	4.4
LG8	17	95.7	6.0
LG9	26	95.1	3.8
LG10	10	94.7	10.5
LG11	14	93.7	7.2
LG12	23	92.2	4.2
LG13	26	87.4	3.5
LG14	15	69.8	5.0
LG15	4	16.1	5.4
LG16	4	15.0	5.0
TOTAL	317	1,468.8	4.6

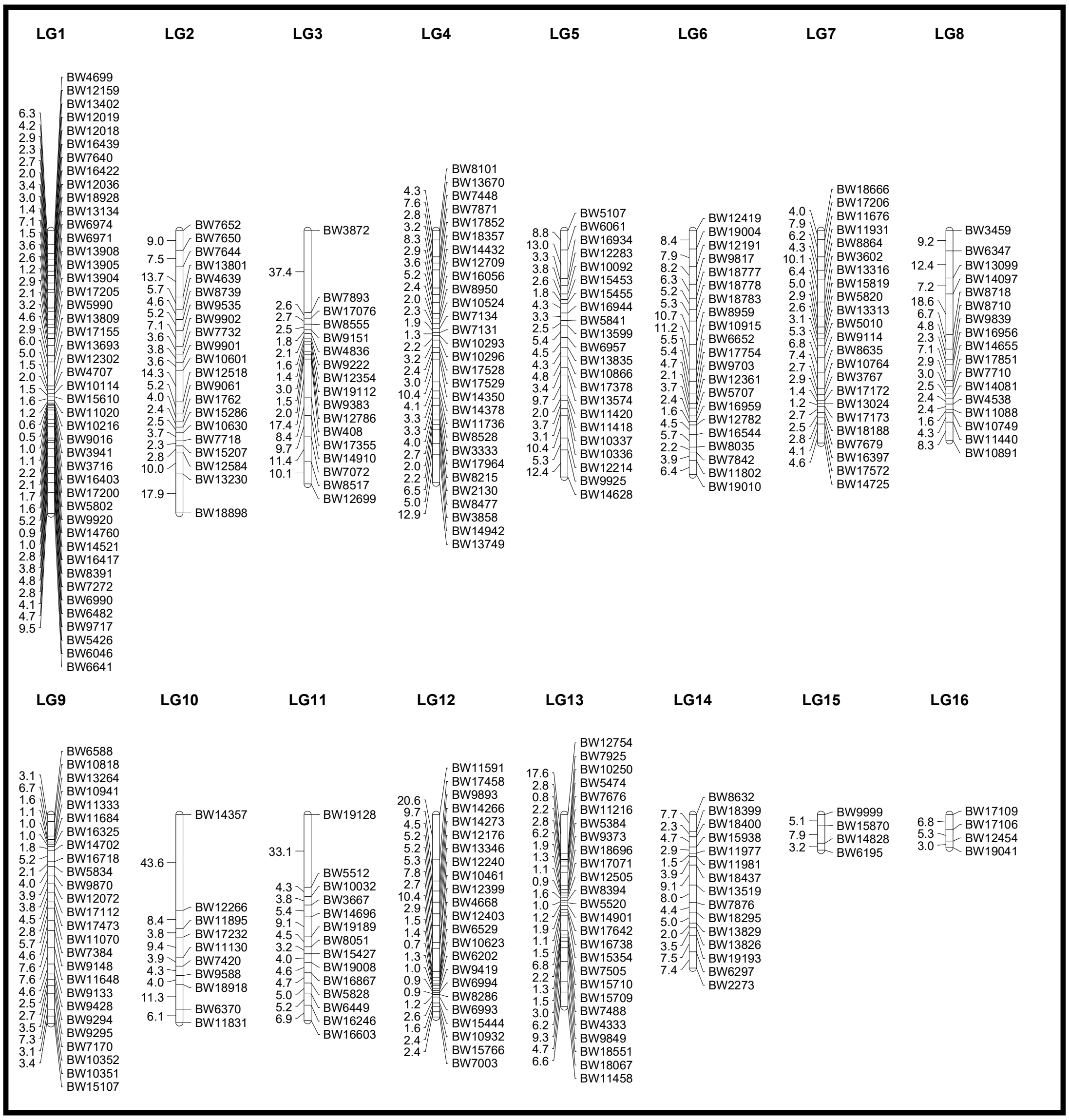


Figure 5.10 Reference-based BWA approach MapA. The marker interval distance is shown on the left (cM), and the marker name is shown on the right of each linkage group.

A second reference-based approach map was constructed by mapping the RAD-Seq data to the *S. rexii* SOAPdenovo2 preliminary genome assembly using the software Stampy. Among the 144 million preprocessed reads, about 134 million were mapped representing 93.4% of the total reads (Table 5.12). From these, a total of 9,185 markers were recovered. However, a large proportion of these markers were removed by filtering and only 503 markers remained. Among these, 16 linkage groups were identified with 338 mapped markers with a total span of 1,567.4 cM (Table 5.12; Figure 5.11). The longest linkage group (LG1) was 154.9 cM, and the shortest (LG16) 21.8 cM, The average marker interval was 4.6 cM (Table 5.13).

Table 5.12 Statistics of the reference-based Stampy approach MapA calculation (data based on 200 BC plants)

	Stampy-MapA
No. of BC plant used	200 plants
Total number of reads (after preprocessing)	144,479,900
No. mapped reads	134,888,035
Mapped read (%)	93.4
Stacks analysis	
Total number of marker recovered	9,185
Mean coverage of marker (×)	22.2
Marker filtering	
No. marker remained after missing genotype filtering	853
No. marker remained after segregation distortion filtering	503
No. markers remained after identical marker filtering	503
Statistics of the final map	
No. of linkage group recovered	16
No. of mapped markers	338
Total map distance (cM)	1,567.4
Average distance between markers (cM)	4.6

Table 5.13 Linkage group statistic summary of reference-based Stampy approach MapA (data based on 200 BC plants)

Linkage group	No. marker	Total distance (cM)	Average marker interval (cM)
LG1	24	154.9	6.7
LG2	32	145.3	4.7
LG3	21	131.3	6.6
LG4	31	114.9	3.8
LG5	27	109.1	4.2
LG6	19	107.2	6.0
LG7	13	100.3	8.4
LG8	35	99.7	2.9
LG9	21	99.3	5.0
LG10	17	98.4	6.2
LG11	24	95.1	4.1
LG12	22	92.5	4.4
LG13	26	86.6	3.5
LG14	17	70.3	4.4
LG15	4	39.9	13.3
LG16	5	21.8	5.5
TOTAL	338	1,566.6	4.6

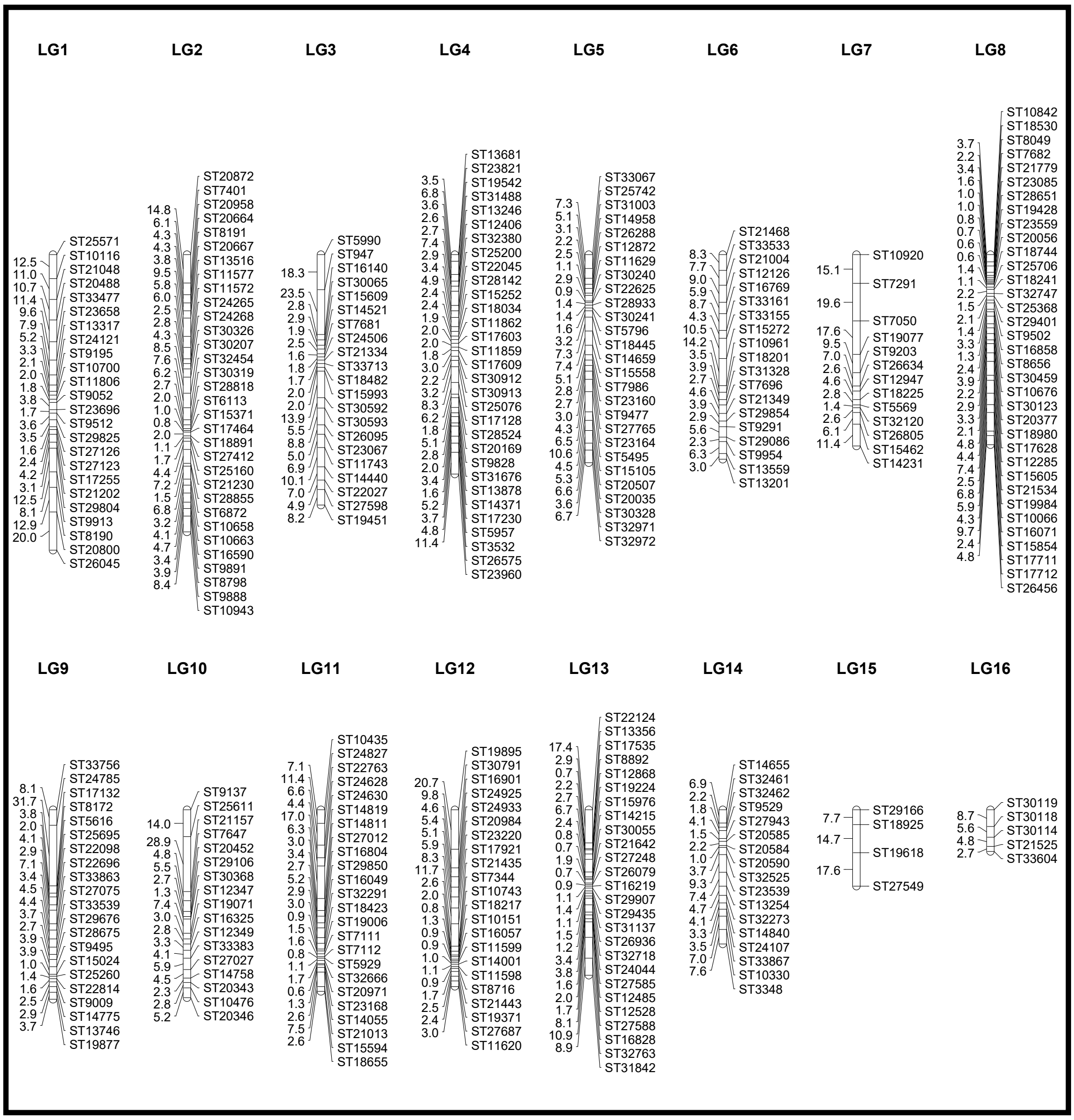


Figure 5.11 Reference-based Stampy approach MapA. The marker interval distance is shown on the left (cM), and the marker name is shown on the right of each linkage group.

The alignment percentage of the RAD-Seq reads of the two parents (*S. rexii* and *S. grandis*) was also examined. For the *S. rexii* data, both BWA and Stampy achieved over 90% of alignment coverage (Table 5.14). BWA aligned 1,230,052 reads and Stampy aligned 1,289,073 reads, corresponding to a 91.7% and 96.1% alignment coverage respectively. On the other hand, BWA struggled to align the *S. grandis* reads to the *S. rexii* genome assembly, with only 755,201 reads aligned (56.3%). This was improved when the software Stampy was used, resulting in 2,176,451 *S. grandis* reads aligned (91.3%) (Table 5.14).

Table 5.14 Comparisons of BWA and Stampy alignment percentages. The RAD-Seq reads of *S. grandis* and *S. rexii* were aligned to the *S. rexii* SOAPdenovo2 genome assembly

Aligner	***S. rexii***	***S. grandis***
	Aligned reads / Aligned reads %	**Aligned reads / Aligned reads %**
BWA	1,230,052 / 91.7%	755,201 / 56.3%
Stampy	1,289,073 / 96.1%	2,176,451 / 91.3%

The mismatch distribution of the alignments between BWA and Stampy were compared (Figure 5.12). The default BWA alignment settings only allowed up to 3 bp mismatches per 51 bp read (Figure 5.12 a), while the Stampy default settings allowed more than 5 bp mismatches per 51 bp read (Figure 5.12 b).

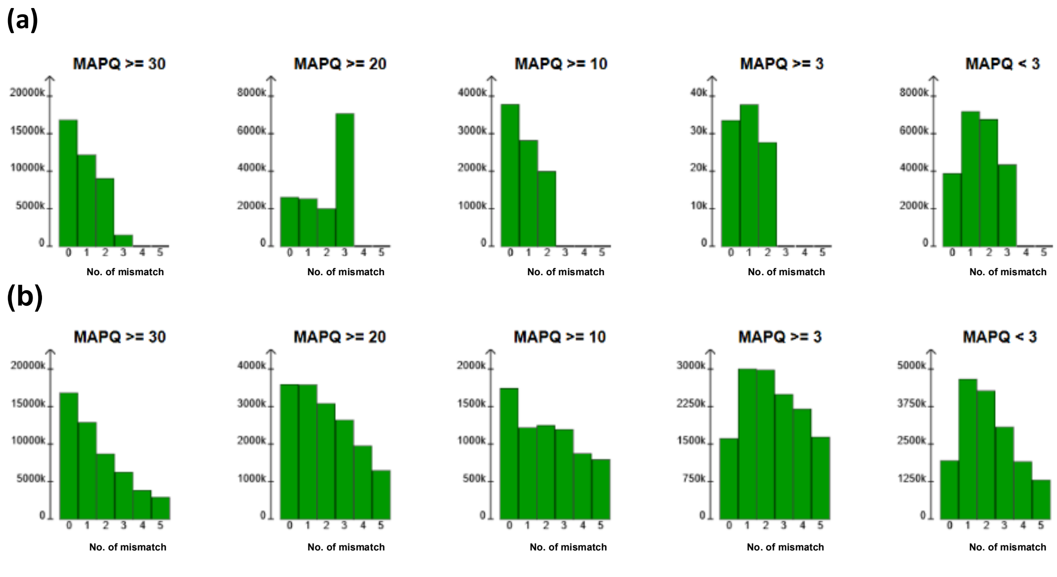


Figure 5.12 Mismatch distributions of BWA and Stampy alignments, with the X-axis showing the number of mismatches per read, and the y-axis showing the number of reads. **(a)** Mismatch distribution of BWA alignment. **(b)** Mismatch distribution of Stampy alignment. MAPQ: Mapping quality score.

Finally, the *de novo* approach and the reference-based Stampy approach were compared in terms of the read count per individual to the proportion of missing genotype, i.e. markers that failed to be genotyped. As shown previously in the *de novo* approach, the proportion of missing genotypes increased nearly exponentially when the read count per individual is below one million (Figure 5.13 blue dots). However, when the same data was analysed using the reference-based approach, more markers were recovered and the proportion of missing genotypes remained below 10% and rarely to below 20%, even when the number of read counts per individual was below 500,000 (Figure 5.13 red dots).

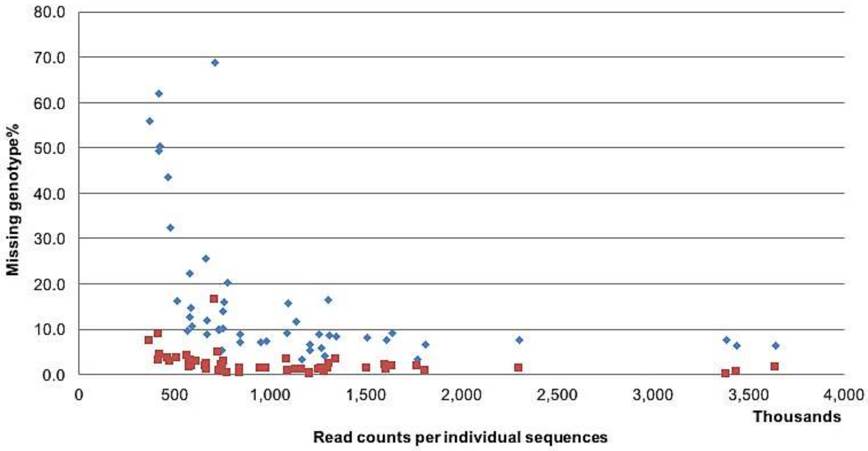


Figure 5.13 Correlation of proportion of missing genotypes to the number of reads per individual, where each dot represents one BC individual. Blue dots: data analysed using the *de novo* approach, as shown in Figure 5.8. Red dots: data analysed using the reference-based Stampy approach. Data for 50 BC individuals used in initial analyses are shown.

5.3.4 MapA calculation – combined approaches

To maximise the number of recovered markers and to increase the resolution of the genetic map, the genotype locus files generated from all three approaches (i.e. *de novo*, BWA and Stampy) were combined. The recalculated genetic map included 14,297 markers in the combined locus genotype file. When markers with more than 20% missing data were excluded, 1,673 markers were left after filtering, and when markers showing strong segregation distortion were also removed, 979 markers remained (Table 5.15). Among these, 180 BWA – Stampy marker pairs showed identical segregation patterns across the 200 BC individuals, and were also found to originate from the same locus in the *S. rexii* genome assembly and shared the same sequences (Appendix 5.2). The Stampy counterpart of these duplicated markers was excluded, which left 799 markers for the genetic map calculation (Table 5.15). On the other hand, no *de novo* approach-generated markers were found sharing identical segregation patterns to other markers. Eventually, 16 linkage groups were identified with 599 mapped markers (Table 5.15, Figure 5.14). The total span of the map was 1,578.2 cM. The longest linkage group (LG1) was 148.0 cM, and the shortest (LG16) 23.7 cM. The average marker interval was 2.6 cM (Table 5.16).

Table 5.15 Statistics of the MapA-combined approaches map calculation (data based on 200 BC plants)

	Combined MapA
Total no. of markers generated from all three approaches	14,297
Marker filtering	
No. marker remained after missing genotype filtering	1,673
No. marker remained after segregation distortion filtering	979
No. markers remained after identical marker filtering	799
Statistics of the final map	
No. of linkage group recovered	16
No. of mapped markers	599
Total map distance (cM)	1,578.2
Average distance between markers (cM)	2.6

Table 5.16 Linkage group statistic summary of combined approach MapA (data based on 200 BC plants)

Linkage group	**No. marker**	**Total distance (cM)**	**Average marker interval (cM)**
LG1	44	148.0	3.4
LG2	83	139.6	1.7
LG3	54	128.6	2.4
LG4	35	125.0	3.7
LG5	43	120.9	2.9
LG6	54	113.7	2.1
LG7	38	113.6	3.1
LG8	34	105.0	3.2
LG9	20	104.9	5.5
LG10	38	97.5	2.6
LG11	31	90.9	3.0
LG12	41	87.5	2.2
LG13	35	72.8	2.1
LG14	34	64.3	1.9
LG15	7	41.3	6.9
LG16	8	23.7	3.4
TOTAL	599	1,577.3	2.6

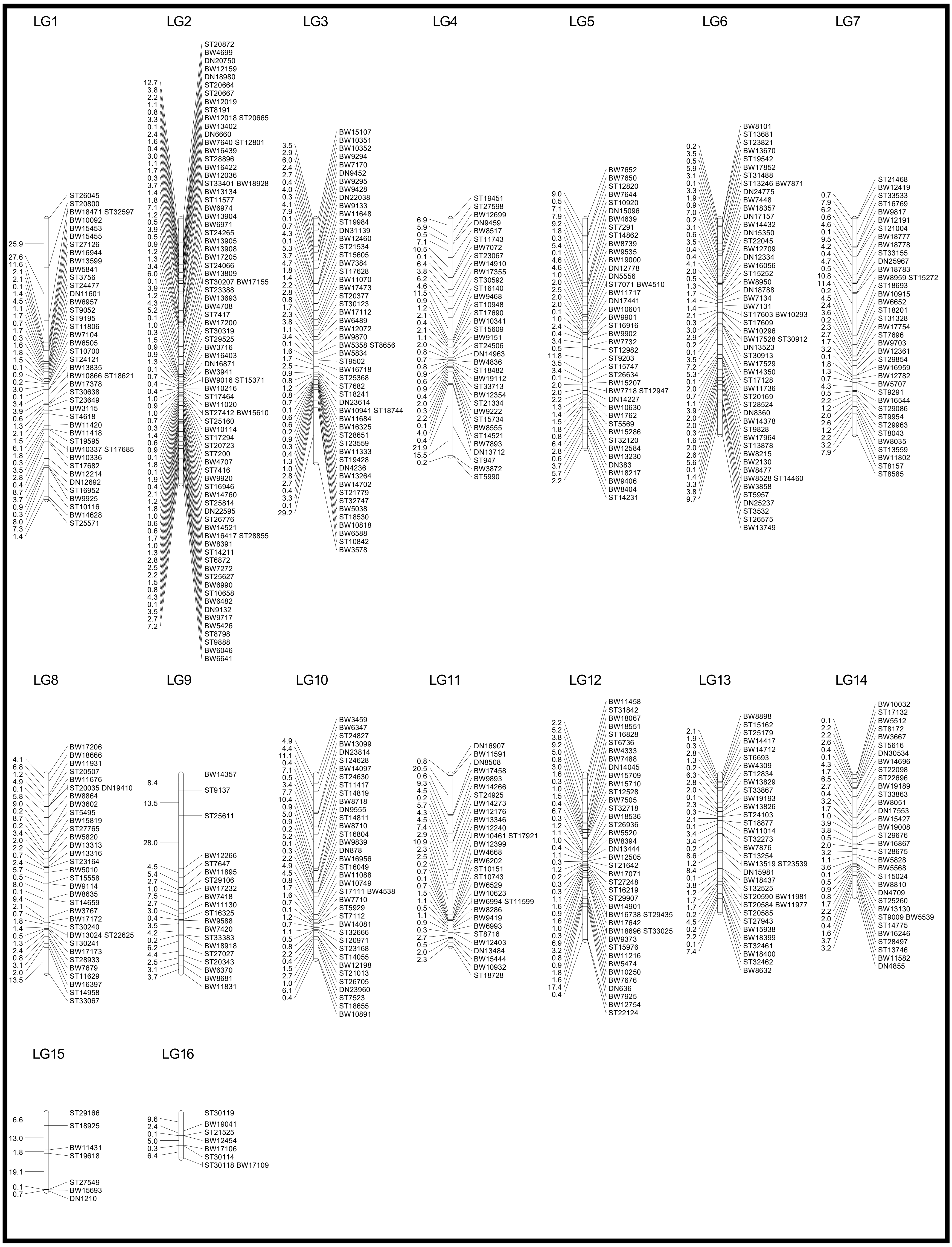


Figure 5.14 Combined approach-MapA. The marker interval distance is shown on the left (cM), and the marker name is shown on the right of each linkage group.

5.3.5 MapB calculation – *de novo* approach optimisation

Four *de novo* analysis-parameters were optimised for the calculation of MapB (Table 5.17). The optimisation results of m, M and N were the same as for the MapA-*de novo* approach, with m=3, M=1 and N=1 giving the highest number of markers (Table 5.17). The number of usable markers (genotyped in 40 out of 50 BC individuals) increased together with larger n values, but the number of mapped markers actually decreased when attempting to reconstruct the genetic map (Table 5.17; Appendix 5.3 a). The maximum markers density was achieved when n=1, with 362 markers mapped. On the other hand, when the n value increased to 8 and 16, only 341 and 337 markers were mapped respectively (Table 5.17; Appendix 5.3 b and c). Thus, the smallest n value of 1 was chosen, and the parameter setting of m=3, M=1, N=1, and n=1 to analyse and genotype the data of 200 BC individuals gave a total of 1,349 markers. The generated locus genotype file was kept for the calculation of the combined approach map.

Table 5.17 Summary of *de novo* analysis optimisation of MapB (based on 50 BC plants)

Stacks parameters				
Minimum stack depth (-m)	Within individual distance (-M)	Mismatch allowed for 2^{ndry} reads (-N)	Mismatch allowed for building catalog (-n)	Usable markers recovered *
default (2)	default (2)	default (M+2=4)	default (1)	531
Optimisation of minimum stack depth (-m)				
1	2	M+2	1	55
2 (default)	**2**	**M+2**	**1**	**531**
3	**2**	**M+2**	**1**	**531**
4	**2**	**M+2**	**1**	**519**
5	2	M+2	1	479
6	2	M+2	1	411
8	2	M+2	1	316
10	2	M+2	1	241
Optimisation of within individual distance (-M)				
2 (default)	**1**	**M+2**	**1**	**538**
2 (default)	2 (default)	M+2	1	531
2 (default)	3	M+2	1	468
2 (default)	4	M+2	1	442
2 (default)	5	M+2	1	423
3	**1**	**M+2**	**1**	**579**
3	2 (default)	M+2	1	531
3	3	M+2	1	502
3	4	M+2	1	476
3	5	M+2	1	458
4	**1**	**M+2**	**1**	**556**
4	2 (default)	M+2	1	519
4	3	M+2	1	490
4	4	M+2	1	464
4	5	M+2	1	448

Table 5.17 continued

Stacks parameters				
Minimum stack depth (-m)	Within individual distance (-M)	Mismatch allowed for 2^{ndry} reads (-N)	Mismatch allowed for building catalog (-n)	Usable markers recovered *
default (2)	default (2)	default (M+2=4)	default (1)	531
Optimisation of mismatch allowed for merging secondary reads (-N)				
2 (default)	1	M+2	1	538
2 (default)	1	M+1	1	545
2 (default)	1	M	1	561
3	1	M+2	1	579
3	1	M+1	1	590
3	**1**	**M**	**1**	**620**
4	1	M+2	1	556
4	1	M+1	1	562
4	1	M	1	592

Optimisation of mismatch allowed for building catalog (-n)

m	M	N	n	Usable markers recovered*	No. marker after filtering†	No. linkage group formed	No. of mapped marker
3	1	M	1	620	486	15	362
3	1	M	8	665	461	15	341
3	1	M	16	694	456	15	337

* markers genotyped in at least 40 out of the 50 BC individuals used for optimisation
† remove markers with strong segregation distortion and similar segregation pattern
Bold text indicates the best results achieved in each step of the optimisation. The parameter showing the overall highest marker counts is highlighted in grey

5.3.6 MapB calculation – reference-based approach using BWA and Stampy aligners

For the approach using the BWA aligner, two major BWA *aln* parameters were tested and optimised using the data of 50 BC individuals (Table 5.18). Different values of the maximum edited distance (-n) were first tested. The higher the value of -n, the more usable markers were recovered: the high maximum edit distance of 12 gave the highest mapping percentage of 76% of input reads (about 69 million reads), and recovered 1,943 usable markers (Table 5.18).

The maximum edit distance in seed (-k) was then tested. Again, the higher the k value the higher the mapping percentage and the number of usable markers recovered. The best result was achieved when using the setting n=12 and k=3, which generated 83% mapping coverage (about 75 million mapped reads) with 2,074 usable markers (Table 5.18).

An attempt was made to optimise an additional parameter, the seed length (-l). However, this leads to dramatic increases of the computational time required for the BWA mapping: for instant, with the setting of n=12, k=3, and l=6, our server only managed to process the reads from two BC individuals per day, i.e. about 7 million read per 24 hour. This analysis was performed on the local RBGE server 'Galvatron', which uses AMD Opteron 6176 SE processors. This was ran using 20 CPUs, suggesting that it would require about 480 CPU hours to align 7 million reads. With a total amount of 90 million reads from 50 BC individuals, it would take at least 6,171 CPU hours (~13 days) to test just one parameter value. If for example, four parameter values are to be tested on 50 BC individuals, it would take more than 50 days for just the mapping step (excluding the time required for genetic map calculation). And for the final mapping of approximately 148 million reads from 200 BC individuals, it would take another 10,149 CPU hours (~21 days) for just the mapping step. The total amount of time required for a proper optimisation plus the actual mapping of 200 BC individuals (>71 days), together with the time required for genetic map calculation, exceeds our available computational and time resources. Thus, the computational time becomes a limitation factor of the analysis, and we decided to proceed the analysis with the two optimised parameters (n and k).

To confirm that the optimised BWA parameters can improve the actual mapping results (i.e. increase the number of mapped markers), genetic maps calculated using the analysis results of the 50 BC individuals were compared with the map calculated using the default settings for these BC individuals (Table 5.18, last row; Appendix 5.4 a). The analysis with default BWA settings resulted in 1,273 markers on 15 linkage groups, and the analysis with optimised parameters of n=12 k=3 resulted in 1,537 mapped markers (Table 5.18, Final mapping results), 264 more than in the default analysis (Table 5.18, Appendix 5.4 b). This suggested that the optimisation does improve the mapping results. Thus, the parameter setting of n=12 k=3 was applied to the analysis of 200 BC individuals, which gave a total of 3,790 markers. The generated locus genotype file was kept for the calculation of the combined approach map.

Table 5.18 Optimisation of the BWA parameters for the reference-based approach of the MapB calculation (data based on 50 BC plants)

BWA parameters		No. reads mapped (% mapped)	Usable markers recovered*
Maximum edit distance (-n)	Maximum edit distance in seed (-k)		
default (3)	default (2)	59,952,269 (66%)	1,716
Optimisation of maximum edit distance (-n)			
3 (default)	2	59,952,269 (66%)	1,716
4	2	63,759,042 (70%)	1,809
6	2	67,263,341 (74%)	1,888
12	**2**	**69,166,302 (76%)**	**1,943**
Optimisation of mismatch allowed within seed (-k)			
12	1	57,110,681 (63%)	1,538
12	2 (default)	69,166,302 (76%)	1,943
12	**3**	**75,653,709 (83%)**	**2,074**

Final mapping results						
n	k	No. reads mapped (% mapped)	Usable markers recovered*	No. marker after filtering†	No. linkage group formed	No. of mapped marker
3 (default)	2 (default)	59,952,269 (66%)	1,716	1,306	15	1,273
12	**3**	**75,653,709 (83%)**	**2,074**	**1,572**	**15**	**1,537**

* Markers genotyped in at least 40 out of the 50 BC individuals used for optimisation

† Remove markers with strong segregation distortion and similar segregation pattern

For the reference-based approach using the Stampy aligner, the aligner was tested using the data of 50 BC individuals under default parameter settings. In total about 91% of the reads (about 83 million) were aligned to the genome, producing 2,115 usable markers (Table 5.19). Among these, 1,594 markers could be mapped on 16 linkage groups (Table 5.19, Appendix 5.5). Since the default Stampy settings already gave the best results among all three approaches tested (Table 5.20), the default settings were applied for the analysis of 200 BC individuals, which resulted in a total number of 4,043 recovered markers. The generated locus genotype file was kept for the calculation of the combined approach map.

Table 5.19 Analysis result of reference-based approach using Stampy (data based on 50 BC plants)

Parameter	**No. reads mapped (% mapped)**	**Usable markers recovered ***	**No. marker after filtering†**	**No. linkage group formed**	**No. of mapped marker**
Default	83,508,405 (91%)	2,115	1,644	16	1,594

* Markers genotyped in at least 40 out of the 50 BC individuals used for optimisation

† Remove markers with strong segregation distortion and similar segregation pattern

Table 5.20 Statistics of number of markers generated from *de novo*, BWA and Stampy approaches for the calculation of MapB. Calculation based on 50 BC individuals

Approach	Parameters used	Usable markers recovered*	Mapped markers

De-novo approach	m=3 M=1
	N=1 n=1
		620	362

Reference-based approach (BWA)	n=12 k=3	2,074	1,537
Reference-based approach (Stampy)	Default	2,115	1,594

* Markers genotyped in at least 40 out of the 50 BC individuals used for optimisation
† Remove markers with strong segregation distortion and similar segregation pattern

5.3.7 MapB calculation – combined approaches

The genotype locus files generated from *de novo*, BWA, and Stampy approaches were combined, giving a total 9,173 markers before filtering. For the calculation of MapB-1, a total of 801 among the 9,173 recovered markers were kept after filtering out markers with excessive missing data (Table 5.21; genotyped in less than 160 among the 200 BC individuals). Among these, 553 were kept after removing markers showing strong segregation distortion (Table 5.21). Finally, 17 linkage groups were identified, with 377 mapped markers and a total span of 1,144.2 cM (Table 5.21; Figure 5.15). The longest linkage group (LG1) was 136.0 cM, and the shortest (LG16) 10.7 cM. The average marker interval was 3.1 cM (Table 5.22).

For the calculation of MapB-2, a lower threshold (i.e. < 30%) of missing data was allowed, which left 1,572 markers after removing markers that were genotyped in less than 140 BC individuals (Table 5.21). Among these, 1,233 markers were kept after removing markers showing strong segregation distortion (Table 5.21). Finally 16 linkage groups were identified, with 836 mapped markers and a total map distance of 1,322.5 cM (Table 5.21; Figure 5.16). The longest linkage group (LG1) was 133.3 cM, and the shortest one (LG13) 51.1 cM, with an average marker interval of 1.6 cM (Table 5.23).

Table 5.21 Statistics summary of MapB-1 and MapB-2 (data based on 200 BC plants)

	MapB-1	**MapB-2**
Total no. of markers generated from all three approaches	9,173	9,173
Marker filtering		
No. marker remained after missing genotype filtering*	801	1,572
No. marker remained after segregation distortion filtering†	553	1,233
Stats of the final map constructed		
No. of linkage group recovered	17	16
No. of mapped markers	377	836
Total map distance (cM)	1,144.2	1,322.5
Average distance between markers (cM)	3.1	1.6

* MapB-1: Remove markers with <20% missing data, MapB-2: Remove markers with <30% missing data

† MapB-1: Remove markers with Chi-square value ≤ 0.0005, MapB-2: Remove markers with Chi-square value ≤ 0.0001

Table 5.22 Linkage group statistic summary of MapB-1 (data based on 200 BC plants)

Linkage group	**No. marker**	**Total distance (cM)**	**Average marker interval (cM)**
LG1	45	136.0	3.1
LG2	13	102.7	8.6
LG3	52	94.7	1.9
LG4	33	79.7	2.5
LG5	20	78.6	4.1
LG6	20	77.0	4.1
LG7	15	72.5	5.2
LG8	17	72.4	4.5
LG9	40	71.8	1.8
LG10	44	71.3	1.7
LG11	16	60.5	4.0
LG12	12	51.1	4.6
LG13	13	50.9	4.2
LG14	16	49.4	3.3
LG15	12	35.2	3.2
LG16	5	10.7	2.7
LG17	2	27.5	27.5
TOTAL	375	1,142.2	3.1

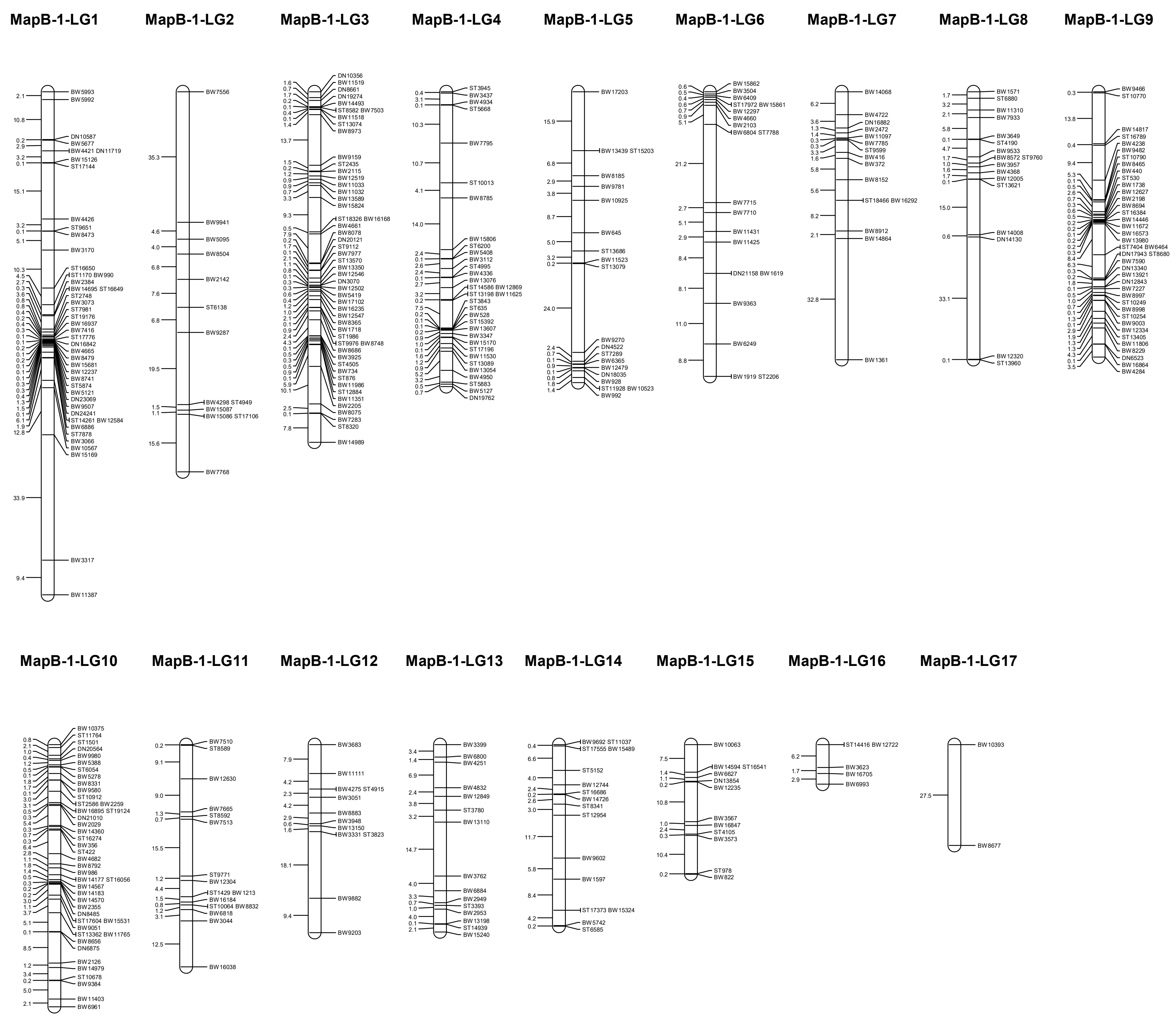


Figure 5.15 MapB-1. Marker interval distances are shown on the left (cM), and the marker names are shown on the right of each linkage group.

Table 5.23 Linkage group statistic summary of MapB-2 (data based on 200 BC plants)

Linkage group	**No. marker**	**Total distance (cM)**	**Average marker interval (cM)**
LG1	89	133.3	1.5
LG2	53	97.6	1.9
LG3	108	109.9	1.0
LG4	50	85.4	1.7
LG5	43	85.1	2.0
LG6	46	82.2	1.8
LG7	32	83.8	2.7
LG8	48	69.2	1.5
LG9	80	75.2	1.0
LG10	62	88.9	1.5
LG11	55	71.1	1.3
LG12	47	85.0	1.8
LG13	38	51.1	1.4
LG14	36	78.0	2.2
LG15	36	66.3	1.9
LG16	13	57.3	4.8
TOTAL	836	1,319.5	1.6

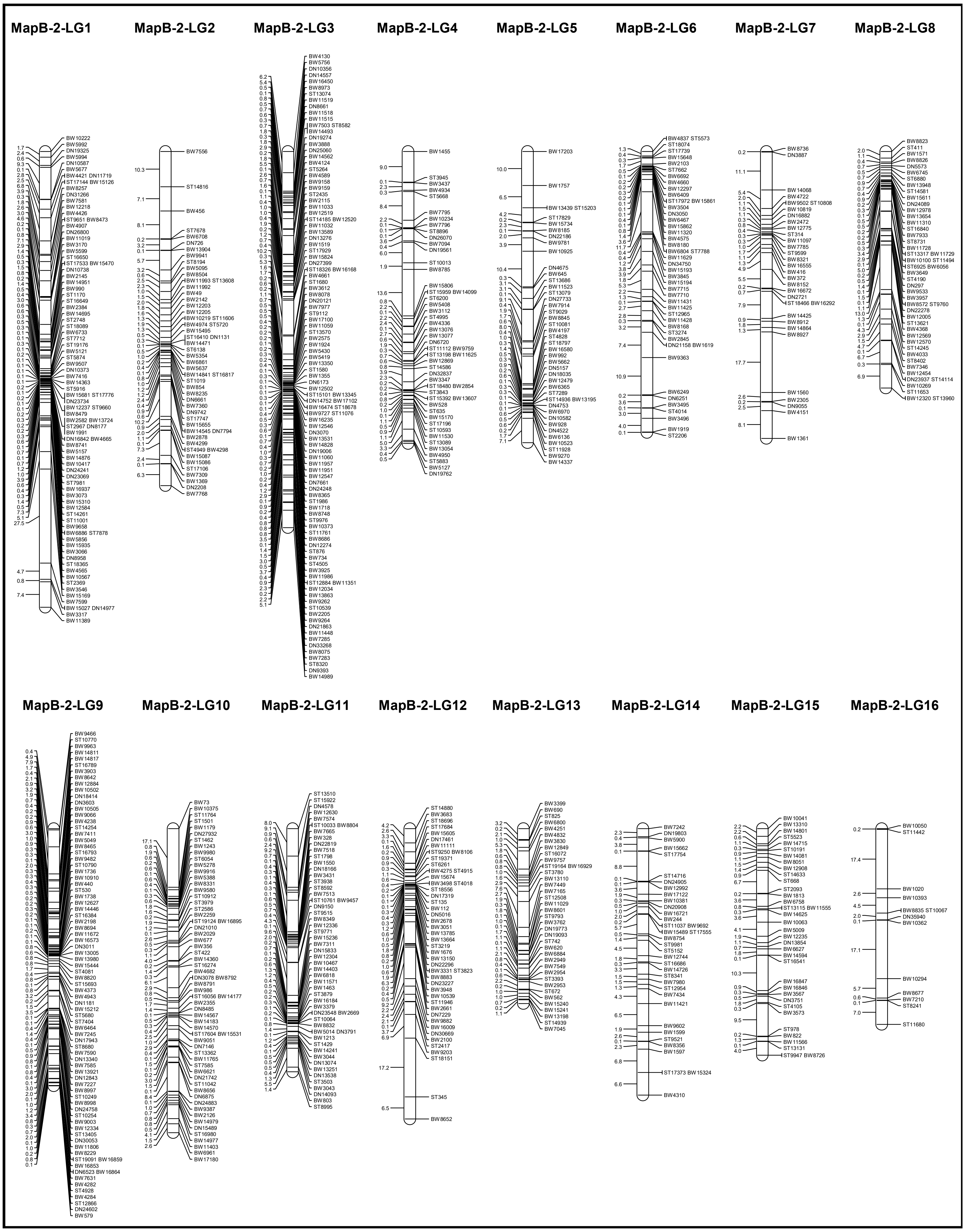


Figure 5.16 MapB-2. Marker interval distances are shown on the left (cM), and the marker names are shown on the right of each linkage group.

In the synteny analysis carried out between MapB-1 and MapB-2, a total of 17 markers were identified to be mapped in MapB-1, but not in MapB-2 (Table 5.24). These markers were located on MapB-1 LG1, LG2, LG8, LG10, LG11, LG14 and LG16. They were often not mapped in MapB-2 due to high mean Chi square values, or were sometimes not grouped during linkage group identification (Table 5.24). Interestingly, the MapB-1 LG16 was not identified in MapB-2 at all, while the MapB-1 LG17 was found to correspond to MapB-2 LG16 (Figure 5.17). In addition to these unmapped markers, inversions of local marker order were found in all linkage groups (Figure 5.17, red open boxes) except for LG2, LG7, LG14, LG16 and LG17. No markers were found to be mapped on different linkage groups between MapB-1 and MapB-2.

Table 5.24 List of markers that were mapped in MapB-1 but not in MapB-2

Marker name	LG	Reason of removal in MapB-2
BW5993	1	High mean Chi square contribution 6.558
BW11387	1	High mean Chi square contribution 13.489
BW9287	2	High mean Chi square contribution 8.256
BW14008	8	High mean Chi square contribution 5.686
DN14130	8	High mean Chi square contribution 6.568
ST10678	10	Discarded in round 3 mapping
BW9384	10	Discarded in round 3 mapping
BW7510	11	High mean Chi square contribution 9.596
ST78589	11	High mean Chi square contribution 8.841
BW16038	11	High mean Chi square contribution 22.178
BW5742	14	High mean Chi square contribution 5.395
ST6585	14	High mean Chi square contribution 7.285
ST14416	16	Excluded when grouping
BW3623	16	Excluded when grouping
BW16705	16	Excluded when grouping
BW6993	16	Excluded when grouping
BW12722	16	Excluded when grouping

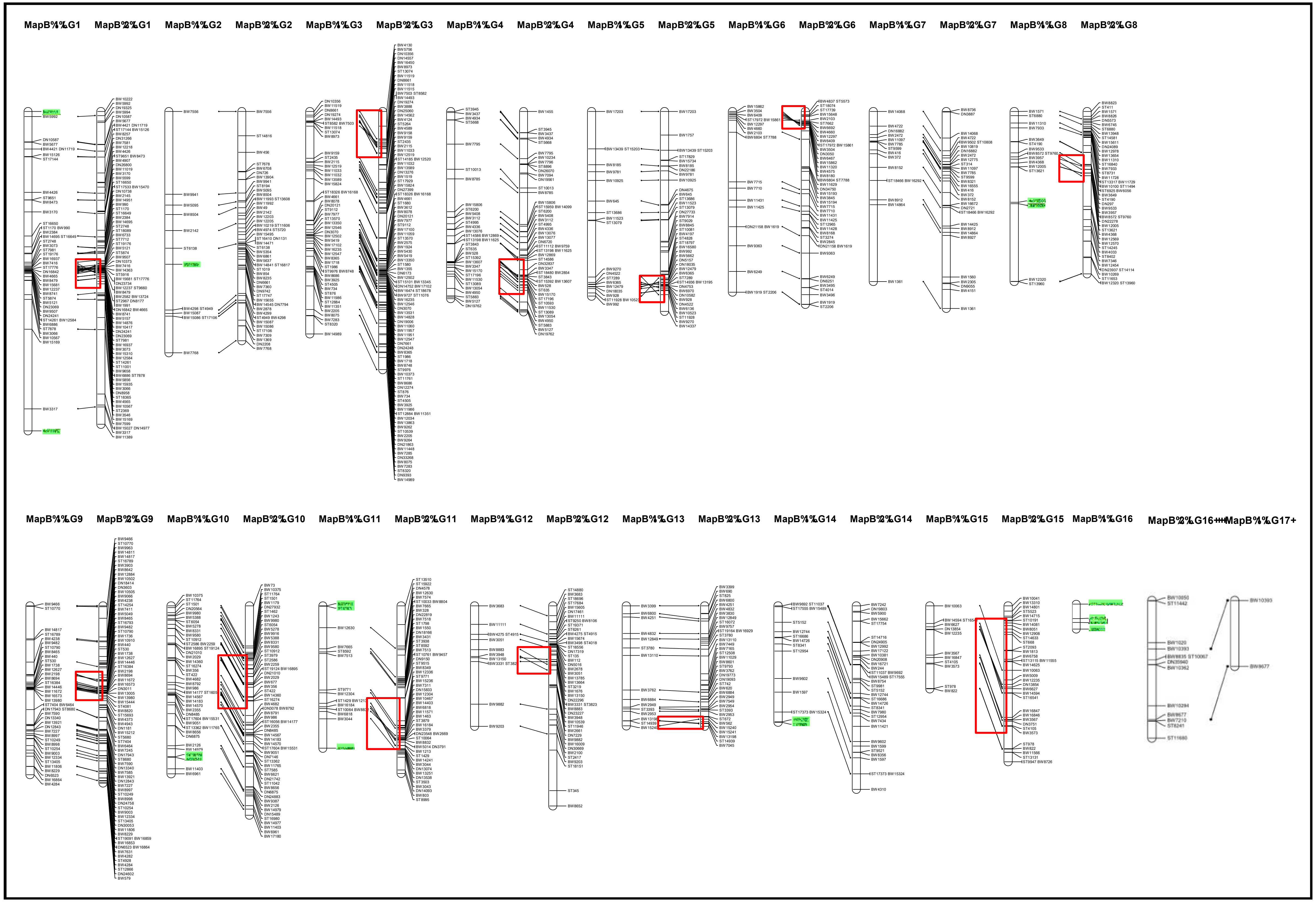


Figure 5.17 Synteny analysis between MapB-1 and MapB-2. Green highlight: markers mapped in MapB-1 but not MapB-2. Red open boxes: local marker rearrangement. Marker names are shown on the right of each linkage group.

The 17 markers identified in Table 5.25 were forcibly included for the calculation of MapB-3, and were kept regardless of their fitness, i.e. high mean Chi square and low N.Nfit values. The result was that MapB-3 had a total of 853 mapped markers distributed across 16 linkage groups (Table 5.25, Figure 5.18). The map spanned 1,389.9 cM. The longest linkage group (LG1) was 134.4 cM and the shortest (LG13) 51.3 cM, with an average marker interval of 1.6 cM (Table 5.26).

Table 5.25 Statistics of MapB-1, MapB-2 and MapB-3

	MapB-1	**MapB-2**	**MapB-3**
Total no. of markers generated from all three approaches	9,173	9,173	9,173
Marker filtering			
No. marker remained after missing genotype filtering*	801	1,572	1,572
No. marker remained after segregation distortion filtering	553	1,233	1,233
Statistics of the final map constructed			
No. of linkage group recovered	17	16	16
No. of mapped markers	377	836	853
Total map distance (cM)	1,144.2	1,322.5	1,389.9
Average distance between markers (cM)	3.0	1.6	1.6

* MapB-1: <20% missing, MapB-2: <30% missing

Table 5.26 Linkage group statistic summary of MapB-3

Linkage group	**No. marker**	**Total distance (cM)**	**Average marker interval (cM)**
LG1	91	134.4	1.5
LG2	54	97.8	1.8
LG3	108	109.9	1.0
LG4	50	85.4	1.7
LG5	43	85.1	2.0
LG6	46	82.2	1.8
LG7	32	83.8	2.7
LG8	50	69.7	1.4
LG9	80	75.2	1.0
LG10	64	90.2	1.4
LG11	58	73.0	1.3
LG12	47	85.0	1.8
LG13	38	51.1	1.4
LG14	38	78.2	2.1
LG15	41	130.7	3.2
LG16	13	57.3	4.8
TOTAL	853	1,389.9	1.6

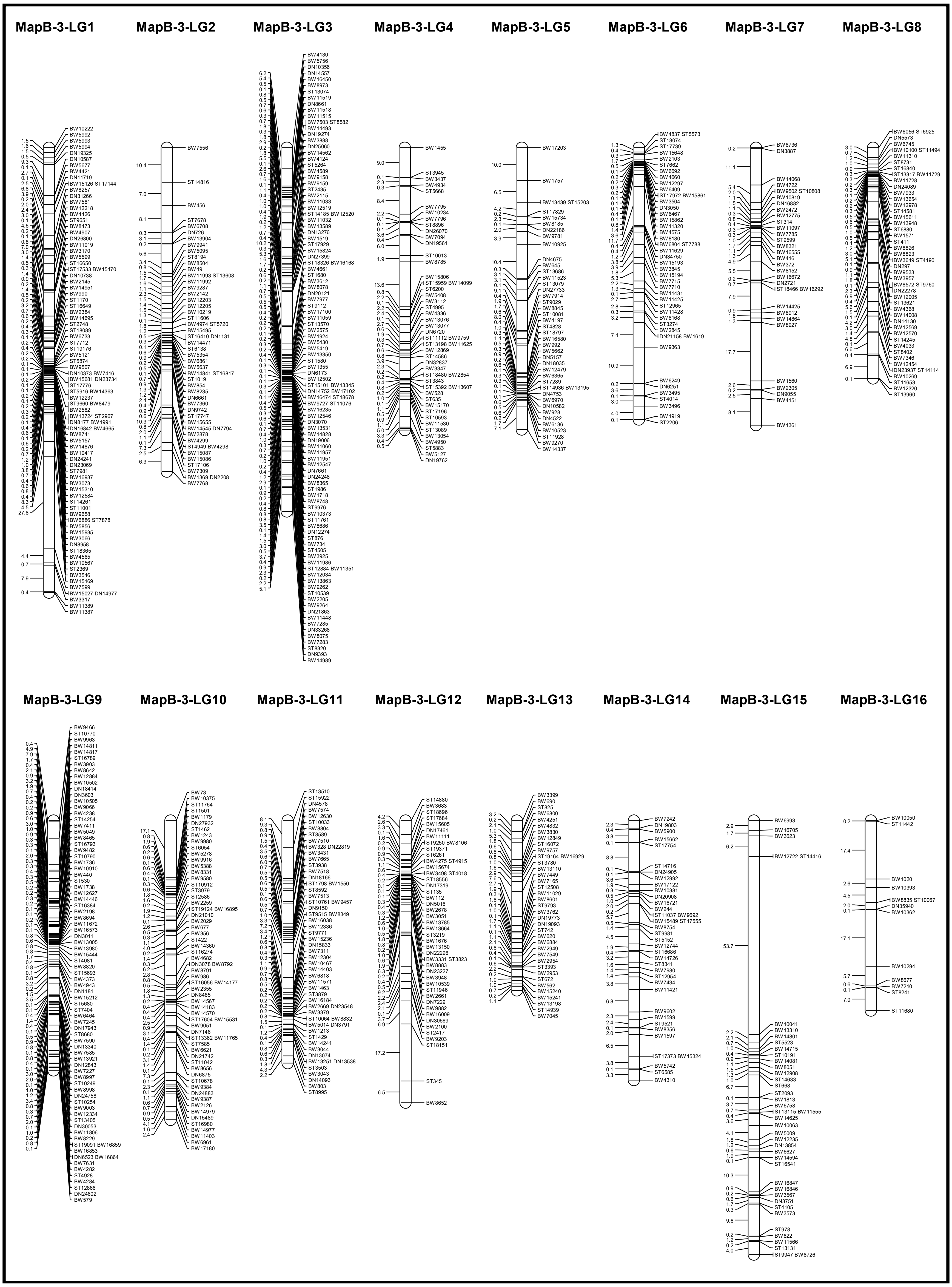


Figure 5.18 MapB-3. Marker interval distances are shown on the left (cM), and the marker names are shown on the right of each linkage group.

Close examination of the 17 markers forcibly added in the MapB-3 confirmed that they were mapped in the correct linkage groups as they were in MapB-1 (Table 5.29). However these markers had the high Chi-square values (>5), except for BW11387 in MapB-3 LG1, BW14008 in LG8, ST10678 and BW9382 in LG10, and the group of five markers in LG15 (Table 5.27). In addition, inclusion of these markers lowered the fitness of the surrounding markers in MapB-3 LG2, LG8, and LG15. For example, in MapB-3 LG2 the marker BW49 had a higher Chi square value of 6.117 than originally where it was below 5 in MapB-2 (Table 5.27). The same is the case in LG8, with the markers BW14008 and DN14130 that were well mapped in MapB-2, but now had a higher mean Chi square value of 5.222 in MapB-3 (Table 5.27).

Table 5.27 Markers added in the calculation of MapB-3

Marker name	LG in MapB-1	LG in MapB-3	Chi square value*	Note
BW5993	1	1	5.97	
BW11387	1	1	1.195	
BW9287	2	2	8.256	BW49 mean Chi square value became 6.117
BW14008	8	8	4.921	BW8823 mean Chi square value became 5.222
DN14130	8	8	6.547	
ST10678	10	10	3.412	
BW9384	10	10	3.594	
BW7510	11	11	7.221	
ST78589	11	11	6.464	
BW16038	11	11	23.665	
BW5742	14	14	5.2	
ST6585	14	14	5.512	
ST14416	16	15	0.143	MapB-1 LG16 is linked to MapB-2 LG15; MapB-1 LG17 is linked to MapB-2 LG16
BW3623	16	15	0.345	
BW16705	16	15	0.239	
BW6993	16	15	1.15	
BW12722	16	15	0.146	

* Indicates the Chi square value in MapB-3

5.3.8 Genetic map synteny analyses

For the comparisons between MapB-1 and MapB-3, an inversion of the marker order was observed in all but LG7, LG14, LG15, and LG16 (Figure 5.19). Interestingly, MapB-1 LG15 and MapB-1 LG16 were found to be linked together in MapB-3 as LG16, although with a relatively long marker interval of 54 cM (Figure 5.19). In addition, the two markers from MapB-1 LG17 were found to be mapped in MapB-3 LG16 (Figure 5.19). The relationship of corresponding linkage groups between MapB-1 and MapB-3 is summarised in Table 5.28.

Table 5.28 Relationship of linkage groups between MapB-1 and MapB-3

Linkage group in MapB-1	Corresponding linkage group in MapB-3
MapB-1_LG1	MapB-3_LG1
MapB-1_LG2	MapB-3_LG2
MapB-1_LG3	MapB-3_LG3
MapB-1_LG4	MapB-3_LG4
MapB-1_LG5	MapB-3_LG5
MapB-1_LG6	MapB-3_LG6
MapB-1_LG7	MapB-3_LG7
MapB-1_LG8	MapB-3_LG8
MapB-1_LG9	MapB-3_LG9
MapB-1_LG10	MapB-3_LG10
MapB-1_LG11	MapB-3_LG11
MapB-1_LG12	MapB-3_LG12
MapB-1_LG13	MapB-3_LG13
MapB-1_LG14	MapB-3_LG14
MapB-1_LG15	MapB-3_LG15
MapB-1_LG16	MapB-3_LG15
MapB-1_LG17	MapB-1_LG16

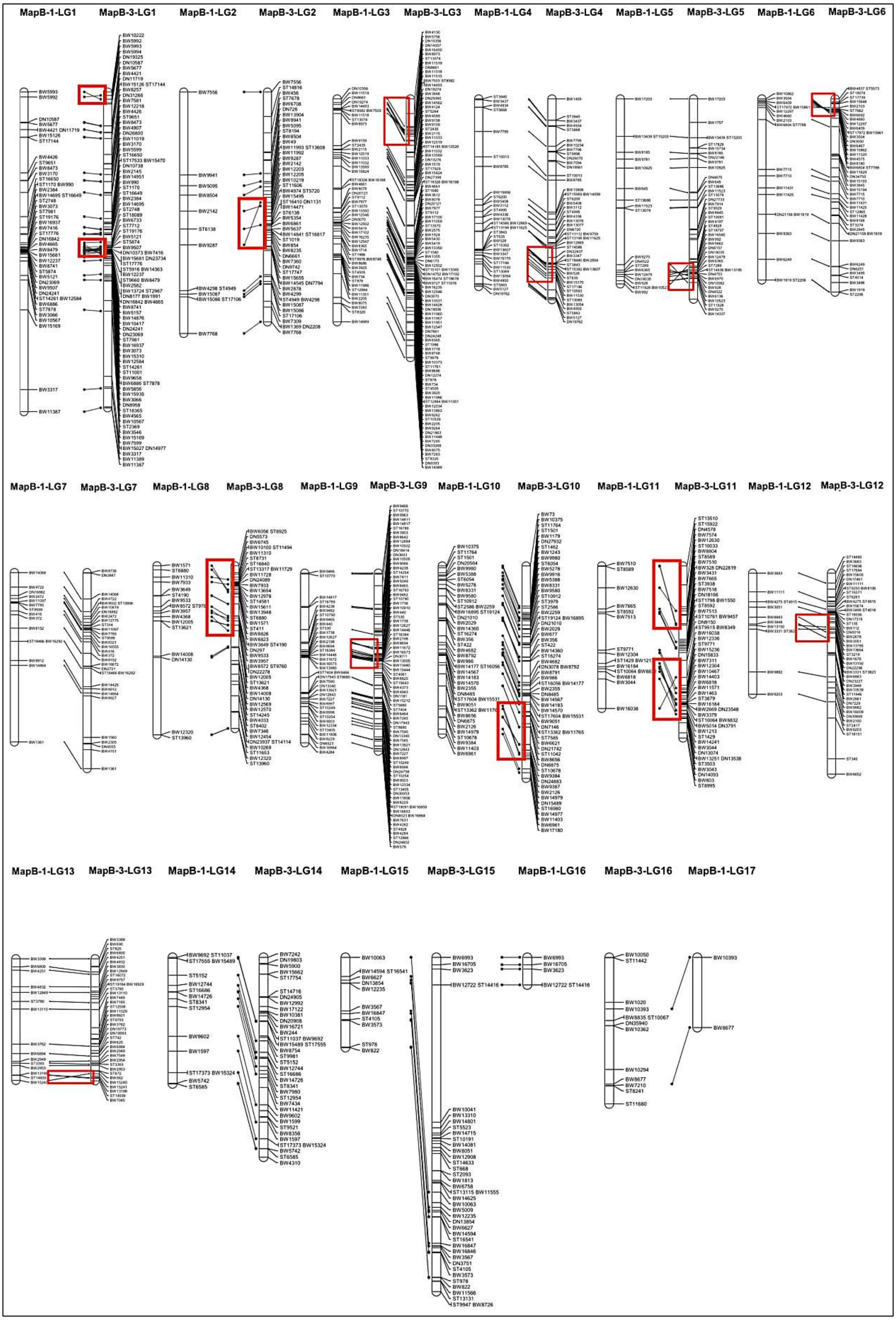


Figure 5.19 Result of the synteny analysis between MapB-1 and MapB-3. Marker names are shown on the right of each linkage group. Red rectangles: inversion of marker order.

For the comparison between MapB-1 and the combined approach-MapA, a total of 223 MapB-1 markers were found present in MapA (Figure 5.20). Inversion of marker orders were found in 10 linkage groups, on MapB-1 LG3 (MapA LG2), LG4 (MapA LG7), LG5 (MapA LG4), LG6 (MapA LG11), LG8 (MapA LG10), LG9 (MapA LG12), LG10 (MapA LG6), LG12 (MapA LG14), LG13 (MapA LG13), and LG14 (MapA LG9) (Figure 5.20). Markers in MapB-1 LG16 were not found in MapA at all, and the MapB-1 LG17 were found to be associated with MapA LG16 (Figure 5.20, Table 5.29).

Table 5.29 Relationship of the linkage groups between MapB-1 and combined approach-MapA

Linkage group in MapB-1	Corresponding linkage group in MapA
MapB-1_LG1	MapA_LG3
MapB-1_LG2	MapA_LG1
MapB-1_LG3	MapA_LG2
MapB-1_LG4	MapA_LG7
MapB-1_LG5	MapA_LG4
MapB-1_LG6	MapA_LG11
MapB-1_LG7	MapA_LG8
MapB-1_LG8	MapA_LG10
MapB-1_LG9	MapA_LG12
MapB-1_LG10	MapA_LG6
MapB-1_LG11	MapA_LG5
MapB-1_LG12	MapA_LG14
MapB-1_LG13	MapA_LG13
MapB-1_LG14	MapA_LG9
MapB-1_LG15	MapA_LG15
MapB-1_LG16	N/A
MapB-1_LG17	MapA_LG16

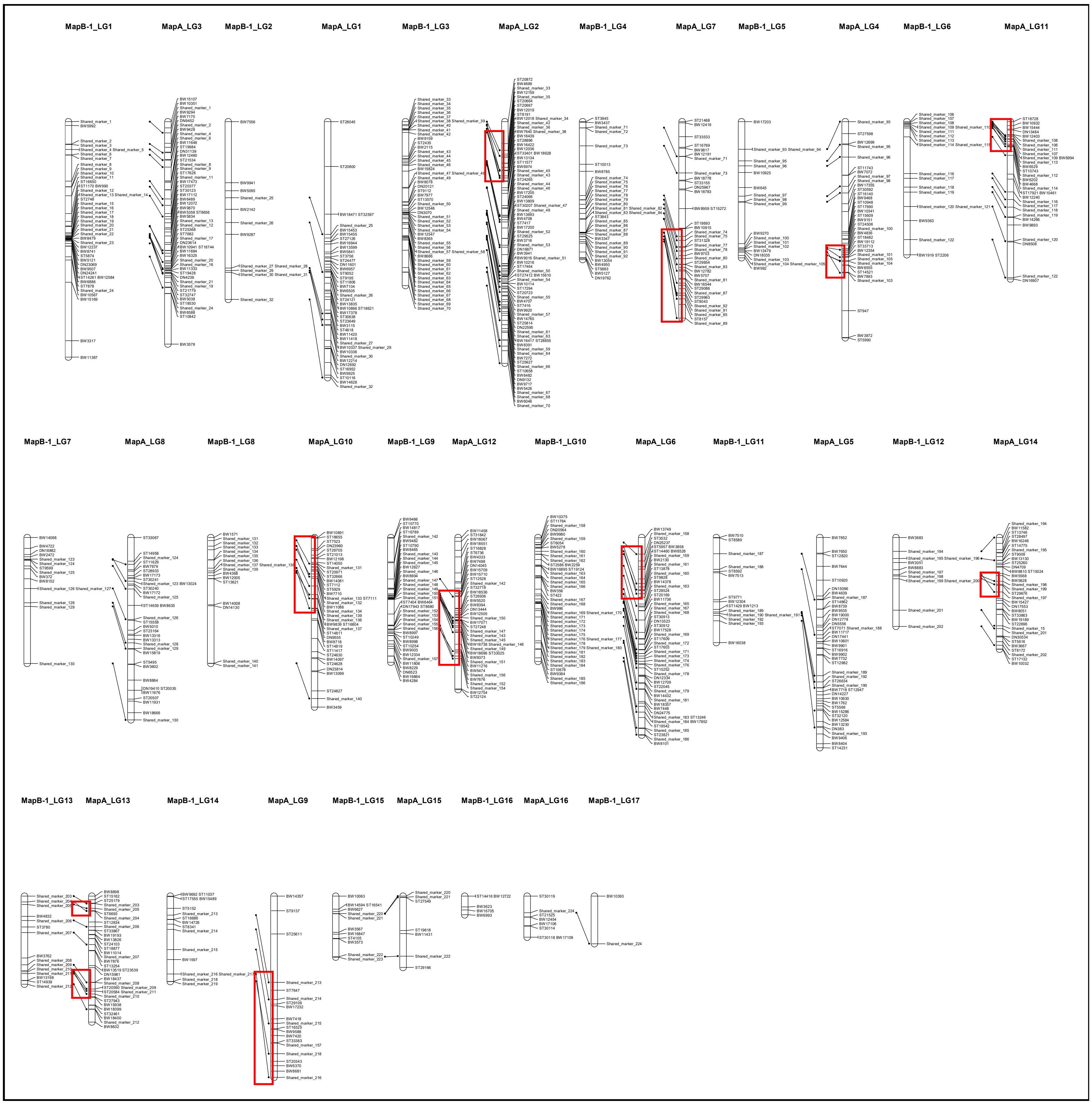


Figure 5.20 Result of the synteny analysis between MapB-1 and the combined approach-MapA. Marker names are shown on the right of each linkage group. Red open boxes: inversion of marker order.

5.4 Discussion

5.4.1 RAD-Seq and data preprocessing

The RAD-Seq read quality check results revealed multiple peaks in the per sequence GC content graph of all the libraries sequenced. Further investigation revealed that the observed GC distribution biases were derived from reads of organellar genome origin, which can be removed by mapping the reads to organellar genome assemblies. However, this leads to over 60% of reads lost after preprocessing (Table 5.6). One possible cause is a high content of chloroplasts and mitochondria in the leaf tissues used for DNA extraction. For instance, in *Arabidopsis* mesophyll cell it was estimated to have up to approximately 100,000 copies of plastid genomes per cell (200 chloroplasts with 500 copies of plastid DNA per chrloplast; Fujie et al., 1994; Pyke and Keech, 1994; Sakamoto et al., 2008). Hence, the organellar genomic DNA was likely to co-precipitate with the nuclear genomic DNA in our plant tissues, and was sequenced and present in the data. Since the aim of this project is the mapping of plant nuclear genome, these organellar genome-derived reads did not significantly contribute to this project and may be considered wasted. One possible improvement for future work in this area could be to keep the plant materials in a dark room prior to DNA extraction to reduce the reproduction of chloroplasts, thus reduce the chloroplast DNA content as well as secondary metabolites (Triboush et al., 1998; Waters and Langdale, 2009).

Another bias observed in the RAD-Seq data was the uneven read counts across the sequenced libraries (Figure 5.6). As shown in this study, libraries with lower read counts had much higher proportions of missing data (Figure 5.8), confirming previous studies (Catchen et al., 2011; Davey et al., 2012). Another possible consequence of insufficient sequencing depths could be genotyping errors, resulting in heterozygous loci being interpreted as homozygous, due to the insufficient depth of one of the allele or other (Davey et al., 2012). In this study, it was attempted to screen out correct and informative markers through various marker quality filtering steps (e.g. exclusion of individuals with very low read counts, removing markers with high proportions of missing data, removal of markers with excessive coverage; Miller et al., 2007; Catchen et al., 2011; Amores et al., 2011). The resultant genetic map here suggested that these attempts were successful in constructing 16 linkage groups with a high number of mapped markers and appropriate fitness values (i.e. Chi-sqaure and N.N.fit values).

5.4.2 Comparison between RAD-Seq data analysis approaches

In the MapA series, genetic maps were calculated based on 200 BC individuals using *de novo* approach, reference-based approach using the BWA aligner, and reference-based approach using the Stampy aligner. *De novo* approach generated 1,361 total markers, which is the least comparing to the other two reference-based approaches (3,751 and 9,185 markers for BWA and Stampy approaches, respectively). This could possibly be explained by the fact that reference-based approaches are better at recovering genotyping data from individuals with low read count (Shafer et al., 2016; Fountain et al., 2016). This is supported by the observation that in *de novo* approach, genotypes were difficult to be recovered from individuals with less than 750 million read counts, resulting in a high proportion of 20% to 70% missing genotypes (Figure 5.13). On the contrary, the reference-based approach using the Stampy aligner recovered more genotyping data, including individuals with less than 500 million read counts (Figure 5.13). The *de novo* approach map was also the least dense (Table 5.30) compared to other maps, and only 10 linkage groups were reconstructed instead of the expected 16. The overall results of less markers and linkage groups recovered suggested that the *de novo* approach alone was insufficient to generate a good genetic map from our dataset.

The two reference-based approaches also differed in their performance in terms of sequence alignment. Both aligners performed similarly when aligning *S. rexii* RAD reads to the *S. rexii* reference genome (Table 5.14; ~90% total reads). However, BWA struggled to align *S. grandis* RAD reads to the reference under default parameters, which resulted in only 56.3% of the total reads mapped (Table 5.14). In comparison, Stampy aligned 91.3% *S. grandis* RAD reads to the *S. rexii* reference genome, more than 1.5 times higher than BWA (Table 5.14). This may explain the overall lower alignment percentage of BWA (62.5% total reads) compared to Stampy (93.4% total reads) in the BWA approach MapA (Table 5.30). Since all BC individuals carry at minimum half of the *S. grandis* chromosomes, the generated RAD reads reflected this property and were thus difficult to map to the reference genome using BWA. The difference in BWA and Stampy alignment percentage also correlated to the number of markers recovered. In MapA, the BWA approach recovered 3,751 markers and the Stampy approach 9,185 markers (Table 5.30).

However, most of the 9,185 markers recovered in the Stampy approach were filtered out, and only 853 markers remained after removing markers with >20% missing data (9.2% of total markers). This suggested that many of the recovered markers were possibly errors, such as sequencing errors or contaminant sequences, which were only present in a few individuals but not in other BC individuals (Catchen et al., 2011). This is possibly due to the fact that Stampy allows more mismatches during the alignment process (Lunter and Goodson, 2011), and more RAD reads with sequencing errors were aligned and mis-judged as informative loci during Stacks analyses. These loci were eventually removed during marker filtering, as they do not constantly appear in all the BC individuals (i.e. high proportion of missing data), or do not follow the expected segregation ratio (segregation distortion). In the end, both BWA and Stampy approaches recovered 16 linkage groups and about 300 markers (Table 5.30). Both maps spanned around 1,500 cM long, and the average marker intervals are both 4.6 cM (Table 5.30). This result suggests that while the two aligners have very different alignment percentages, the number of markers recovered and the final genetic mapping is similar.

In the calculation of the MapA series, the combined approach was found to generate the genetic map with highest resolution and longest map distance (Table 5.30). As shown in the MapA calculation, the combined approach-MapA had 599 mapped markers spanning 1,578.2 cM, with an average marker interval of 2.6 cM, nearly twice as dense as the BWA or Stampy maps alone (Table 5.30; marker interval 4.6 cM in the maps of both approaches). Some *de novo* assembled markers were also mapped, suggesting that *de novo* approach may contribute to recovering markers outside the currently available genome assembly (Wang et al., 2013). Thus, the combined approach is valuable in recovering as many markers as possible, and this approach was taken for the calculation of the MapB series.

In MapB series maps, different marker filtering strategies were applied for the calculation of MapB-1, MapB-2, and MapB-3. The markers used for constructing MapB-1 was processed under stringent filtering, while the markers used for MapB-2 and MapB-3 were processed with a more relaxed filtering allowing higher proportion of missing data and segregation distortion (Table 5.5). This resulted in difference in number of mapped markers of each map, with in MapB-2 and MapB-3 more than twice the number of markers were mapped (836 markers and 853 markers, respectively) than in MapB-1 (377 markers). The low number of markers recovered in MapB-1 was most likely due to a high proportion of missing data (Table 5.30; before filtering 9,173 markers, after filtering 801 markers). Hence, by lowering the threshold of filtering markers with missing data, the genetic map density was improved in MapB-2 and MapB-3. However, using markers with excessive amounts of missing data (>20% missing) has been shown to result in ~50% chances to misplace the marker order or produce false linkages in simulation data (Hackett and Broadfoot, 2003). This suggests that the result of MapB-2 and MapB-3 should be treated carefully, and lowering the threshold for missing data is not a permanent solution to increase the map density as it also increases the chance of constructing incorrect genetic maps. Instead, resequencing of the samples with lower depth of coverage would be the optimal way to improve the map quality.

Table 5.30 Summary of the statistics of the main results of the genetic map reconstruction in this chapter

	De novo **MapA**	**BWA-MapA**	**Stampy-MapA**	**Combined-MapA**	**MapB-1***	**MapB-2***	**MapB-3***
No. of BC plant used	200 plants	200 plants	200 plants	200 plants	200 plants	200 plants	200 plants
Total number of read (after preprocessing)	144,479,900	144,479,900	144,479,900	-	-	-	-
No. analysed / mapped reads	130,186,134	90,233,330	134,888,035	-	-	-	-
No. analysed / mapped reads (%)	90.1	62.5	93.4	-	-	-	-
Stacks analysis							
Total number of marker recovered	1,361	3,751	9,185	-	9,173	9,173	9,173
Mean coverage of marker (×)	13.6	23.3	22.2	-	-	-	-
Marker filtering							
No. marker kept after missing genotype filtering	121	699	853	1,673	801	1,572	1,572
No. marker kept after segregation distortion filtering	62	414	503	979	553	1,233	1,233
No. markers kept after identical marker filtering	62	414	503	799	553	1,233	1,233
Final map statistics							
No. of linkage group recovered	10	16	16	16	17	16	16
No. of mapped markers	55	317	338	599	377	836	853
Total map distance (cM)	716.0	1,468.6	1,567.4	1,578.2	1,144.2	1,322.5	1,389.9
Average distance between markers (cM)	13.0	4.6	4.6	2.6	3.0	1.6	1.6

* - MapB-1, MapB-2 and MapB-3 were all calculated based on combined approach

5.4.3 Comparison between MapA and MapB maps

The genetic maps (MapA and MapB) calculated in this study show different advantages: the combined approach-MapA had the longest total distance with 1,578.2 cM; the MapB-3 had the highest number of 853 mapped markers and highest resolution (average marker interval = 1.6 cM); the MapB-1 may have the most reliable marker order and sequences as it was constructed under the most stringent conditions (i.e. usage of the filtered *S. rexii* genome and the stringent marker filtering strategy).

A major difference between the three maps concerned the number of linkage groups they recovered. Combined approach-MapA and MapB-3 both had 16 linkage groups, consistent with 16 pairs of chromosomes of *Streptocarpus* (Figure 5.14, Figure 5.18). On the other hand MapB-1 showed 17 linkage groups, with a small LG17 that contained only two markers (Figure 5.15, Table 5.22). The synteny analysis between MapB-1 and MapB-3 showed that MapB-1 LG17 actually corresponded to MapB-3 LG16, and MapB-1 LG16 was linked to LG15 (Figure 5.19, Table 5.28). The same analysis between MapB-1 and combined approach-MapA suggested that LG17 corresponded to MapA LG16, while the markers on MapB-1 LG16 could not be identified in combined approach-MapA (Figure 20, Table 5.29). It can be speculated that the stringent marker filtering criteria of MapB-1 filtered out too many markers, and those which linked the LG17 to other linkage groups. Such linkages were supported in combined approach-MapA and MapB-3, suggesting that these two maps may be better to illustrate the actual linkage groupings in this genome area.

A second difference observed among the diverse maps concerned the number of markers recovered. Comparing between MapB-1 and combined approach-MapA, MapB-1 only had 377 markers mapped while in combined approach-MapA 599 markers were mapped (Table 5.30). As these two maps were constructed using the marker filtering strategies of same stringency (i.e. removing markers with >20% missing data and removing markers showing strong segregation distortion), it is likely that the difference came from (1) the *S. rexii* reference genome used, which in combined approach-MapA is the unfiltered genome assembly that was used as reference, and in MapB-1 the contaminant-filtered genome assembly was used. (2) In combined approach-MapA the default BWA mapping parameters were used, while in MapB-1 the *de novo* and BWA mapping parameters were further optimised. As shown in this study that the optimisation of the *de novo* analysis and BWA alignment parameters were proven to improve the map marker density (Table 5.17 and Table 5.18, respectively), a more possible explanation is that some of the markers recovered in combined approach-MapA are actually contaminant sequences which inflated the marker density of combined approach-MapA. An examination of the combined approach-MapA markers could be done, by performing BLAST searches of the marker sequences against the nucleotide database (nt) of NCBI to identify possible contaminant sequences.

The third difference observed concerned the order of markers. Inversions of marker orders were observed in most linkage groups between MapB-1 and MapB-3, and between MapB-1 and combined approach-MapA (Figure 5.19 and Figure 5.20). This is possibly a result of the genetic mapping algorithm chosen here, i.e. regression mapping. In regression mapping, the order of markers is determined by minimising the sum of the squared deviation of the distance between two adjacent markers (Van Ooijen and Jansen, 2013). In other words, the marker order that generates the shortest linkage group is favoured, implying that the determination of marker order may change every time a new marker is added (Van Ooijen and Jansen, 2013). It is thus unsurprising that the three maps discussed here show marker inversions. However, it should be noted that marker order in MapB-3 might be the least reliable among the three maps due to its' less-stringent marker filtering and forced addition of some markers for map calculation: in MapB-3 the forcibly added markers on LG1, LG2, LG8, LG11, and LG14 showed >5 Chi-square goodness-of-fit values (Table 5.27), implying that the quality of the map was lower (Van Ooijen and Jansen, 2013). Moreover, it is known that inclusion of markers with excessive amounts of missing data (>20% missing) during map calculation leads to incorrect map ordering and overestimation of map distances (Hackett and Broadfoot, 2003), which is the case of MapB-3 as it included markers with up to 30% missing data in its' calculation.

On the other hand, there was no marker found to be grouped incongruently in the synteny analyses among the three genetic maps discussed here (i.e. grouped in different linkage groups between two maps analysed). This implies that the overall marker grouping is highly reliable.

5.4.4 Difficulties in reconstructing linkage groups LG15 and LG16

Amongst all the genetic maps calculated, there were always one or two linkage groups that were difficult to reconstruct. These are the LG15 and LG16 of all the MapA maps and in MapB-1, which were always shorter than 50 cM and contained fewer than 10 markers. One possible explanation for this poor mapping results is the evolutionary distance between *S. rexii* and *S. grandis* (Nishii et al., 2015). Recombination suppression is known to occur between species that have undergone chromosomal rearrangements, and for recombination to happen the two genomes should show similarities in gene order or chromosome homology (Jackson, 2011; Ren et al., 2018). Even though *S. rexii* and *S. grandis* are from sister clades, the BWA show poor alignment percentage when aligning *S. grandis* RAD-Seq reads to the *S. rexii* genome (Table 5.14). This suggests that the genome sequences between the two species share high proportion of heterologous sequences, which may contributed to lesser frequency of recombination hence the chromosomes can only be mapped partially.

On the other hand, MapB-2 and MapB-3 both showed better resolved LG15 and LG16 with longer than 50 cM genetic distance and contained more than 10 markers (Table 5.23, Table 5.26). The difference between MapB-2, MapB-3 and the MapA, MapB-1 is that the former two maps allowed higher proportion of missing genotypes (up to 30% missing). It can thus be speculated that the markers on LG15 and LG16 are actually presented in our RAD-Seq data, but were not mapped in MapA and MapB-1 due to high proportion of missing genotypes. The most plausible reason for having missing genotypes is the low read counts in many of the libraries sequenced (Appendix 5.1 a; Catchen et al., 2011; Davey et al., 2012). (Hackett and Broadfoot, 2003). By performing additional RAD-Seq experiments, it can be possible to increase the read counts per libraries, and in turn improve the genotyping result and the number of markers recovered to increase the resolution of LG15 and LG16.

5.4.5 Comparison of the *Streptocarpus* genetic map to other Gesneriaceae maps

Currently, there are three genetic maps available for species in the Gesneriaceae family, for the New World genus *Rhytidophyllum* (Alexandre et al., 2015), the Asian genus *Primulina* (Feng et al., 2016), and the African genus *Streptocarpus* (this study). Using MapB-1 as the representative *Streptocarpus* genetic map (for it was constructed under the most stringent strategy), the statistics of these genetic maps can be compared (Table 5.31). The *Rhytidophyllum* map was built using a Genotyping-by-Sequencing approach (GbS), and resulted in 559 mapped markers across 16 linkage groups (Alexandre et al., 2015). The *Primulina* map was built using a SNP massARRAY derived from an Expressed Sequence Tags (EST-SNP massARRAY) genotyping method, with 215 markers in 18 linkage groups (Feng et al., 2016). While the current *Streptocarpus* MapB-1 does not have the highest number of markers, it is the densest map with the average marker interval of 3.0 cM (Table 5.31). As these three genera are geographically and phylogenetically widely separated (Möller et al., 2009; Weber et al., 2013), comparative studies or synteny analyses of the three genetic maps may provide interesting evolutionary insights.

Table 5.31 Genetic maps of the Gesneriaceae family

Taxon	**Genotyping method**	**No. LG**	**No. markers**	**Total map distance**	**Average marker interval**	**Reference**
Streptocarpus[1]	RAD-Seq	17	377	1,144.2	3.0	This study
Rhitidophyllum[2]	GbS	16	559	1,650.6	3.4	Alexandre et al., 2015
Primulina[3]	EST-SNP massARRAY	18	215	3,774.7	17.6	Feng et al., 2016

1 – *Streptocarpus grandis* and *S. rexii*, 2 – *Rhitidophyllum auriculatum* and *R. rupicola*, 3 – *Primulina eburnea*

5.4.6 Conclusion

In this chapter, the first genetic map of the genus *Streptocarpus* was constructed using a RAD-Seq genotyping method. Several genetic maps were constructed throughout the study, with their information content may be complementary to each other. In particular, *de novo* and reference-based map-building approaches were compared, and the combined approach was found to generate the genetic map of highest map density. Further improvement of the map resolution can be made through more sequencing experiments in future studies to improve the number of marker recovered. Nevertheless, these genetic maps provide the basis for the QTL mapping in Chapter 6.

Chapter 6 Studies of marker-trait association – morphological variations and QTL mapping in the *Streptocarpus* backcross population

6.1 Introduction

6.1.1 Quantitative trait loci (QTL) and binary trait loci (BTL) mapping overview

Most morphological traits, such as height and body weight, show continuous variation in phenotypic values. These traits are often regulated by multiple genes with smaller effects, in combination with interactions with environmental factors. The genetic regions associated with these traits are called quantitative trait loci (QTL). On the other hand, Mendelian traits are regulated by a single or a few genes, and segregation follows a discrete pattern, often binary, according to Mendelian laws. The genetic loci associated with these traits are called Mendelian loci or binary trait loci (BTL) (Lynch and Walsh, 1998; Mauricio, 2001; Coffman et al., 2005). QTL or BTL mapping (abbreviated henceforth as QTL mapping for simplicity) describes the process of locating these loci on a genetic map. This is achieved through collecting genotype and phenotype data from an experimental population, and analysing their correlations (Lynch and Walsh, 1998; Broman and Sen, 2009).

Identifying the causative loci could address questions such as how much phenotypic variation is due to genetic variation, and how much does each one of the loci contribute to the difference in phenotypic values observed. The loci information also narrows down the candidate region aiding isolation of the causative genes (Lynch and Walsh, 1998; Broman and Sen, 2009). QTL mapping of important agricultural traits has helped the selective breeding process to improve crops (Causse et al., 2002; Lanceras et al., 2004; Wang et al., 2016). QTL mapping was adopted in evolutionary biology to identify genetic regions related to important morphological changes related to fitness or ecological importance (Bradshaw et al., 1998; Gailing, 2008; Wessinger et al., 2014; Alexandre et al., 2015; Feng et al., 2018). Taking QTL studies of Gesneriaceae for example, in the genus *Rhytidophyllum*, a QTL study identified the loci regulating floral shape and nectar volume in relation to the formation of their hummingbird-specific pollination syndrome (Alexandre et al., 2015). QTLs for floral and leaf shape traits were identified for the genus *Primulina* to study the differentiation of two ecologically distinct sister species that grow sympatrically but have different morphologies and occupy contrasting microhabitats (Feng et al., 2018).

This study carries out QTL mapping of morphological variation in the genus *Streptocarpus.* The diverse morphological characters of this genus are well documented and preliminary knowledge about their genetic inheritance is available (Reviewed in Chapter 1;

Oehlkers, 1938; Lawrence et al., 1939; Oehlkers, 1942; Lawrence, 1947; Lawrence and Sturgess, 1957; Lawrence, 1957; 1958; Oehlkers, 1966). In particular, the rosulate / unifoliate trait, and some of the floral pigmentation traits were suggested to be inherited in Mendelian fashion implying that loci with major effects may be found (Oehlkers, 1938; 1942; Lawrence, 1957). QTL mapping will aid the identification of these loci and shed light on the genetic basis of these interesting traits, which may ultimately enhance our understanding on how this highly diverse genus has evolved.

In terms of methodology, QTL mapping starts with constructing a genetic map and collecting phenotype data from the mapping population to study the trait segregation (Sehgal et al., 2016). For a Mendelian trait, the segregation ratio observed within the population should follow Mendelian laws of segregation; for a quantitative trait, the distribution of the phenotype should be statistically tested for their distribution (i.e. parametric or nonparametric), as this will affect the selection of the QTL model in later analyses (Broman and Sen, 2009). Phenotypic correlations evaluate how tightly two traits tend to co-segregate, suggesting pleiotropic effects (Lynch and Walsh, 1998). With the genetic maps and phenotype data at hand, QTL mapping can be performed.

Commonly used mapping approaches include standard interval mapping (SIM; Lander and Botstein, 1989) and composite interval mapping (CIM; Zeng, 1994). SIM performs QTL model fitting along the intervals between two genotyped markers and tests the result using a maximum-likelihood method. CIM is based on SIM but incorporates ‘cofactors’ in the analysis, a group of markers which show significant association with the trait. This reduces the genetic background noise hence improves the power of QTL detection, distinguishing closely linked QTLs (Broman and Sen, 2009). After identifying the genetic loci, QTL models can be fitted to estimate (1) the proportion of phenotypic variance explained by the loci, (2) the effect size of each loci, and (3) the interaction between loci (Broman and Sen, 2009). In addition, effect plots of the identified loci can be graphed for direct comparison of the average phenotypic values between different genotypes. For example, if ‘marker A’ was found to be a potential QTL in a BC population, the average phenotypic value of individuals with homozygous genotype at ‘marker A’ should be statistically different from that of the individuals with heterozygous genotype (Broman and Sen, 2009). On the contrary, if effect plots show no difference between the two genotypes, then the identified QTL may be a false signal. A general workflow of QTL mapping is summarised in Figure 6.1.

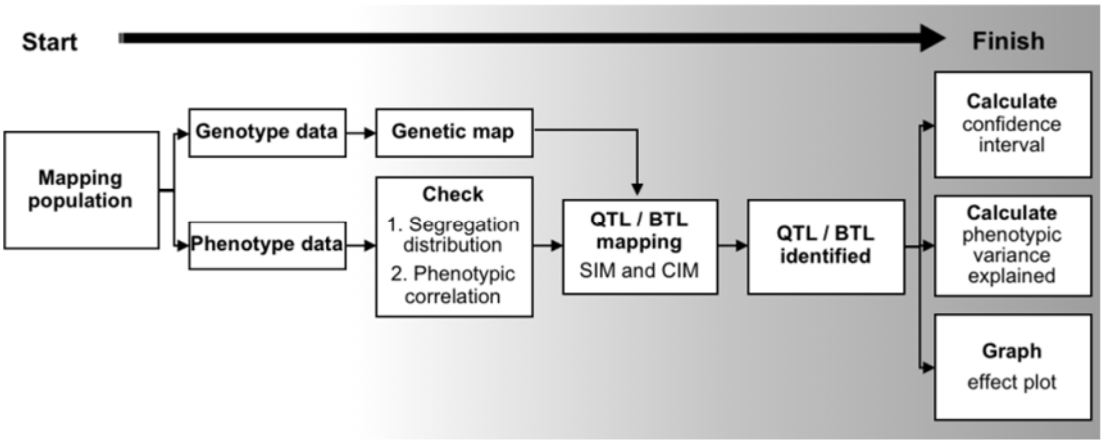


Figure 6.1 General workflow of QTL mapping using SIM and CIM methods

6.1.2 Morphological differences between *S. rexii* and *S. grandis* and their hybrids

The mapping population and the genetic maps constructed in Chapter 5 represented the basis for QTL mapping, as the two species used to construct the BC population show contrasting phenotypes. The morphologies of *S. rexii* and *S. grandis* were briefly described in Chapter 1 and illustrated (Figure 1.5). Here a more detailed background on their morphological, developmental, and genetic differences are provided.

Streptocarpus rexii is a perennial plant with an excentric rosulate growth form, and has open-tube type flowers with pollination chambers and a purple anthocyanin stripe pigmentation in the corolla (Jong and Burtt, 1975; Möller et al., 2018). During embryogenesis, the *S. rexii* embryo does not develop a shoot apical meristem (SAM) between the cotyledons (Jong, 1970; Mantegazza et al., 2007). The seedlings also lack a SAM and the cotyledons develop unequally in size (anisocotyly) due to the activity of a basal meristem (BM) at the proximal end of the lamina of the macrocotyledon (Jong, 1970; Mantegazza et al., 2007; Nishii and Nagata, 2007). In anisocotylous seedlings at around 21 to 35 days-after-sown (DAS), the groove meristem (GM) first emerges as a group of densely staining cells between the two cotyledons at the base of the macrocotyledon. With further development, the GM is organised and possesses a tunica-corpus-like meristem structure, and is located at the groove at the junction of lamina and petiolode of the macrocotyledon (Nishii and Nagata, 2007). The formation of the first phyllomorph occurs at around 65 DAS (Nishii et al., 2010a), as the GM transforms into the bulge stage (i.e. forming a bulge of small meristematic cells), followed by the dome-shaped GM stage (i.e. a bulge partly covered with trichomes), followed by the formation of the first phyllomorph (Nishii and Nagata, 2007).

Streptocarpus grandis is a monocarpic plant (i.e. dies after flowering and fruiting) with a unifoliate growth form, and has open-tube type flowers with broad cylindrical tubes that have purple pigmentation blotches and yellow spots (Möller et al., 2018; Figure 1.5 c and d). The embryo and early seedling stages of *S. grandis* are similar to those of *S. rexii* (Jong, 1970). The onset of anisocotyly begins at about 16 DAS and becomes apparent at 30 DAS, when fan-shaped BMs can be observed at the proximal end of the macrocotyledon (Jong, 1970; Imaichi et al., 2000). At about the same time, the formation of a GM is observed on the petiolode of macrocotyledons and is distinguished from the surrounding tissue by smaller cell sizes. The GM increases gradually in size with the enlarging macrocotyledon and petiolode, and a tunica-corpus-like meristem structure is established (Imaichi et al., 2000). The inflorescence meristem later initiates from the GM with multiple inflorescence primordia arising in acropetal order (Jong, 1970, 1978; Imaichi et al., 2000).

Occasionally an additional phyllomorph may form at the base of the first inflorescence towards the end of the flowering season, termed 'accessory phyllomorph', 'subtending phyllomorph' or 'supplementary phyllomorph' (Oehlkers 1956; Jong, 1978; Dubuc-Lebreux, 1978; Nishii et al., 2012a). This development was documented in several unifoliate species including *S. grandis*, *S. wendlandi, S. michelmorei* and *S. goetzei*, and external hormone treatment with gibberellin can enhance the production of more accessory phyllomorphs (Dubuc-Lebreux, 1978; Nishii et al., 2012a). It is unknown whether the accessory phyllomorph originates from the GM or from a separate blastogen (Jong, 1978).

Hybrid plants of crosses *S. grandis* × *S. rexii* all show a rosulate growth form, and the (*S. grandis* × *S. rexii*) × *S. grandis* backcross (BC) population were reported to segregate into a Mendelian ratio of rosulate to unifoliate ratio of 3:1 (Oehlkers, 1938, 1942). It was suggested that the rosulate phenotype is regulated by an early acting locus (E) and a late acting locus (L); the early locus producing a rosulate to unifoliate 1:1 ratio in six-months-old BC plants, and the late locus producing a rosulate to unifoliate 3:1 ratio in nine-months-old BC plants (Oehlkers 1942). *Streptocarpus rexii* is hypothesised to carry the dominant alleles at both loci (E/E and L/L), and *S. grandis* the recessive alleles (e/e and l/l; Oehlkers, 1938; 1942). However, there can be great variations in the time the rosulate phenotype appears, and it may take longer time observe the 3:1 ratio (Oehlkers 1942). On the other hand, more recent studies using other rosulate × unifoliate crossing combinations, including *S. rexii* × *S. wittei* and *S. rexii* × *S. dunnii*, suggest the distinction between rosulate and unifoliate may not be clear in the BC population, and ambiguous phenotypes can be found (Harrison, 2002; Harrison et al., 2005). For instance, some backcross individuals may have two macrocotyledons, and if the phenotype was not scored at an early stage the second macrocotyledon may result in the plant to be scored as a rosulate phenotype. The formation of an accessory phyllomorph can also cause confusion, as it is morphologically similar to a phyllomorph in rosulates but the trait is actually inherited from the unifoliate parent, and it was suggested that they should not be scored as a rosulate phenotype (Harrison, 2002). At least six morphological types were observed in BC populations previously, including single leaf unifoliates, plants with two macrocotyledons, plants with one main leaf with some very small additional leaves, plants with two main leaves and some small ones, plants with more than two main leaves but not fully rosulate, and fully rosulate plants (Harrison, 2002).

In addition to the rosulate and unifoliate phenotypes, other traits were described for *S. grandis* × *S. rexii* F2 populations, but not BC populations (Oehlkers, 1942). A Mendelian segregation 3:1 ratio was observed for the inflorescence colour, midrib colour, and absence / presence of striped pigmentation in the flower (Oehlkers, 1942). Other traits such as presence or absence of yellow spots were studied in other *Streptocarpus* hybrids, and were reviewed in Chapter 1 (Table 1.2 and Table 1.3).

6.1.3 Objectives of this chapter

Overall, the *S. grandis* × *S. rexii* BC population can be used to study the genetic inheritance of *Streptocarpus* morphological traits, and QTL mapping can shed light on the underlying genetics of these traits. In this chapter, the vegetative and floral characters were studied for the parental lineages and the BC population, including the growth habit, floral dimension traits, flowering time, and flower pigmentation patterns. The segregation and correlation patterns between the phenotypes were investigated, and QTL mapping of the measured traits performed using the genetic maps constructed in Chapter 5. In particular, the main focus was the mapping of the rosulate / unifoliate loci, with SIM and CIM of four different scoring methods performed on three different genetic maps (i.e. MapA, MapB-1, and MapB-3) to retrieve as much information as possible. For other traits measured, SIM analyses were performed using the main genetic map MapB-1 as this map represent the most stringently filtered genetic map (i.e. based on contaminant-free genome assembly and with strigent marker filtering strategy). Finally, we identified the genome scaffolds that fallen within the QTL found that are associated with rosulate / unifoliate trait, and performed genome annotation on the scaffolds in search of candidate genes.

6.2 Materials and methods

A flowchart summarising the whole analysis process carried out in this chapter can be found in Appendix 6.1.

6.2.1 Plant materials

The (*S. grandis* × *S. rexii*) × *S. grandis* backcross population used for phenotyping and QTL mapping consisted of 233 plants and was the same as described in section 5.2.1 (Figure 6.2). For the phenotype scoring of the parental materials, four plants of *S. rexii* (accession 20150819*A) were used, which were propagated from a single plant using leaf cuttings and thus have the same qualifier *A. For *S. grandis*, eight *S. grandis*^{BC} plants (accession 20150821) and one *S. grandis*^{F1} plant (accession 20151810) were available at the time of this experiment and were all used for phenotype scoring. For the *S. grandis* × *S. rexii* F1 hybrid, three leaf-cutting-propagated plants of the original F1 lineage (accession 20071108*J) were used (Table 6.1). All plant materials were sown, propagated and maintained in the living research collection at the Royal Botanic Garden Edinburgh.

Table 6.1 List of parental and backcross materials used in the study

Taxon	Accession	Qualifier	Date sown	No. plants	Note
Streptocarpus rexii	20150819	A	17.01.2015	4	Used for genome sequencing
Streptocarpus grandis^{BC}	20150821	A, B, C, H, I, K, M, O	17.01.2015	8	Used for genome sequencing
Streptocarpus grandis^{F1}	20151810	O	27.07.2015	1	
Streptocarpus grandis^{F1} × *S. rexii*	20071108	J	27.08.2007	3	
(*S. grandis* × *S. rexii*) × *S. grandis*^{BC}	20150825	A - IS	17.01.2015	233	Used for genetic mapping

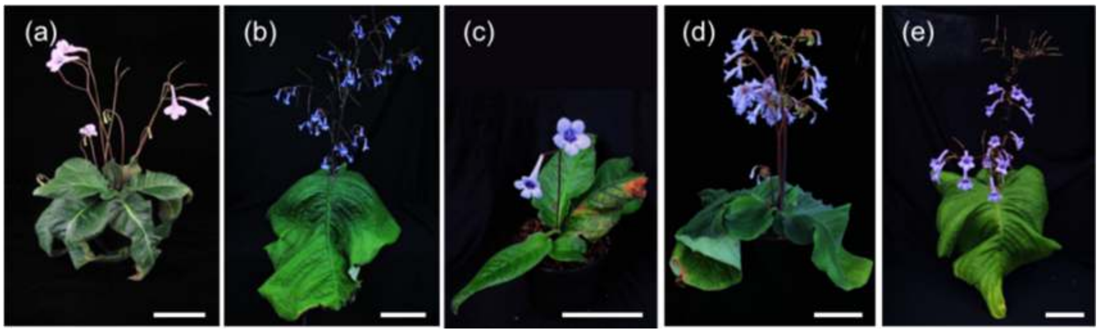


Figure 6.2 *Streptocarpus* materials used in the study. **(a)** *S. rexii*. **(b)** *S. grandis*. **(c)** *S. grandis* × *S. rexii* F1. **(d)** Example of a rosulate (*S. grandis* × *S. rexii*) × *S. grandis* BC plant. **(e)** Example of a unifoliate BC plant. Bars = 10 cm.

6.2.3 Morphology scoring of the parental plants

Floral and vegetative characters of the parental materials were measured. For the scoring of floral characters, fresh flowers were collected from *S. rexii* (4 plants, totally 17 flowers), *S. grandis*^{BC} (8 plants, totally 12 flowers), *S. grandis*^{F1} (1 plant, 10 flowers), and S. *grandis* × *rexii* F1 hybrids (Table 6.3; 3 plants, totally 17 flowers). Photos of the collected flowers were taken including side view, top view, front five, and dissected view (removal of the adaxial side of the corolla) using a Canon G12 camera (Canon Inc., Tokyo, Japan) (Figure 6.3 a, b, c, d). All images taken have a scale ruler for standardisation. The images were analysed in ImageJ v1.48 (Schneider et al., 2012), and the each character measured as summarised in Table 6.4 and Figure 6.3. The list of traits measured was based on previous studies (Lawrence, 1957; Oehlkers, 1967; Harrison et al., 1999; Chou, 2008). For measuring the quantitative floral traits the scale in the photos was used, and the *Straight* line tool in ImageJ used for length measurements. For binary traits, visual inspection was made directly on the photo. The pigmentation traits were scored using the dissected flower photos (Figure 6.3 d). Statistical tests were carried out for quantitative traits in R v3.3.0 (R Development Core Team, 2008): The Wilcoxon-rank-sum test (Package 'wilcox.test' in R; Bauer, 1972) was used for comparing the quantitative data between *S. grandis*^{F1} and *S. grandis*^{BC}, and between *S. rexii* and *S. grandis* (*S. grandis*^{BC} and *S. grandis*^{F1}); and Dunn's post-hoc test (Package 'dunn.test' in R; Dunn, 1964) was used for three-ways comparisons among *S. rexii*, *S. grandis*, and the F1 hybrid.

Table 6.3 Number of plants and flowers collected for the trait measurement in parental materials

Taxon	Accession	No. plants	No. flowers collected
Streptocarpus rexii	20150819	4	17
Streptocarpus grandis^{BC}	20150821	8	12
Streptocarpus grandis^{F1}	20151810	1	10
Streptocarpus grandis × *S. rexii* F1	20071108	3	17

(Next page) Figure 6.3 Illustration of the floral pictures taken using a flower of *S. rexii* as example. **(a)** Flower side view. **(b)** Flower top view. **(c)** Flower face view. **(d)** Flower dissected view. The dorsal corolla tube was dissected off, and the pistil removed. **(e)(f)(g)(h)** Schematic illustration of the flower photos, with the traits measured in each photo indicated with dotted lines. **(e)** Side view. **(f)** Top view. **(g)** Face view. **(h)** Dissected view. The numbers assigned to each trait correspond to Table 6. (1) Corolla length. (2) Undilated tube length. (3) Dilated tube length. (4) Undilated tube height. (5) Dilated tube height. (6) Undilated tube width. (7) Dilated tube width. (8) Corolla face height. (9) Tube opening height, outer. (10) Tube opening height, inner. (11) Corolla face width. (12) Tube opening width, outer. (13) Tube opening width, inner. (14) Pistil length. (15) Ovary length. (17) Calyx length. (18) Stamen length. (19) Filament length, attached part. (21) Ventral tube length. (22) Ventral lobe length. (23) Dorsal tube length. (24) Dorsal lobe length. Bar = 2.5 cm.

Figure 6.3 Illustration of the floral pictures taken using a flower of *S. rexii* as example. Full legent given on prevopus page.

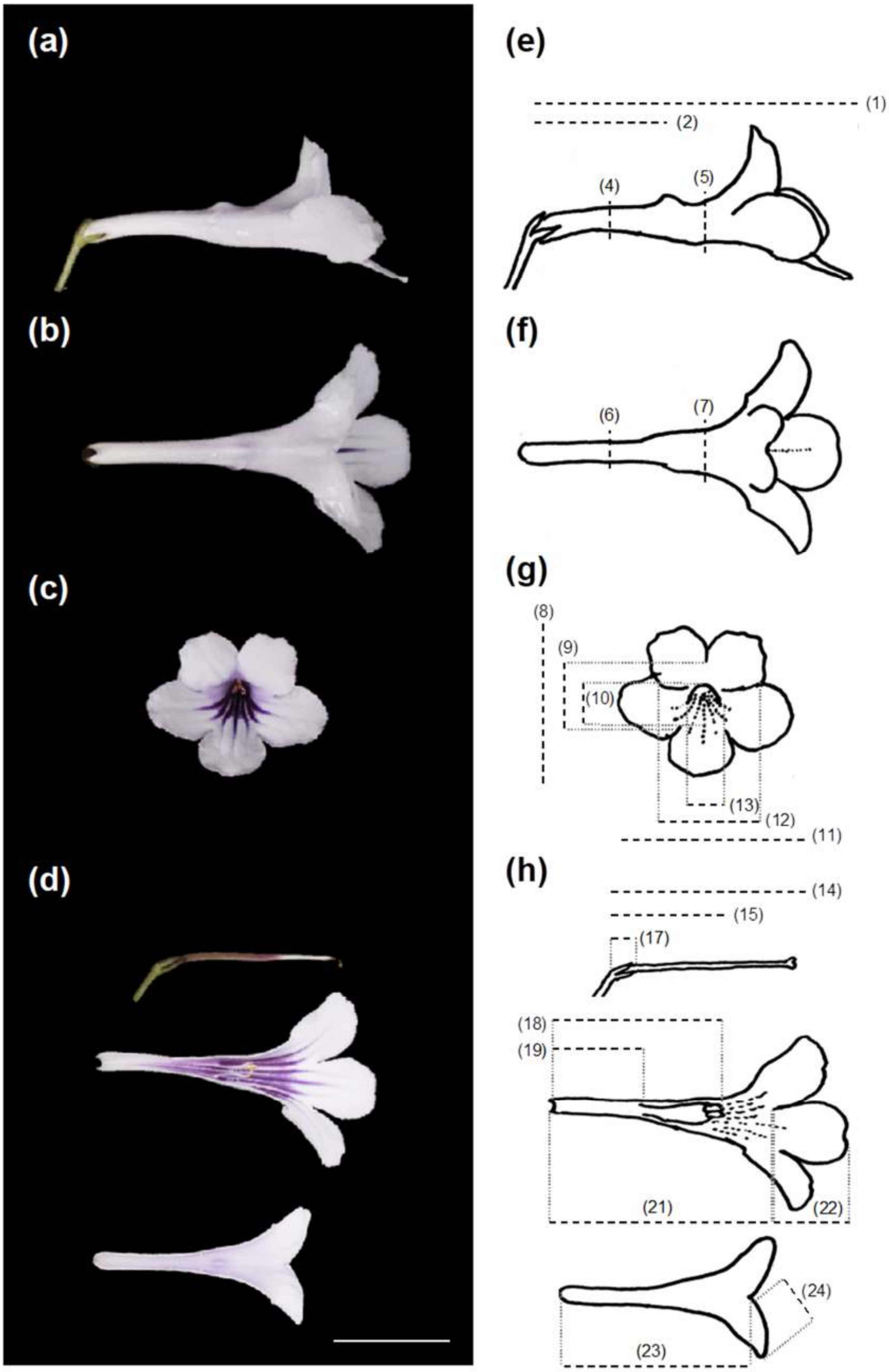


Table 6.4 List of characters measured in the parental *S. rexii*, *S. grandis* and F1 materials

Trait No.	Trait name	Data type*	Trait description
I. Flower dimensions			
1	Corolla length	Q	Length of whole corolla (corolla tube + lobe)
2	Undilated tube length	Q	Length of undilated part of corolla tube
3	Dilated tube length	Q	Length of dilated part of corolla tube (trait 1 – trait 2)
4	Undilated tube height	Q	Height of undilated part of corolla tube
5	Dilated tube height	Q	Height of dilated part of corolla tube
6	Undilated tube width	Q	Width of undilated part of corolla tube
7	Dilated tube width	Q	Width of dilated part of corolla tube
8	Corolla face height	Q	Height of front facing corolla
9	Tube opening height (outer)	Q	Height of corolla tube entrance†
10	Tube opening height (inner)	Q	Height of corolla tube entrance
11	Corolla face width	Q	Width of front facing corolla
12	Tube opening width (outer)	Q	Width of corolla tube ‡
13	Tube opening width (inner)	Q	Width of corolla tube entrance
14	Pistil length	Q	Length of pistil
15	Ovary length	Q	Length of ovary (purple part of the pistil)
16	Style length	Q	Length of style (trait 14 – trait 15)
17	Calyx length	Q	Length of calyx
18	Stamen length	Q	Length of whole stamen (includes filament and anther)
19	Filament length (attached)	Q	Length of part of filaments that is fused to the corolla tube
20	Filament length (free)	Q	Total filament length minus the fused part (trait 18 – trait 19)
21	Ventral tube length	Q	Length of ventral tube
22	Ventral lobe length	Q	Length of ventral lobe

Table 6.4 continued

Trait No.	Trait name	Data type*	Trait description
23	Dorsal tube length	Q	Length of dorsal tube
24	Dorsal lobe length	Q	Length of dorsal lobe
II. Other floral traits			
25	Flowering time	Q	Time of first flower, unit: days after sowing
26	Lateral lobe pigmentation	B	Presence or absence of the pigmentation on lateral lobe
27	Ventral lobe pigmentation	B	Presence or absence of the pigmentation on ventral lobe
28	Yellow spot	B	Presence or absence of the yellow spot
III. Vegetative traits			
29	Rosulate/unifoliate scoring	B	Rosulate or unifoliate
30	Two macrocotyledons	B	With or without two macrocotyledons
31	Days to 1st leaf	Q	Time to first leaf initiation, unit: days after sowing

* Q – Quantitative data, B – binary data. † the height from the joint between two dorsal lobe to the line between the two joints of the lateral lobe and ventral lobe. ‡ The width between the two joints of the dorsal lobe and the lateral lobe.

6.2.4 Morphology scoring and examination of trait distribution in the BC mapping population

The morphology of 233 plants of the BC mapping population was assessed as described previously as the parental material in section 6.2.3. For floral characters, photos of two to three flowers per BC individuals were taken (Figure 6.3).

The vegetative habit of the BC plants was observed by eye once every week and photos of each plant taken once every month from 15 April 2015 to 30 May 2016. As reported above and also observed in the present work, categorising the phyllomorphs in some BC plants was challenging due to their variability in occurrence (such as the morphologies described in Harrison 2002). As a result, we classified any additional phyllomorph observed (i.e. any newly produced ones in addition to the two cotyledons) into six types (Table 6.5, Figure 6.4). To aid the categorisation, the weekly visual observations and monthly photo records were used.

Table 6.5 Description of the types of additional phyllomorphs observed in the (*S. grandis*^{F1} × *S. rexii*) × *S. grandis*^{BC} backcross population

Type	Name	Description of the type
1	True phyllomorphs	Additional phyllomorphs originating from the position of the GM at the base of the preceding, usually cotyledonary, phyllomorph; the additional phyllomorphs were sessile, i.e. did not have an elongated stalk (Figure 6.4 a).
2	Accessory phyllomorphs	Subtending a series of acropetally forming inflorescences and its petiolode was attached to the base of the preceding, usually cotyledonary, phyllomorph in the position of the GM; the petiolode usually has an elongated stalk (Figure 6.4 b).
3	Bract-like phyllomorphs	Produced from a “node” in the position of the GM at the base of a late developing inflorescence, and is usually located near the base of the cotyledonary phyllomorph and is sessile (Figure 6.4 c).
4	Ambiguous phyllomorphs	Originating from the base of the cotyledonary phyllomorph in position of the GM but were presumed to have been buried underground during repotting and under-developed (Figure 6.4 d).
5	Adventitious phyllomorphs	Produced along the base of the acropetally formed row of inflorescences and did not originated from the position of the GM (Figure 6.4 e).
6	Paired accessory phyllomorphs	This morphology is similar to the type 2 accessory phyllomorphs, but possessed two opposite phyllomorphs both bearing inflorescences in acropetal succession (Figure 6.4 f).

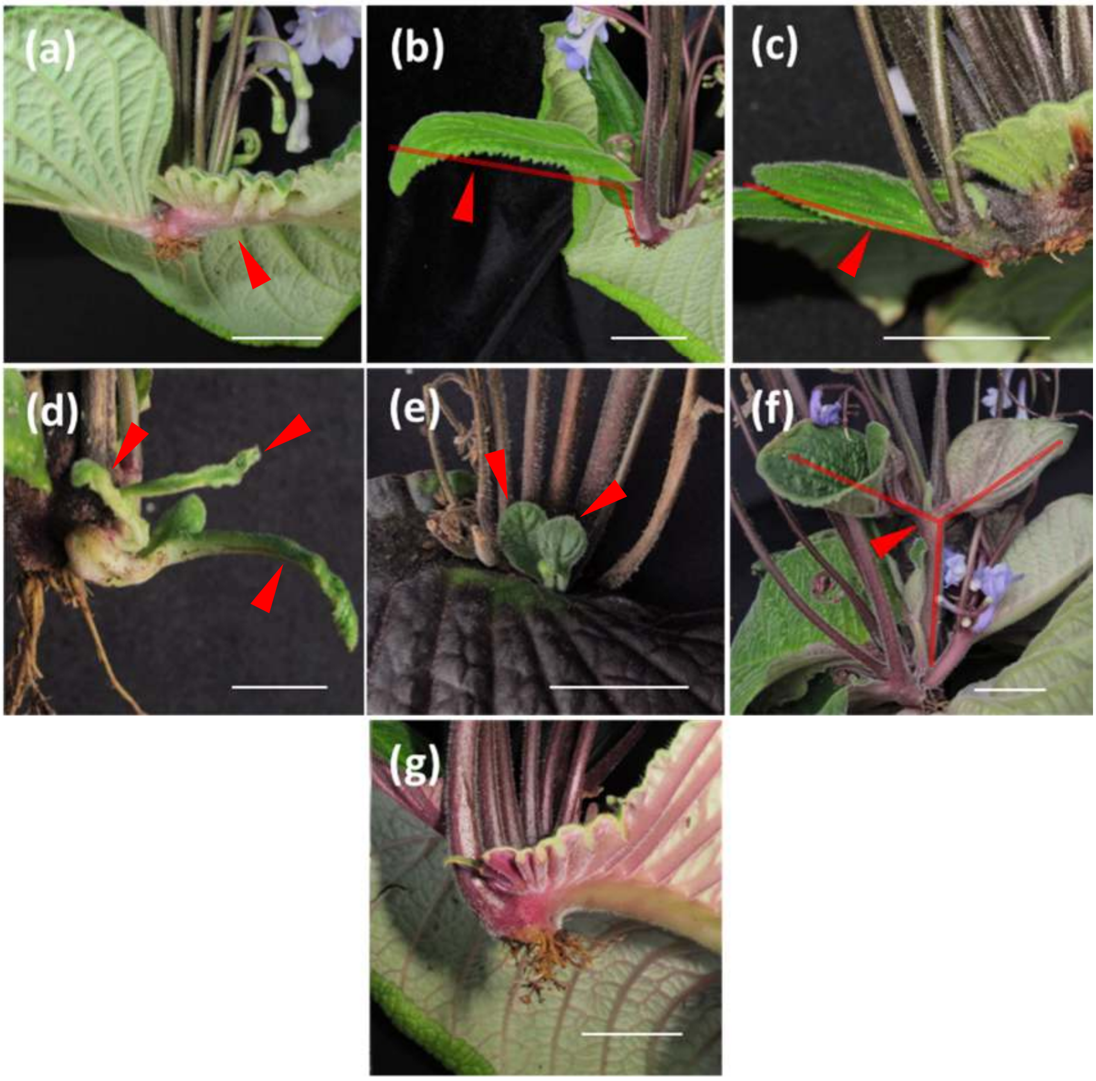


Figure 6.4 Examples of vegetative phenotypes observed in the BC population. These pictures were taken after removing the plants from the pots and before pressing them into herbarium specimen, thus the morphology of each phyllomorph can be more accurately captured. **(a)** Type 1, true phyllomorph. **(b)** Type 2, accessory phyllomorph. **(c)** Type 3, bract-like phyllomorph. **(d)** Type 4, ambiguous phyllomorph. The leaf buds were buried under soil after repotting and were typically under-developed, i.e. less than 5 mm. **(e)** Type 5, adventitious phyllomorph, produced along the row of inflorescences. **(f)** Type 6, paired accessory phyllomorphs. **(g)** Unifoliate. Red arrow heads and red lines indicate the additional phyllomorphs observed which is the main distinguishing feature. Bars = 2 cm.

Because of the difficulties in categorising some plants in the BC population due to the type of additional phyllomorphs they produced, four different scoring methods were devised used differing in the categorisation of these ambiguous phenotypes (Table 6.6):

Method 1 – Plants with type 1 phyllomorphs were scored as rosulate. Those with only type 2, 3, 4, 5 and/or 6 and true unifoliates were all scored as unifoliate.

Method 2 – Plants with type 1 and/or type 2 phyllomorphs were scored as rosulate. Plants with only type 3, 4, 5 and/or 6 and true unifoliates were scored as unifoliate.

Method 3 – Plants with type 1 phyllomorphs were scored as rosulate. Plants with only type 2, 3, 4, 5 and/or 6 were scored as unknown. Only plants without any additional phyllomorphs were scored as unifoliate.

Method 4 – Plants with type 1, 2, 3, 4 or 6 phyllomorphs were scored as rosulate. Those plants with type 5 phyllomorphs were scored as unknown, and plants without any additional phyllomorphs were scored as unifoliate.

Table 6.6 Scoring methods for the QTL analysis of the vegetative habit trait for the *S. grandis*^{F1} × *S. rexii*) × *S. grandis*^{BC} backcross population.

Type	Primary phyllomorph	Method 1	Method 2	Method 3	Method 4
1	True	R	R	R	R
2	Accessory	U	R	?	R
3	Bract-like	U	U	?	R
4	Ambiguous	U	U	?	R
5	Adventitious	U	U	?	?
6	Paired accessory	U	U	?	R
7	Unifoliate	U	U	U	U

Eventually, all BC plants were processed into herbarium voucher specimens to preserve their morphology. Prior to pressing, photos of each plant were taken with particular focus on the basal part, which was easier after the plants were removed from the pots.

6.2.5 Phenotypic distribution, segregation ratio, and phenotypic correlation in the BC population

The distribution of the different phenotypes observed in the BC population was visualised using R v3.3.0. The normality of the distribution (i.e. whether the data fits a normal distribution or not) was checked using Shapiro-Wilk test in R (function 'shapiro.test'; Royston, 1982). The segregation ratio of binary traits (traits 26, 27, 28, 29, 30) were examined by Chi-square tests using the QuickCalc Chi-square test function (GraphPad Software, Inc. Accessed 5 March 2019. Available at https://www.graphpad.com/quickcalcs/chisquared1.cfm). For phenotypic correlations, Spearman correlation coefficients were calculated between each pair of quantitative traits, using the 'cor.test' function in R (Hollander and Wolfe, 1973; Best and Roberts, 1975). The Spearman correlation was chosen instead of Pearson's correlation, as some of the measured traits showed non normal distributions and some were binary. Thus, the data themselves did not meet the assumptions of Pearson’s correlation (i.e. the data should be continuous and follow a normal distribution; Hollander and Wolfe, 1973). The results of the correlation tests were visualised in R using the function ‘pairs’. The commands used are summarised in Box 6.1.

Box 6.1 Commands used for the phenotypic correlation analysis and visualisation of the results

```
## Phenotpyic correlation
## Ref :
## http://stackoverflow.com/questions/31709982/how-to-plot-in-r-a-
## correlogram-on-top-of-a-correlation-matrix
## http://stat.ethz.ch/R-manual/R-devel/library/graphics/html/pairs.html

## 1. Load data

data=read.csv("[INPUT_FILE.csv]")

## 2. Set heatmap color range
## from left to right:
## negative correlation, no correlation, positive correlation

colorRange <- c('green3', 'white', 'red3')
myColorRampFunc <- colorRamp(colorRange)

## 3. Set panel.smooth for printing scattered plot
## Copy and paste the whole command block below

panel.smooth<-function (x, y, col = "grey", bg = NA, pch = 18,
                        cex = 0.8, col.smooth = "black", span = 2/3,
                        iter = 3, ...)
{
  points(x, y, pch = pch, col = col, bg = bg, cex = cex)
  ok <- is.finite(x) & is.finite(y)
  if (any(ok))
    lines(stats::lowess(x[ok], y[ok], f = span, iter = iter),
          col = col.smooth, ...)
}

## 4. Set panel.hist for histogram
## Copy and paste the whole command block below

panel.hist <- function(x, ...)
{
  usr <- par("usr"); on.exit(par(usr))
  par(usr = c(usr[1:2], 0, 1.5) )
  h <- hist(x, plot = FALSE)
  breaks <- h$breaks; nB <- length(breaks)
  y <- h$counts; y <- y/max(y)
  rect(breaks[-nB], 0, breaks[-1], y, col = "white", ...)
}

## 5. Set panel.cor.value for printing the correlation coefficient
## and print the degree of correlation in heatmap
## Copy and paste the whole command block below

panel.cor.value <- function(w, z, digits = 2, cex.cor, ...)
{
  ## Heat map part
  ################
  correlation <- round(cor(w, z,method="spearman",use="pairwise"),2)
  col <- rgb( myColorRampFunc( (1+correlation)/2 )/255 )
  ## Also, square the value to avoid visual bias due to "area vs diameter"
  radius <- sqrt(abs(correlation))
  radians <- seq(0, 2*pi, len=50)
  ## 50 is arbitrary
  x <- radius * cos(radians) * 10
  y <- radius * sin(radians) * 10
  ## ‘*10’ is to fill the whole square same as above
  x <- c(x, tail(x,n=1))
  y <- c(y, tail(y,n=1))
  ## make them full loops
  par(new=TRUE)
  plot(0, type='n', xlim=c(-1,1), ylim=c(-1,1), axes=FALSE, asp=1)
  polygon(x, y, border=col, col=col)
  ################
  ## Correlation coefficient part
  ################
  usr <- par("usr"); on.exit(par(usr))
  par(usr = c(0, 1, 0, 1))
  # correlation coefficient
  r <- cor(w, z, method="spearman",use = "pairwise")
  ## I added ‘use=“pairwise”’ to deal with missing value
  txt <- format(c(r, 0.123456789), digits = digits)[1]
  txt <- paste("r= ", txt, sep = "")
  text(0.5, 0.6, txt)
  # For  P-value calculation
  p <- cor.test(w, z, method="spearman")$p.value
  txt2 <- format(c(p, 0.123456789), digits = digits)[1]
  txt2 <- paste("p= ", txt2, sep = "")
  if(p<0.01) txt2 <- paste("p ", "<0.01", sep = "")
  text(0.5, 0.4, txt2)
  ################
}

## 6. Plot
pdf('Correlation plot.pdf',width=20,height=20)
pairs(data, lower.panel = panel.smooth, diag.panel = panel.hist,
      upper.panel = panel.cor.value, gap = 0.5, text.panel = NULL)
dev.off()
```

6.2.6 QTL mapping

QTL mapping was performed using the R package ‘qtl’ v1.39-5 (Broman and Sen, 2009) in R v3.3.0. The genetic maps reported in Chapter 5 were used for the analysis, in specific the combined approach-MapA, MapB-1, and MapB-3. The analysis focused particularly on the mapping of the rosulate / unifoliate loci, which was performed on all three genetic maps using both SIM and CIM methods (Table 6.7). On the other hand, the mapping of all other traits was solely performed on MapB-1 using the SIM method, as this map is composed by the most stringently filtered markers and hence is the most accurate genetic map (Table 6.7).

Table 6.7 Details of the genetic maps used for QTL mapping

	Combined-MapA	MapB-1	MapB-3
Genetic map statistics			
No. of linkage group recovered	16	17	16
No. of mapped markers	599	377	853
Total map distance (cM)	1,578.2	1,144.2	1,389.9
Average distance between markers (cM)	2.6	3.0	1.6
Usage in QTL mapping			
Used for rosulate / unifoliate trait mapping	yes	yes	yes
Used for quantitative trait mapping	no	yes	no
Used for other binary trait mapping	no	yes	no

The function ‘calc.genoprob’ of the ‘qtl’ package was used to calculate the underlying genotype at every 1 cM using the Haldane map function (Haldane, 1919). The function ‘scanone’ was then used for the SIM method to calculate the likelihood that the genetic regions were associated with the trait variations, and the results were visualized as a LOD curve. The model selection for the SIM analysis was based on the type of distribution of the measured traits (described in section 6.2.3). For quantitative traits showing normal distributions, extension of the Haley-Knott regression method was used (Feenstra et al., 2006); for quantitative traits showing nonparametric distribution, the model “np” was selected (Kruglyak and Lander, 1995); for binary traits, including the rosulate / unifoliate trait, the model ‘binary’ was selected (Xu and Atchley, 1996; Broman, 2003). For CIM for the rosulate / unifoliate trait, the function ‘cim’ was used (Broman and Sen, 2009). The genome-wide LOD threshold was determined in 5,000 permutation tests, i.e. option ‘n.perm’ in the function ‘scanone’, and the value corresponding to 0.05 false discovery rate was chosen as the LOD threshold (Broman and Sen, 2009). LOD curves showing ‘peaks’ (a LOD score higher than the obtained threshold) were examined and their Bayes confidence intervals calculated using the ‘qtl’ package function ‘lodint’ (Manichaikul et al., 2006). The percentage of phenotypic variance explained was calculated using the ‘fitqtl’ function. The effect plots of the measured loci were generated using the ‘effectplot’ function. All commands and functions used are summarised in Box 6.2.

Box 6.2 Commands used for QTL mapping using the ‘qtl’ package in R

```
## 1. Install package

install.packages("qtl")
library(qtl)

## 2. Load input data. The missing genotype are denoted as – or NA

data=read.cross(format="csv",file="[R/QTL_INPUT_FILE.csv]",na.strings=c("
    -","NA"),genotypes=c("b","h"),estimate.map=FALSE,convertXdata=FALSE)

## 3. QTL mapping
# 3.1 Generation of missing genotype using HMM model with 1 cM iteration

data = calc.genoprob(data, step=1, stepwidth="fixed")

# 3.2 SIM with the traits listed in column 1 in the input file
# First line: for binary traits
# Second line: for quantitative traits
# Third line: for nonparametric traits
morph.bin = scanone(data,pheno.col=1,model="binary")
morph.ehk = scanone(data,pheno.col=1,model="normal",method=”ehk”)
morph.np = scanone(data,pheno.col=1,model="np")

# 3.3 CIM
Cim.bin=cim(data,pheno.col=1)

# 3.4 Check LOD distribution of all LGs
plot(morph.bin)
plot(morph.ehk)
plot(morph.np)

# 3.5 Permutation test to calculate LOD threshold, with 5,000 permutations

morph.perm.bin = scanone(data,pheno.col=1,model="binary",n.perm=5000)
morph.perm.ehk =
    scanone(data,pheno.col=1,model="normal",method=”ehk”,n.perm=5000)
morph.perm.np = scanone(data,pheno.col=1,model="binary",n.perm=5000)

# 3.6 Obtain the LOD threshold corresponding to 0.05 false discovery rate
quantile(morph.perm.bin,0.95)
quantile(morph.perm.ehk,0.95)
quantile(morph.perm.np,0.95)

# 3.7 Calculate Bayes confidence interval based on the LOD threshold
bayesint(morph.bin,[LG],[LOD_THRESHOLD],expandtomarkers=TRUE)
bayesint(morph.ehk,[LG],[LOD_THRESHOLD],expandtomarkers=TRUE)
bayesint(morph.np,[LG],[LOD_THRESHOLD],expandtomarkers=TRUE)

## 4. Fitting QTL models and calculated the variance explained%
# Example for binary trait with one locus detected
qc=[LG_OF_DETECTED_LOCI]
qp=[POSITION_OF_MARKER]
qtl=makeqtl(data,qc,qp,what="prob")
lod=fitqtl(data,pheno.col=[TRAIT_COLUMN],qtl,formula=y~Q1,model="binary")
summary(lod)
```

```
# Example for quantitative trait with five loci detected
qc=[LG_OF_DETECTED_LOCI]
qp=[POSITION_OF_MARKER]
qtl=makeqtl(data,qc,qp,what="prob")
lod=fitqtl(data,pheno.col=[TRAIT_COLUMN],qtl,formula=y~Q1*Q2*Q3*Q4*Q5
    ,model=”normal”,method="ehk")
summary(lod)

# Example for nonparametric trait
qc=[LG_OF_DETECTED_LOCI]
qp=[POSITION_OF_MARKER]
lod=fitqtl(data,pheno.col=[TRAIT_COLUMN],qtl,formula=y~Q1*Q2*Q3*Q4*Q5
    ,model=”np”)
summary(lod)

## 5. Effect plot of a specific marker (within the detected QTLs)
effectplot(data,pheno.col=[TRAIT_COLUMN],mname=”[MARKER_NAME]”)
```

6.2.7 Genome annotation for the rosulate / unifoliate loci

To search for candidate genes related to the rosulate / unifoliate trait, the BTL regions identified from the above section were examined. Genetic markers which fell within the BTL regions (from both SIM and CIM mapping results of all three maps) were listed, and their corresponding genome scaffolds retrieved (the genome scaffolds where the marker sequences were derived from). The retrieved sequences were annotated using the web-based pipeline MEGANTE (Numa and Itoh, 2014 Release 2018-02), with *Nicotiana tabacum* chosen as reference for the gene prediction and annotation. The retrieved sequences were all from reference-based approach markers whereas no genome information was available for *de novo*-approach-derived markers.

6.2.8 Scaffold to scaffold alignments

To identified the relationships between genome scaffolds (i.e. whether the scaffold from one genome assembly can be align to a scaffold from another genome assembly), scaffold-to-scaffold alignment was carried out using the D-GENIES web-tool under default settings (Cabanettes and Klopp, 2018). Three different alignments were performed: (1) using scaffolds of MapA as query sequence (which came from the preliminary SOAPdenovo2 *S. rexii* assembly), and scaffolds of MapB-1 / MapB-3 as target sequence (which was the filtered ABySS2 *S. rexii* assembly); (2) using scaffolds of MapA as query sequence, and the *S. grandis* genome assembly (the filtered ABySS2 *S. grandis* assembly) as target sequence; (3) using scaffolds of MapB-1 / MapB-3 as query sequence, and the *S. grandis* genome assembly as target sequence. Alignment (1) was used to identify shared-sequences between the two *S. rexii* genome assemblies; alignment (2) and (3) were used to identify the corresponding allelic sequences from the *S. grandis* genome assembly.

6.3 Results

6.3.1 Morphological variation between *S. rexii*, *S. grandis*, and their F1 hybrid

The two *S. grandis* lineages used (i.e. *S. grandis*^{F1} and *S. grandis*^{BC}) showed no significant difference between the two lineages in most of the 25 quantitative traits measured, except for 7 traits: corolla length ($P < 0.01$), corolla tube length ($P < 0.01$), corolla face width ($P < 0.01$), ventral and dorsal tube length ($P < 0.01$), pistil length ($P < 0.01$), ovary length ($P < 0.01$), and flowering time ($P < 0.01$) (Appendix 6.2). In general, the flower of *S. grandis*$^{\mathrm{F1}}$ lineage was 0.2 – 0.7 cm longer and wider than the flower of *S. grandis*^{BC} (Appendix 6.2). In terms of flowering time, the recorded flowering time of *S. grandis*^{F1} lineage is earlier (on average 265 DAS) than that of the *S. grandis*^{BC} lineage (on average 377 DAS; Appendix 6.2).

Statistical comparisons were carried out on the three parental lineages, i.e. *S. rexii*, *S. grandis*^{F1}, and F1 hybrid (Table 6.8 and Figure 6.7). Amongst all the floral dimension traits measured, *S. rexii* usually showed the larger trait values (Table 6.8; Figure 6.7, red boxes) while *S. grandis* usually had the lowest trait values (Figure 6.7, blue boxes). The F1 hybrid values usually fell between those of *S. rexii* and *S. grandis* (Figure 6.7, purple boxes), and sometimes more closely resembled *S. rexii* (e.g. corolla length and dilated tube length; Figure 6.7 a and c). The differences between *S. rexii* and *S. grandis* were statistically significant for most of the traits, except for ‘undilated tube height’ (Appendix 6.3). In general, the *S. rexii* flower was larger than or was similar to the *S. grandis* flower in most the floral traits measured.

Three way comparisons were also conducted for the three parents (Appendix 6.4). The ‘undilated tube height’ was identified to be not statistically different among all three parents (Appendix 6.4; Trait 4), while in several other traits, the trait value of the F1 hybrid was more similar to that of *S. rexii* (Appendix 6.4; Trait 1, 3, 6, 8, 9, 10, 16, 17, 21, 22, 24).

In terms of flowering time, *S. rexii* flowered on average at 237 days after sowing (DAS), which is earlier than *S. grandis* (329 DAS; Table 6.8). For the F1 lineage, because the plants used in this study originated from leaf cuttings of a plant originally sown and grown in 2007 (accession 20071108), the ‘flowering time (DAS)’ data were not available. Pigmentation on the lateral lobes was observed in all three lineages (Figure 6.8 a). Pigmentation on the ventral lobe was observed in *S. rexii* and the F1 hybrid, but not in *S. grandis* (Figure 6.8 b). A yellow spot was observed in *S. grandis*, but not in *S. rexii* or the F1 hybrid (Figure 6.8 c). Plants with two macrocotyledons were not observed among the parental lineage materials.

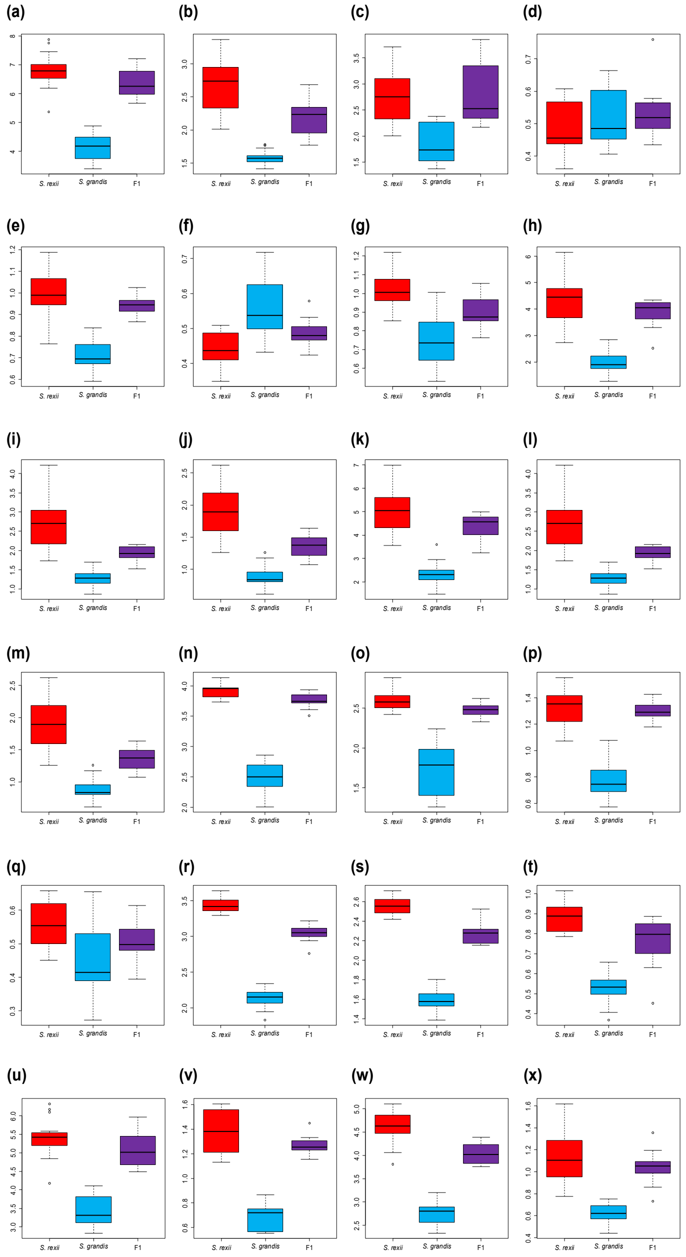


Figure 6.7 Box plots of floral quantitative traits measured in the parental lineages. Red: *S. rexii*. Blue: *S. grandis*. Purple: F1 hybrid. Unit = cm. **(a)** Corolla length. **(b)** Undilated tube length. **(c)** Dilated tube length. **(d)** Undilated tube height. **(e)** Dilated tube height. **(f)** Undilated tube width. **(g)** Dilated tube width. **(h)** Corolla face height. **(i)** Tube opening height, outer. **(j)** Tube opening height, inner. **(k)** Corolla face width. **(l)** Tube opening width, outer. **(m)** Tube opening width, inner. **(n)** Pistil length. **(o)** Ovary length. **(p)** Style length. **(q)** Calyx length. **(r)** Stamen length. **(s)** Filament length, attached part. **(t)** Filament length, free part. **(u)** Ventral tube length. **(v)** Ventral lobe length. **(w)** Dorsal tube length. **(x)** Dorsal lobe length.

Table 6.8 Summary of results of the morphometric measurements and flowering time of the parental lineages

Species	N	Corolla length	Undilated tube length	Dilated tube length	Undilated tube height	Dilated tube height	Undilated tube width	Dilated tube width
S. rexii	17	6.79 ± 0.60	2.65 ± 0.42	2.74 ± 0.46	0.49 ± 0.08	1.01 ± 0.11	0.44 ± 0.05	1.03 ± 0.10
*S. grandis**	22	4.13 ± 0.46	1.58 ± 0.10	1.86 ± 0.38	0.52 ± 0.09	0.71 ± 0.06	0.55 ± 0.09	0.74 ± 0.13
F1 hybrid	17	6.35 ±- 0.46	2.19 ± 0.26	2.90 ± 0.62	0.53 ± 0.07	0.95 ± 0.05	0.49 ± 0.04	0.90 ± 0.08

Species	Corolla face height	Tube opening height (outer)	Tube opening height (inner)	Corolla face width	Tube opening width (outer)	Tube opening width (inner)	Pistil length	Ovary length	Style length
S. rexii	4.29 ± 0.86	1.94 ± 0.49	1.51 ± 0.34	5.06 ± 0.91	2.73 ± 0.67	1.88 ± 0.41	3.91 ± 0.10	2.60 ± 0.12	1.31 ± 0.14
*S. grandis**	1.97 ± 0.44	1.08 ± 0.24	0.93 ± 0.21	2.29 ± 0.52	1.29 ± 0.22	0.88 ± 0.16	2.49 ± 0.23	1.71 ± 0.31	0.79 ± 0.15
F1 hybrid	3.86 ± 0.49	1.73 ± 0.21	1.42 ± 0.17	4.36 ± 0.50	1.91 ± 0.20	1.34 ± 0.18	3.77 ± 0.12	2.47 ± 0.08	1.30 ± 0.07

Species	Calyx length	Stamen length	Filament length (attached)	Filament length (free)	Ventral tube length	Ventral lobe length	Dorsal tube length	Dorsal lobe length	Flowering time (DAS)
S. rexii	0.56 ± 0.07	3.44 ± 0.10	2.55 ± 0.08	0.88 ± 0.07	5.39 ± 0.51	1.39 ± 0.17	4.62 ± 0.33	1.13 ± 0.22	237 ± 0.00
*S. grandis**	0.45 ± 0.11	2.13 ± 0.12	1.59 ± 0.11	0.53 ± 0.08	3.44 ± 0.40	0.69 ± 0.10	2.74 ± 0.24	0.62 ± 0.08	329 ± 77.08
F1 hybrid	0.51 ± 0.06	3.05 ± 0.11	2.28 ± 0.11	0.77 ± 0.11	5.08 ± 0.47	1.26 ± 0.07	4.04 ± 0.22	1.04 ± 0.14	-

Species	Lateral lobe pigmentation	Ventral lobe pigmentation	Yellow spot	Rosulate / Unifoliate	Two macrocotyledons	1st leaf time
S. rexii	present	present	absent	Rosulate	N/A	65 DAS
*S. grandis**	absent	absent	present	Unifoliate	N/A	N/A
F1 hybrid	present	present	absent	Rosulate	N/A	65 DAS

* The values for *S. grandis* are averages taken from both lineages. *S. grandis*$^{\mathrm{F1}}$ and *S. grandis*$^{\mathrm{BC}}$

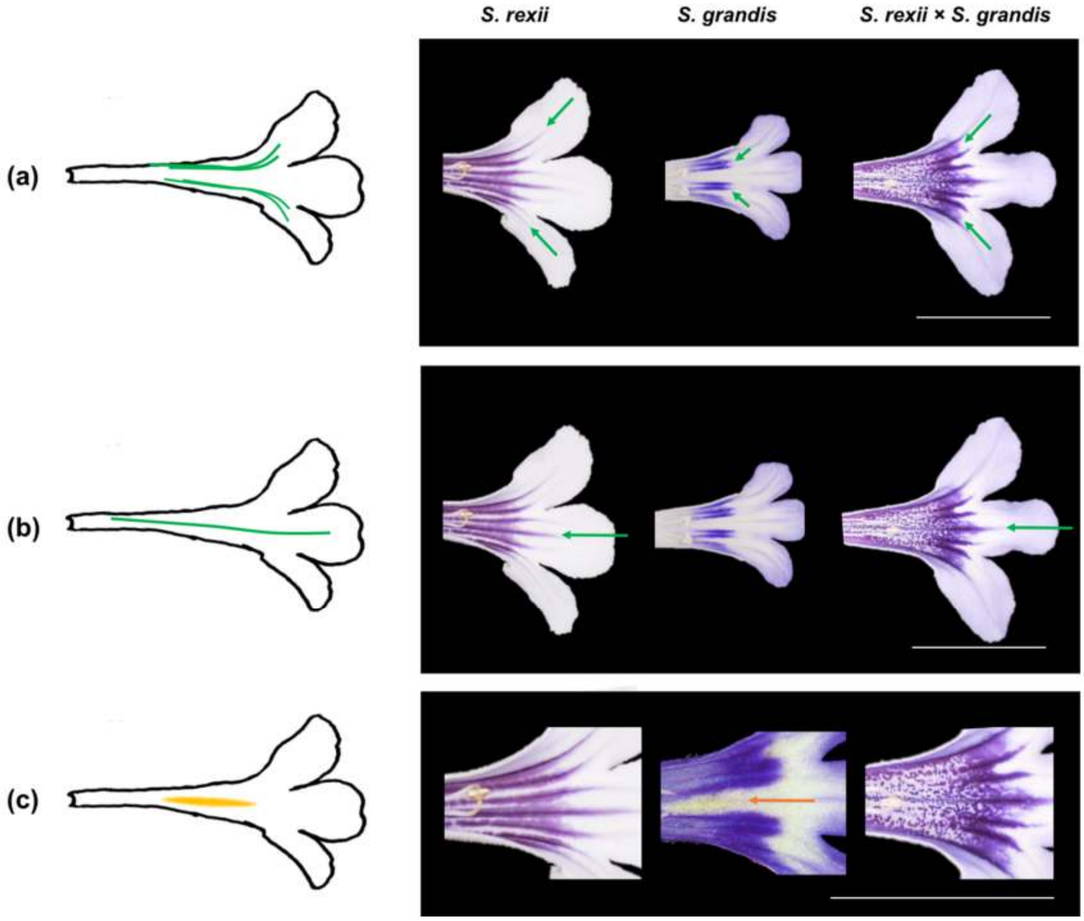


Figure 6.8 Floral pigmentation of the parental lineages. **(a)** Pigmentation on the lateral lobes. **(b)** Pigmentation on the ventral lobe. **(c)** Yellow spot on the ventral side of the corolla tube. Green arrows: pigmentation on the lateral and ventral corolla. Orange arrow: yellow spot on the ventral corolla. Bars = 2.5 cm.

6.3.2 Segregation of morphological variations in the backcross population

The morphology of the 200 BC plants were measured and the segregation patterns were examined (Appendix 6.5, 6.6, 6.7). In terms of the segregation of the vegetative habits, when scoring using Method 1 (i.e. score all ambiguous morphologies as unifoliate) the ratio of rosulate to unifoliate was 107:93 (Table 6.9). The ratio did not fit the expected 3:1 ratio (Chi-square test: P < 0.0001), but conformed to a 1:1 ratio (Chi-square test: P = 0.3222),. Method 2 (i.e. score rosulate and accessory phyllomorphs as rosulate, and others as unifoliate) scoring gave a ratio of rosulate to unifoliate of 142:58, which fitted the expected 3:1 ratio (Chi-square test: P = 0.1914,). Method 3 (i.e. score all ambiguous morphologies as unknown) scoring gave a ratio of 123:25, which deviated from the expected 3:1 ratio (Chi-square test: P = 0.0227) but conformed to a 4:1ratio (Chi-square test: P = 0.3445). The scoring of Method 4 (i.e. score plants with any phyllomorphs produced from the GM as rosulate, and others as unifoliate or unknown) gave a ratio of 154:44, which best fitted a 3:1

ratio (Chi-square test: P = 0.3667). In addition, the ratio of plants with presence:absence of accessory phyllomorphs in the BC population was 104:77, which slightly deviated from the Mendelian ratio of 1:1 (Chi-square test: P = 0.0448).

Table 6.9 Result of rosulate / unifoliate trait scoring in the BC population using four different methods

	N	**Rosulate (R)**	**Unifoliate (U)**	**Unknown**	**R:U ratio**	***P*-value (Chi-square test)**
Method 1	200	107	93	0	1:1	0.3222
Method 2	200	142	58	0	3:1	0.1914
Method 3	200	123	25	52	4:1	0.3445
Method 4	200	154	44	2	3:1	0.3667

Among the 200 BC plants used for genetic mapping, 14 plants did not produce flowers and thus their floral data were not available (Appendix 6.5; i.e. qualifier G, AH, AS, BQ, BY, CE, DJ, DO, DZ, FR, GF, IF, IJ, IQ). The distributions of the measured quantitative traits indicated that 14 showed a normal distribution and another 12 skewed nonparametric distributions (Table 6.10 and Appendix 6.6, 6.7).

The flower pigmentation patterns among the BC plants showed great variation and a gradient of pigmentation intensity and stripiness (Figure 6.9). The variation could roughly be categorised into nearly absence of stripes on the corolla floor (Figure 6.9 a), a pattern that resembled *S. grandis* flowers, with two short double stripes in the throat (Figure 6.9 b), a pattern somewhat more similar to the F1 plant flowers with seven stripes (Figure 6.9 c), and a very intensive pigmentation blotch covering the entire corolla floor (Figure 6.9 d). The least-pigmented phenotype (Figure 6.9 a) and the most intensely pigmented phenotype (Figure 6.9 d) were only observed in the BC population but not in the parental materials (Figure 6.8) and represented new phenotypes. There was some variation in the two more-intensively pigmented phenotypes (i.e. Figure 6.9 c and d) that made their categorisation sometimes difficult.

In terms of segregation patterns, the presence and absence of pigmentation on the lateral lobe segregated as 161:25, a non-Mendelian ratio (i.e. Figure 6.9 a are scored as absence; Figure 6.9 b, c, d are scored as presence). The presence and absence of the ventral lobe pigmentation stripe was 102:85 (i.e. Figure 6.9 a and b are scored as absence; Figure 6.9 c and d are scored as presence), which fitted a Mendelian ratio of 1:1 (Chi square test: P = 0.2138). The yellow spot on the ventral lobe observed in the backcross population (Figure 6.10 a) was sometimes more intensive in colour compared to the *S. grandis* parent (Figure 6.10 b). However, in the case where the purple pigmentation in the corolla tube was very intensive (i.e. Figure 6.9 d), the purple pigmentation may have covered the area where the yellow spot was located, making the yellow spot trait indeterminable. For these plants the yellow spot trait was scored as unknown (Appendix 6.5). The ratio of presence and absence of the yellow spot was 76:110, and deviated from a 1:1 ratio (Chi square test: P = 0.0127). Two macrocotyledons were rarely encountered in a seedling, only in 6 plants among the 200 BC plants observed (Appendix 6.5).

Table 6.10 Summary of the results of the morphology scoring for the BC population

Trait No.	Trait (trait unit)	Average value	Note
1	Corolla length (cm)	4.57 ± 0.56	Normal distribution
2	Undilated tube length (cm)	1.46 ± 0.19	Normal distribution
3	Dilated tube length (cm)	3.28 ± 0.43	Normal distribution
4	Undilated tube height (cm)	0.52 ± 0.07	Non-normal distribution
5	Dilated tube height (cm)	0.80 ± 0.09	Normal distribution
6	Undilated tube width (cm)	0.56 ± 0.08	Non-normal distribution
7	Dilated tube width (cm)	0.90 ± 0.11	Non-normal distribution
8	Corolla face height (cm)	2.65 ± 0.47	Normal distribution
9	Tube opening height (outer) (cm)	1.20 ± 0.20	Non-normal distribution
10	Tube opening height (inner) (cm)	0.97 ± 0.18	Non-normal distribution
11	Corolla face width (cm)	2.99 ± 0.46	Normal distribution
12	Tube opening width (outer) (cm)	1.50 ± 0.22	Non-normal distribution
13	Tube opening width (inner) (cm)	1.01 ± 0.17	Non-normal distribution
14	Pistil length (cm)	3.03 ± 0.25	Normal distribution
15	Ovary length (cm)	1.99 ± 0.20	Normal distribution
16	Style length (cm)	1.03 ± 0.13	Non-normal distribution
17	Calyx length (cm)	0.54 ± 0.13	Non-normal distribution
18	Stamen length (cm)	2.70 ± 0.25	Normal distribution
19	Filament length (attached) (cm)	2.00 ± 0.21	Normal distribution
20	Filament length (detached) (cm)	0.70 ± 0.11	Non-normal distribution
21	Ventral tube length (cm)	3.76 ± 0.46	Normal distribution
22	Ventral lobe length (cm)	0.99 ± 0.16	Normal distribution
23	Dorsal tube length (cm)	3.07 ± 0.33	Normal distribution
24	Dorsal lobe length (cm)	0.91 ± 0.16	Normal distribution
25	Flowering time (DAS)	247 ± 77.36	Non-normal distribution

Table 6.10 continued

Trait No.	Trait (trait unit)	Average value	Note
26	Lateral lobe pigmentation	present:absent = 161:25	Non-Mendelian ratio
27	Ventral lobe pigmentation	present:absent = 102:85	Mendelian ratio = 1:1
28	Yellow spot	present:absent = 76:110	Non-Mendelian ratio
30	Accessory phyllomorph	present:absent = 104:77	Non-Mendelian ratio
31	Two macrocotyledons	present:absent = 6:193	Non-Mendelian ratio

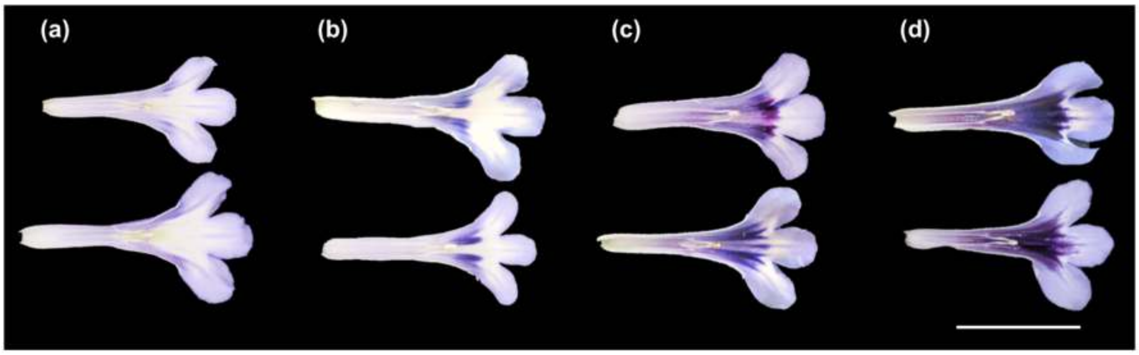


Figure 6.9 Examples of the floral pigmentation patterns observed in the BC population. Flowers were dissected by cutting between the lateral and dorsal corolla lobes on both sides. The lower parts of the corolla including the lateral and ventral lobes are shown. **(a)** Flower lacking stripe pigmentation on both lateral and ventral lobes. **(b)** Flower with four short lateral stripes on the three lobes, lacking the middle stripes. **(c)** Flower with seven long stripe pigmentation on the lateral and ventral lobes. **(d)** Flower with seven intensive stripe pigmentation showing as a blotch. Bar = 2.5 cm.

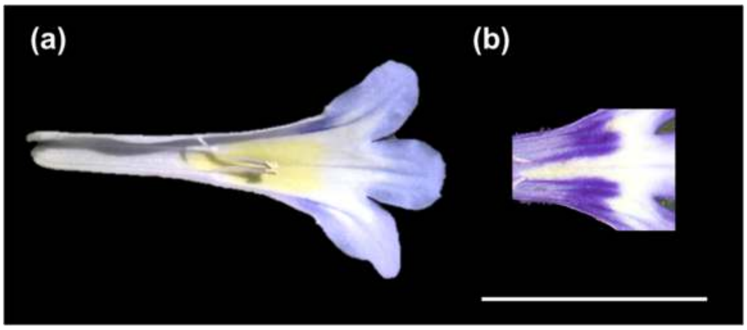


Figure 6.10 Examples of the yellow spot phenotype observed in **(a)** BC plants, and in **(b)** parental lineages. Bar = 2.5 cm.

Analysis of the correlations between the measured traits indicated that all floral dimension traits were significantly positively correlated with each other. The correlation of ‘style length’ and ‘filament length, detached’ was less significant (Figure 6.11; traits 16 and

20). On the other hand, the ‘flowering time’ (trait 25) was negatively correlated with floral dimensions (i.e. the later the plant flowered the smaller the flower size). At the same time ‘flowering time’ was positively correlated with all four vegetative habit scoring methods (traits 29-32; i.e. early flowering plants were more likely to be rosulate). The two floral pigmentation traits, ‘lateral’ and ‘ventral pigmentation’ (traits 26 and 27), were positively correlated to each other. But the two traits were negatively correlated with the ‘yellow spot’ (trait 28). The ‘ventral pigmentation’ trait was also positively correlated to ‘tube opening height’ (traits 9 and 10).

In terms of vegetative traits, the four vegetative habit scoring methods were all positively correlated with each other (Figure 6.11; traits 29 - 32). The presence of ‘two macrocotyledons’ (trait 34) was positively correlated only with Method 1 scoring, but not with other traits. In addition, various correlations were found between the vegetative traits and floral dimensions, such as a positive correlation between scoring Method 4 and ‘tube opening height, outer’ trait (trait 9).

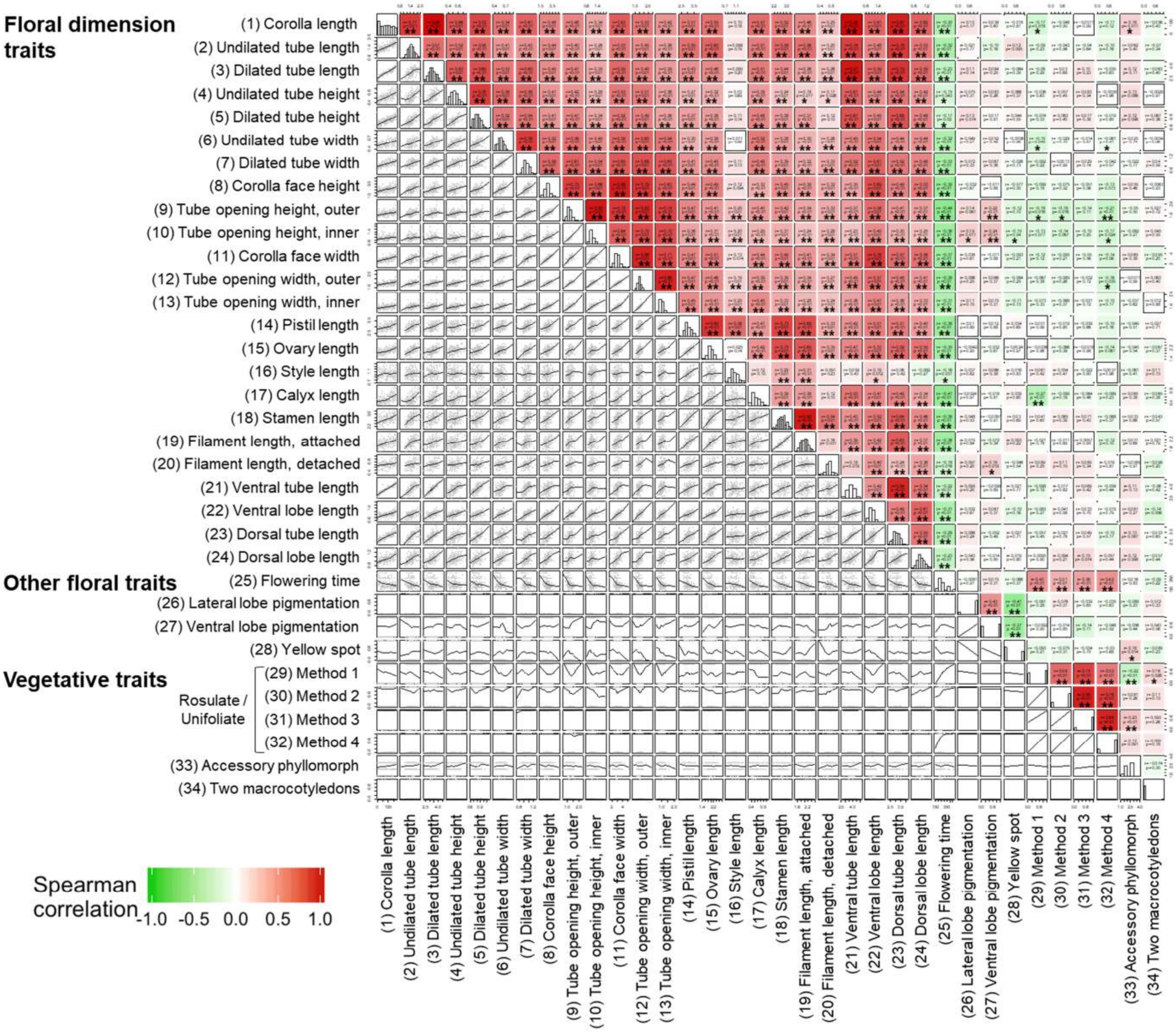


Figure 6.11 Pairwise correlation comparisons of the measured traits in the BC population. The graph has three parts: the **upper triangle** (coloured correlations), the **diagonal** (with histograms), and the **lower triangle** (scatter plots and trend lines). The colour scale at the bottom left of the graph shows the degree of correlation, with positive correlations in red, and negative correlations in green. The **upper triangle** shows the degree of correlation, the Spearman correlation coefficient (r), and *P*-values. Asterisks indicate the degree of significance of the correlation (**P* ≤ 0.05, ** *P* ≤ 0.01). **Diagonal line** is composed of histograms of the distribution of each of the trait values measured in the BC population. **Lower triangle** shows scattered plots of the measured trait values (grey dots) and polynomial regression lines (black).

6.3.3 Mapping of the rosulate / unifoliate loci

In the OTL analysis using MapA, BTL signals were detected on LG1, LG7, and LG9 (Table 6.11; Figure 6.12 a). The maximum LOD score of the detected loci ranged from 3.24 (Method 4, LG9) to 6.28 (Method 2, LG9) while the calculated LOD threshold was between 3.05 and 3.10. Among the detected loci, the one on LG9 was detected in the analysis of all four scoring methods (Figure 6.12 d). Mapping of scoring Method 2 and 4 detected a signal on LG1 (Figure 6.12 b), and only the mapping of Method 4 scoring detected a signal on LG7 (Figure 6.12 c). The confidence intervals varied from the smallest of 12.9 cM (Method 4, LG1) to as large as the whole linkage group (148.04 cM; Method 2, LG1). The highest percentage of variance explained was obtained in the mapping using scoring Method 4 (24.77%, Table 6.11). The CIM results showed similar LOD curves to those of SIM (Figure 6.13), except for the mapping of scoring Method 4 where only the locus on LG7 was detected (Figure 6.13 d). The BTL signals found in CIM were identical with those found by SIM and no additional BTL region was discovered (Figure 6.13). The effect plots of the detected loci are summarised in Appendix 6.8 a.

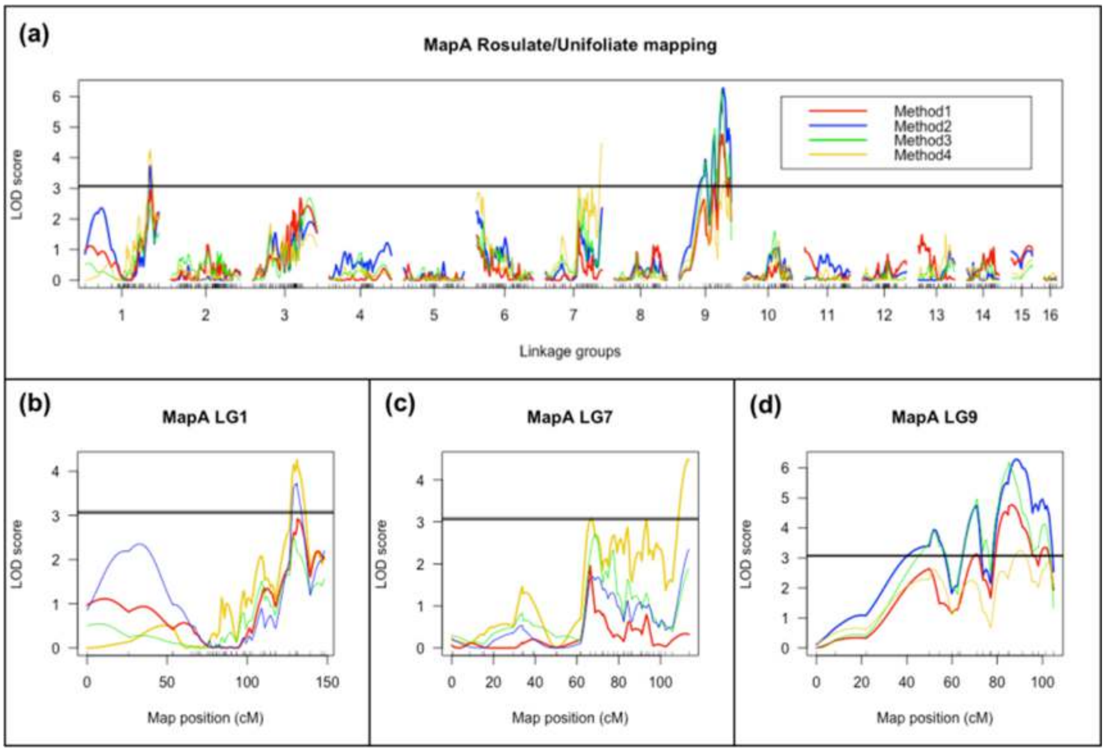


Figure 6.12 LOD curves of the rosulate / unifoliate SIM results using MapA. **(a)** LOD curve by standard interval mapping of all linkage groups. **(b)** LOD curve of LG1. **(c)** LOD curve of LG7. **(d)** LOD curve of LG9. Red: LOD score of scoring Method 1. Blue: LOD score of scoring Method 2. Green: LOD score of scoring Method 3. Yellow: LOD score of scoring Method 4. Black horizontal lines: LOD score thresholds.

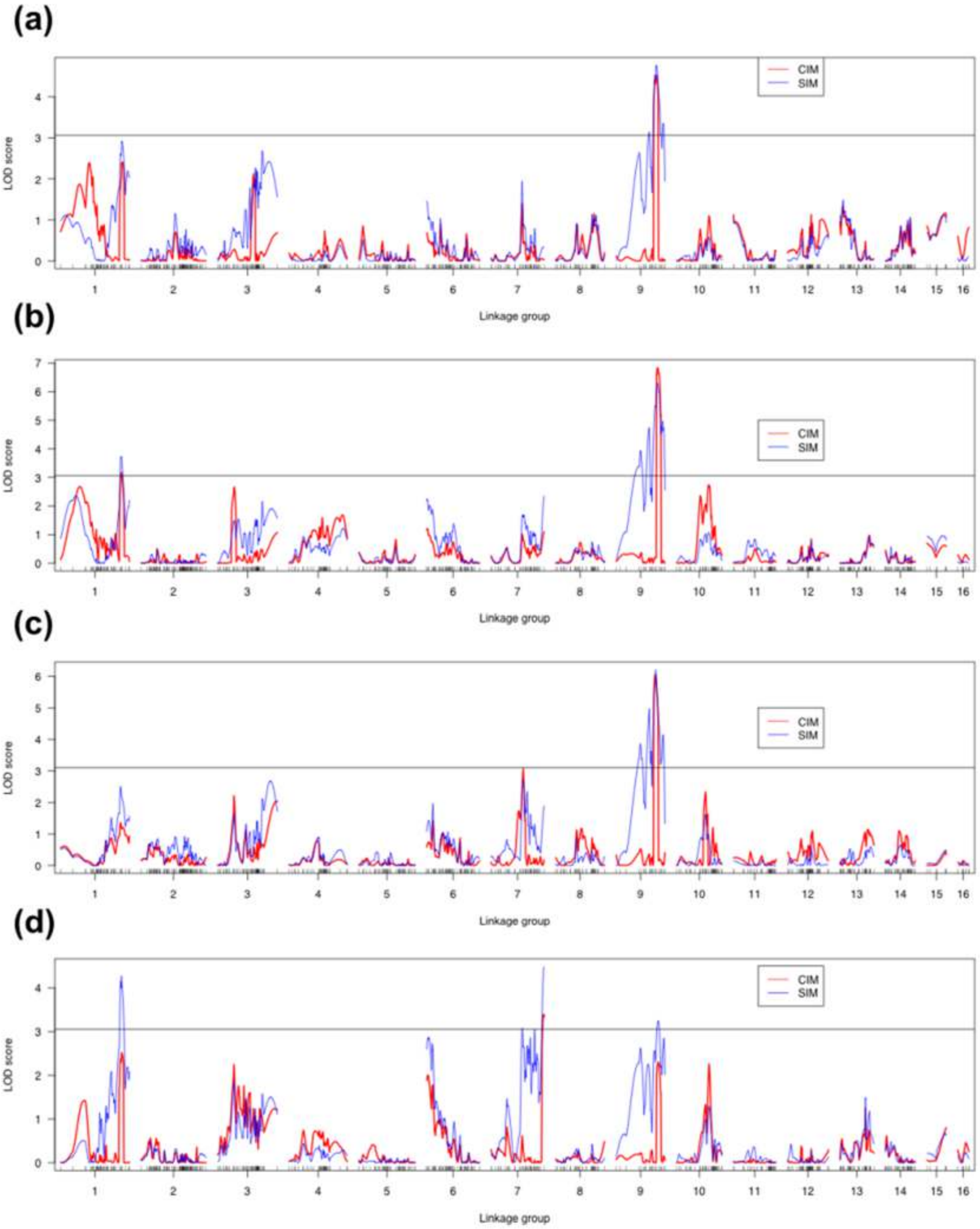


Figure 6.13 LOD curves of the rosulate / unifoliate CIM results using MapA. **(a)** Scoring Method 1. **(b)** Scoring Method 2. **(c)** Scoring Method 3. **(d)** Scoring Method 4. Red lines: CIM LOD curves. Blue lines: SIM LOD curves.

Slightly different results were obtained in the SIM using MapB-1, with one additional BTL site found on LG10 (Table 6.12). Overall, BTL signals were detected in LG2, LG4, LG10, and LG14 (Figure 6.14 a), equivalent to MapA LG1, LG7, LG6, and LG9, respectively. The LOD scores obtained in MapB-1 mapping were lower than those of MapA, with the lowest value of 3.17 (Method 4, LG2) and highest of 4.94 (Method 2, LG14). The

LOD thresholds obtained were the same as the mapping in MapA, ranging from 3.05 to 3.10 (Table 6.12). The BTL loci found in LG14 (equivalent to MapA LG9) were again detected in the mapping of all four scoring methods (Figure 6.14 e). Scoring Method 4 also detected the loci on LG2 (equivalent to MapA LG1; Figure 6.14 b) and LG4 (equivalent to MapA LG7; Figure 6.14 c), and the additional locus on LG10 (equivalent to MapA LG6; Figure 6.14 d) which was not found previously. The size of the confidence intervals varied from 8.65 cM (Method 4, LG10) to 49.01 cM (Method 4, LG14). The highest percentage of variance explained was obtained in the mapping using scoring Method 4, and was higher than the result when using MapA (38.66%; Table 6.12). The CIM results of the MapB-1 mapping showed similar LOD curves to SIM and the same BTL regions were identified (Figure 6.15). The CIM of scoring Method 1, 2, and 3 received the maximum LOD scores of 3.98, 5.12, and 4.75 respectively, higher than the SIM results (Figure 6.15 a). But CIM of scoring Method 4 detected no BTL at all (Figure 6.15 d). The effect plots of the detected loci are summarised in Appendix 6.8 b.

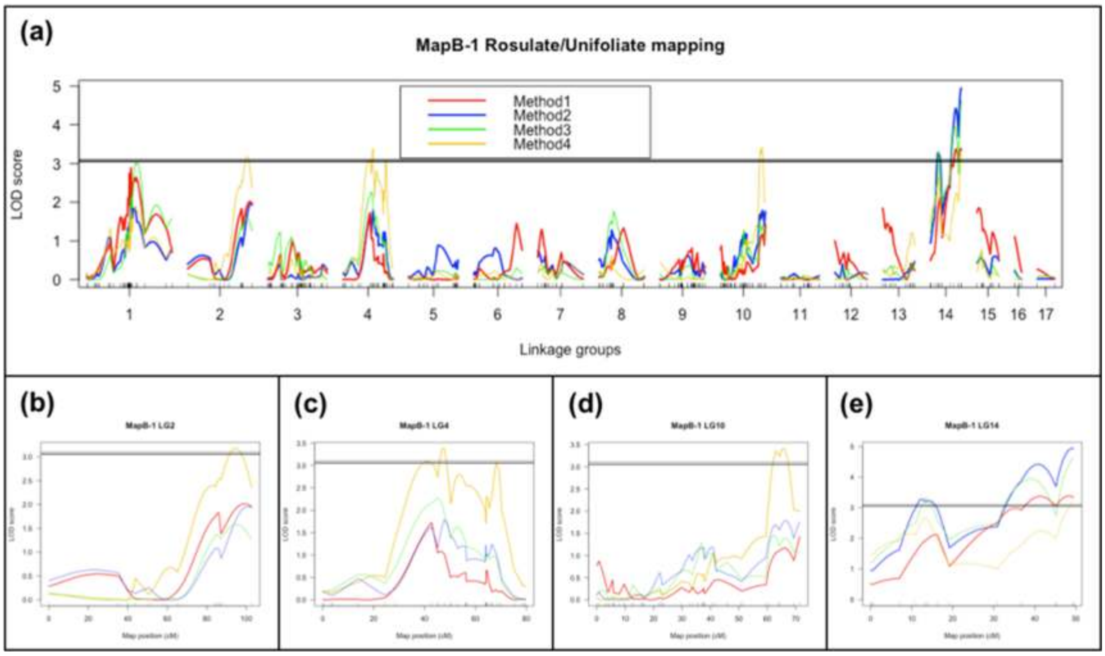


Figure 6.14 LOD curves of the rosulate / unifoliate SIM results of the MapB-1. (a) LOD curves by SIM of all linkage groups. (b) LOD curves of the LG2. (c) LOD curves of the LG4. (d) LOD curves of the LG10. (e) LOD curves of the LG14. Red: LOD curve of Method 1. Blue: LOD curve Method 2. Green: LOD curve of Method 3. Yellow: LOD curve of Method 4. Black horizontal lines: LOD score threshold.

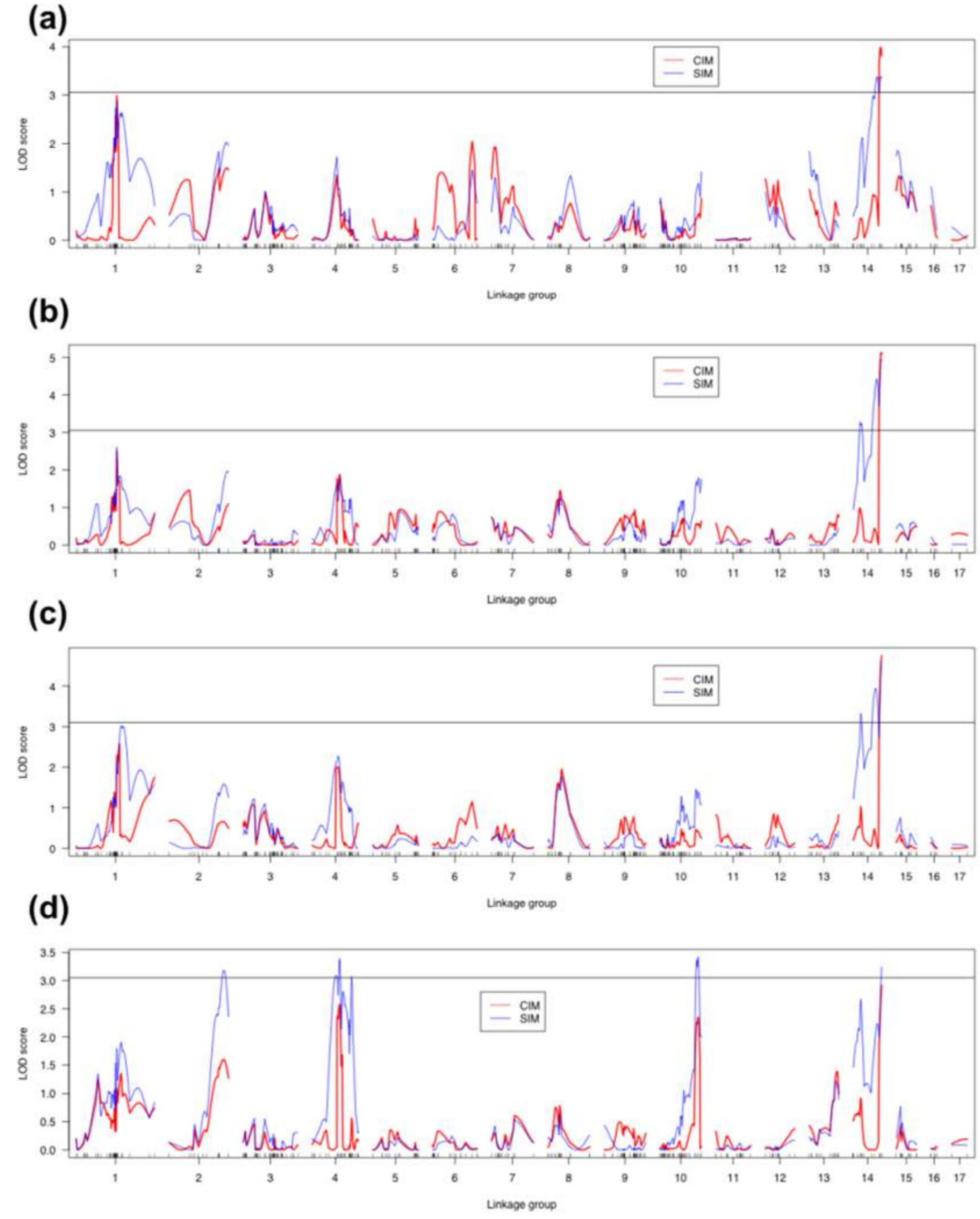


Figure 6.15 LOD curves of the rosulate / unifoliate CIM results of the MapB-1. (a) Scoring Method 1. (b) Scoring Method 2. (c) Scoring Method 3. (d) Scoring Method 4. Red lines: CIM LOD curves. Blue lines: SIM LOD curves.

The mapping results using MapB-3 were more similar to those of MapA, with BTL signals detected on LG2, LG4, and LG14 (Table 6.13; Figure 6.16 a). The signal in LG14 (equivalent to MapA LG9) was again detected in all four scoring methods (Figure 6.16 d). The signal in LG2 was detected in both scoring Method 2 and 4 (Figure 6.16 b), and the signal in LG4 was only detected in scoring Method 4 (Figure 6.16 c). The LOD score obtained was similar to MapB-1 and lower than that of the mapping in MapA, with the lowest LOD score of 3.33 (Method 4, LG14) and highest LOD score of 5.22 (Method 2, LG14). The size of the confidence intervals ranged from 16.06 cM (Method 2, LG2) to 53.09 cM (Method 4, LG14). The highest percentage of variance explained was obtained in the mapping using scoring Method 4, but the value was lower than in the results of both MapA and MapB-1 (Table 6.13; 19.72%). The CIM results showed a slightly different pattern (Figure 6.17), and in scoring Method 1 an additional locus was found on LG1 (corresponding to MapA LG3) with a maximum LOD score of 4.19 (Figure 6.17 a). This region gave a very narrow confidence interval of 6.93 cM, and together with the LG14 locus detected in SIM they explained 17.15% of the trait variance (Table 6.13). No additional locus was found in other CIM results (Figure 6.17). The effect plots of all the detected loci are summarised in Appendix 6.8 c.

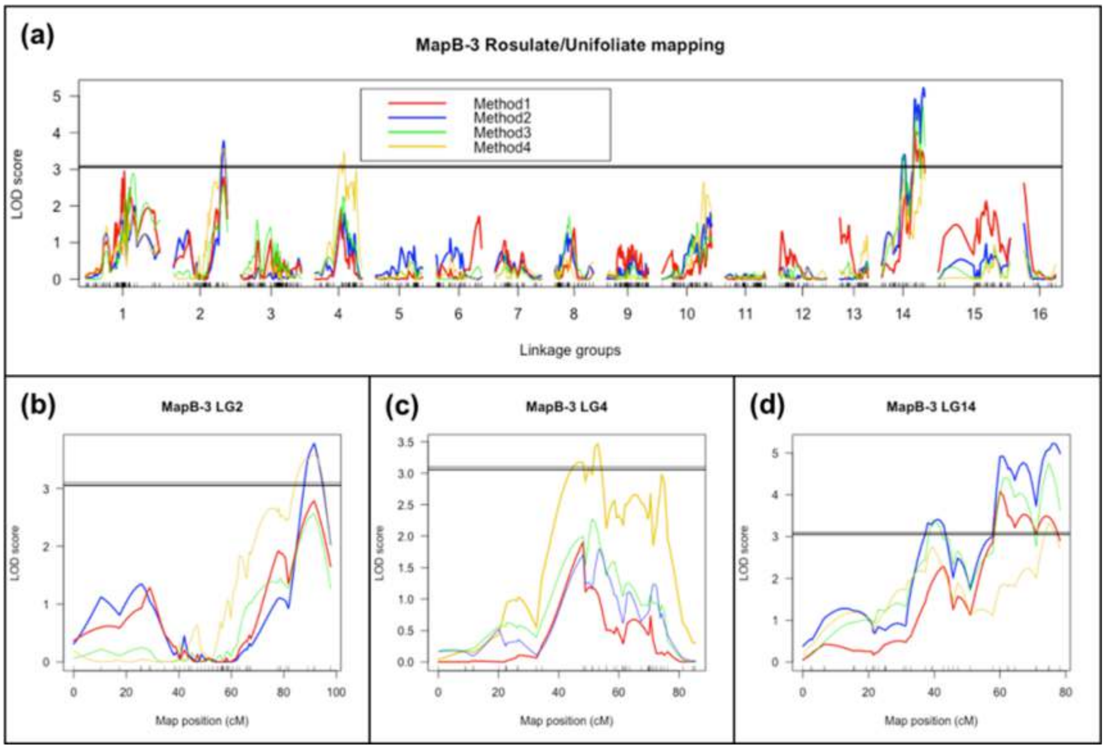


Figure 6.16 LOD curves of the rosulate / unifoliate mapping results of the MapB-3. (a) LOD score by standard interval mapping of all linkage groups. (b) LOD score of the LG2. (c) LOD score of the LG4. (d) LOD score of the LG14. Red: LOD score of Method 1. Blue: LOD score Method 2. Green: LOD score of Method 3. Yellow: LOD score of Method 4. Black horizontal lines: LOD score threshold.

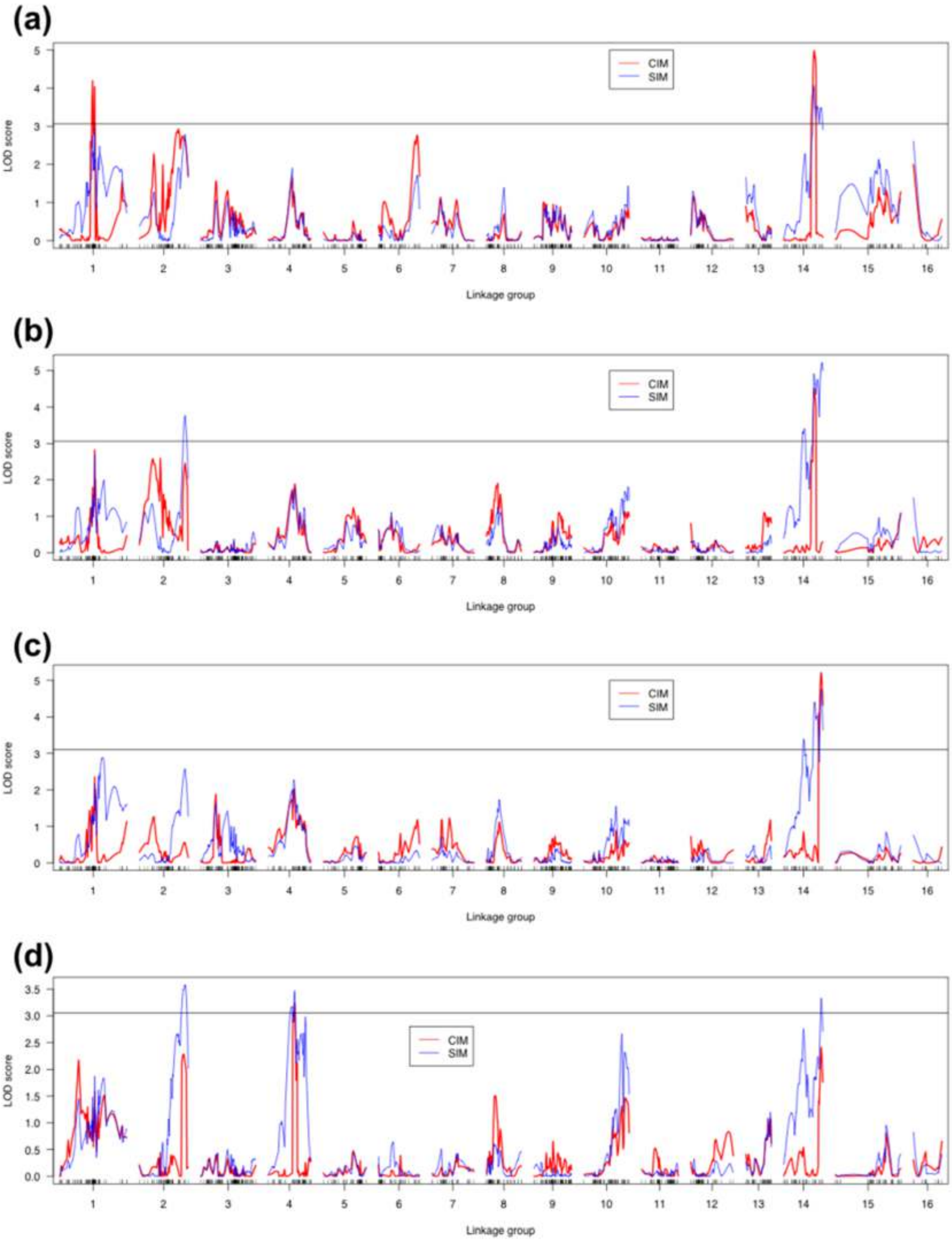


Figure 6.17 LOD curves of the rosulate / unifoliate CIM results of the MapB-3. (a) Scoring Method 1. (b) Scoring Method 2. (c) Scoring Method 3. (d) Scoring Method 4. Red lines: CIM LOD curves. Blue lines: SIM LOD curves.

Table 6.11 Mapping of the rosulate / unifoliate loci on MapA

	Marker name	LG	Marker position (cM)	LOD threshold	Max LOD score	Bayes CI* (cM)	CI* size (cM)	Variance explained† (%)
Method 1	c9.loc86	9	86	3.06	4.77	77.12 – 104.94	27.81	10.68
Method 2	ST16952	1	131	3.06	3.73	0.00 – 148.04	148.04	19.56
	c9.loc88	9	88	3.06	6.28	80.56 – 101.20	20.63	
Method 3	BW18918	9	85	3.10	6.20	63.54 – 91.18	27.64	17.55
Method 4	ST16952	1	131	3.05	4.27	126.43 – 139.33	12.90	24.77
	ST8585	7	114	3.05	4.48	61.75 – 113.64	51.89	
	c9.loc90	9	90	3.05	3.24	21.89 – 104.94	83.04	

* CI: confidence interval. † Showing the combined variance explained of all the loci and inter-loci interactions

Table 6.12 Mapping of the rosulate / unifoliate loci on MapB-1

	Marker name	LG	Marker position (cM)	LOD threshold	Max LOD score	Bayes CI* (cM)	CI* size (cM)	Variance explained† (%)
Method 1	C14.loc48	14	48	3.06	3.38	19.19 – 49.45	30.25	7.72
Method 2	C14.loc49	14	49	3.06	4.94	30.87 – 49.45	18.58	11.05
Method 3	BW5742	14	49.3	3.10	4.62	10.99 – 49.45	38.46	13.68
Method 4	C2.loc95	2	95	3.05	3.17	65.12 – 102.78	37.66	38.66
	C4.loc47	4	47	3.05	3.38	28.66 – 70.23	41.57	
	C10.loc66	10	66	3.05	3.41	60.66 – 69.32	8.65	
	ST6585	14	49.45	3.05	3.23	0.44 – 49.01	49.01	

Table 6.13 Mapping of the rosulate / unifoliate loci on MapB-3

	Marker name	LG	Marker position (cM)	LOD threshold	Max LOD score	Bayes CI* (cM)	CI* size (cM)	Variance explained† (%)
Method 1	BW5121[1]	1	4.19	3.06	4.19	64.64 – 71.57	6.93	17.15
	BW1599	14	59.99	3.06	4.07	57.70 – 78.20	20.49	
Method 2	DN2208	2	91.61	3.06	3.77	81.82 – 97.88	16.06	16.84
	C14.loc76	14	76	3.06	5.22	59.99 – 78.20	18.20	
Method 3	ST6585	14	74.93	3.10	4.74	39.45 – 78.20	38.74	13.95
Method 4	C2.loc92	2	92	3.05	3.57	67.47 – 97.88	30.41	19.72
	C4.loc53	4	53	3.05	3.47	34.54 – 53.00	18.45	
	ST6585	14	74.93	3.05	3.33	25.11 – 78.20	53.09	

[1] The marker BW5121 on LG1 was found in CIM results

6.3.4 Mapping of the floral and other vegetative trait loci

Since MapB-1 was constructed under the most stringent filtering condition and the above mapping analysis showed an overall similar pattern in all three genetic maps, the mapping of other traits was performed solely on MapB-1.

Table 6.14, Figure 6.18 and Figure 6.19 summarise the standard interval mapping results. In terms of floral characters, most had 1 to 4 QTLs associated with dimension traits. Exceptions were the ‘Dilated tube height’, ‘Style length’, and ‘Filament length (detached)’, which no QTL was found (Table 6.14). The QTLs of corolla-length related traits (i.e. corolla length, dilated and undilated tube length) were only found on LG2, with small effect sizes between 9.29% and 12.47% variance explained. The QTLs of tube height and tube width (i.e. dilated and undilated tubes) were found on LG1, LG2, and LG14, with small effects between 7.7% and 19.36% variance explained. The QTL for corolla face-related traits (i.e. corolla face height, tube opening height, corolla face width, and tube opening width) were found in LG2, LG3, LG6, and LG9, again with low proportions of variance explained between 5.47% and 18.53%. QTLs of pistil length-related traits (i.e. pistil and ovary length) located differently from those of the corolla-related traits, which were found on LG2, LG6, LG7, LG9, and LG14. In particular, the four loci identified for ‘Pistil length’ explained 39.85% of the phenotypic variance, which was highest among the floral dimension traits examined. QTLs of stamen-related traits (i.e. stamen and filament length) were found on LG2, LG7, and LG12. Finally, the QTLs of tube and lobe length-related traits (i.e. length of dorsal and ventral tube / lobe) were found on LG2, LG6, LG8, LG9, and LG12. Overall, the floral dimension-related QTLs were distributed across 9 linkage groups, including LG1, LG2, LG3, LG6, LG7, LG8, LG9, LG12, and LG14. In particular, QTLs on LG2 and LG14 were detected for multiple traits: LG2 locus was reported among 20 floral dimension traits, and LG14 locus was reported in 4 floral dimension traits (Table 6.15). The Bayes confidence interval of the detected QTLs ranged from 15.56 cM (LG2 locus, Table 6.14) up to 72.48 cM (LG7 locus for the stamen length; Table 6.14). However, the percentage of trait variance explained was low in most of the traits mapped (i.e. ~5% - 25%; Table 6.14). Two exceptions were the QTLs for ‘Pistil length’ and ‘Dorsal lobe length’, with the detected QTLs explaining 39.85% and 30.79% of the variance, respectively (Table 6.14).

The mapping of the ‘Flowering time’ trait identified 3 QTLs with major effects that explained 50.88% of the trait variance (Table 6.14). The 3 QTLs were located on LG1, LG2 and LG14, with the locus on LG2 showing a high LOD score of 14.24. The region overlapped with other floral dimension traits (Figure 6.18). In particular, the confidence regions of the LG2 and LG14 QTLs were relatively specific, with a size of 15.56 cM and 13.23 cM respectively. On the other hand, the confidence interval on LG1 was less specific (75.43 cM).

QTLs related to the pigmentation traits were detected on the LG3, LG9 and LG10 (Table 6.14). For the pigmentation on the lateral lobes, two BTLs on LG3 and LG10 were identified with the highest LOD score of 7.29 and 6.36. The two loci have relatively specific confidence interval size of 28.56 cM and 6.15 cM, respectively, and contributed to 29.92% of the trait variance (Table 6.14). The locus on LG3 showed a very high LOD score of 37.21 and a specific confidence interval size of 18.28 cM (Table 6.14). This locus also contributed to high proportion of 59.94% of the trait variance, the highest percentage found in this study (Table 6.14). Finally, the LG3 and LG10 loci with an additional LG9 locus were found correlated to the yellow spot trait (Table 6.14). The LG10 locus showed a particularly high LOD score of 11.53, and a specific confidence interval size of 8.4 cM. The three loci combined contributed 45.58% of the trait variance.

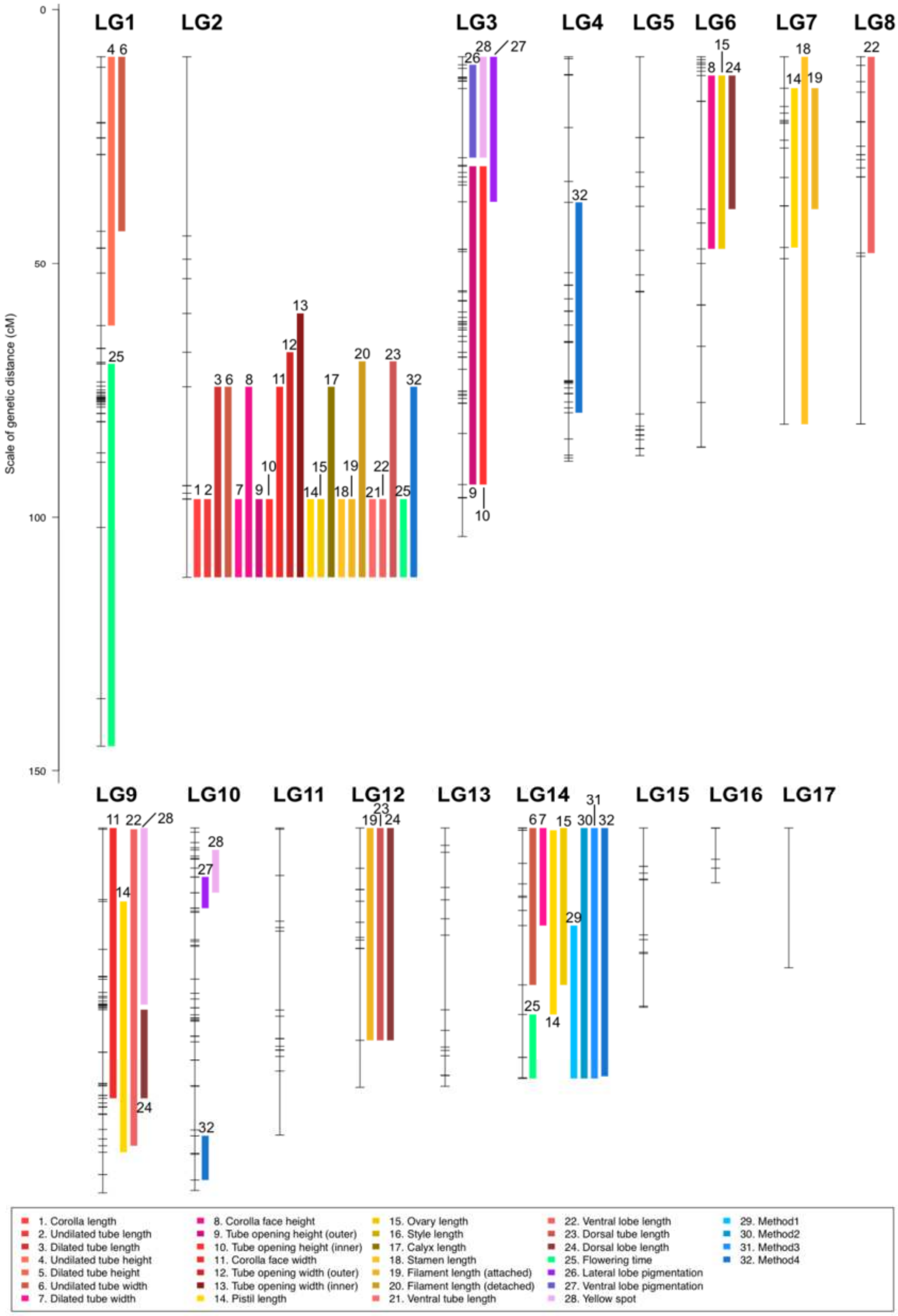
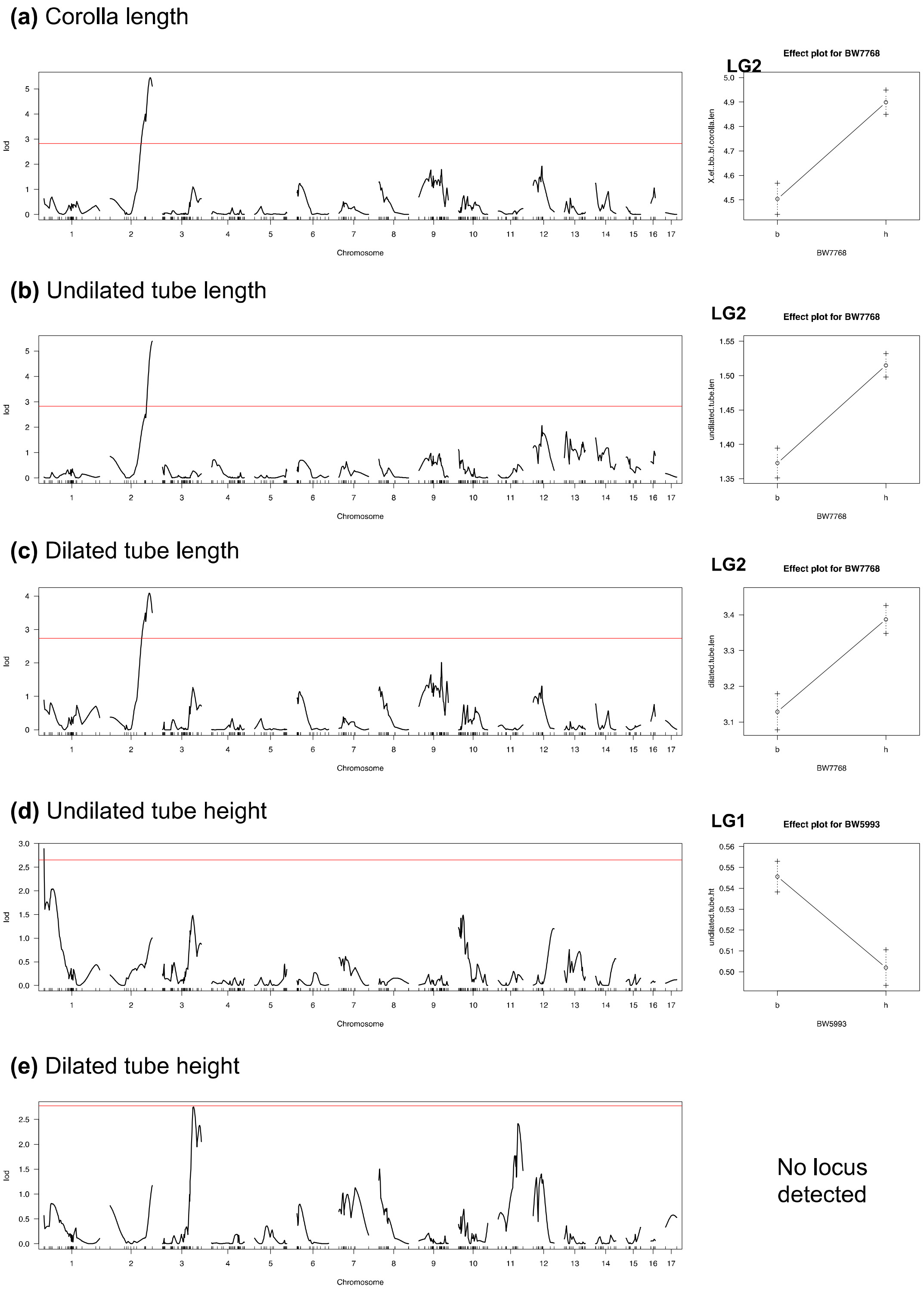


Figure 6.18 Summary of the QTL / BTL confidence intervals identified in the MapB-1.



(a) Corolla length

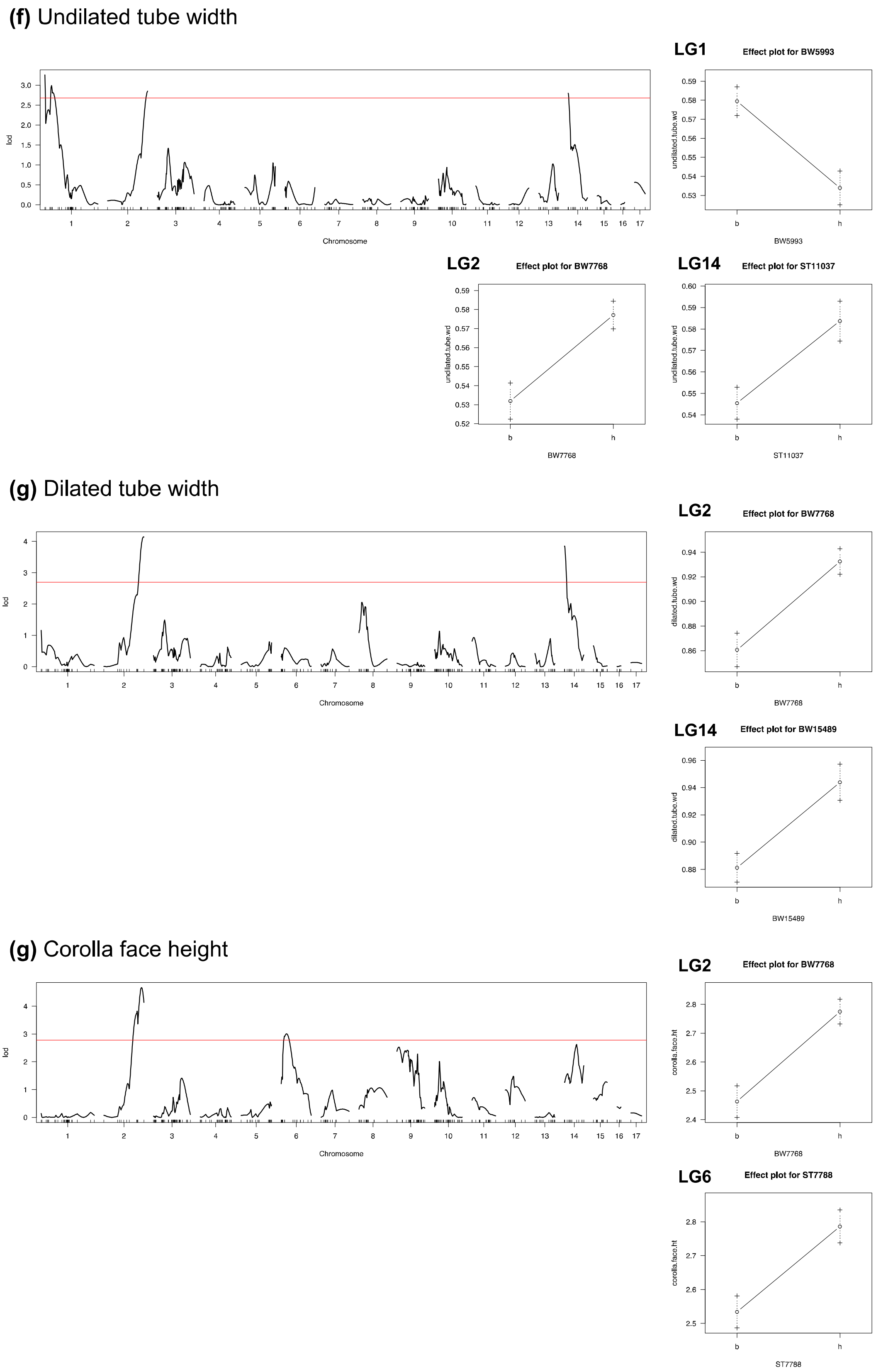
(b) Undilated tube length

(c) Dilated tube length

(d) Undilated tube height

(e) Dilated tube height

Figure 6.19 LOD curves and effect plots of the standard interval mapping on MapB-1. Full legend given on page 250.

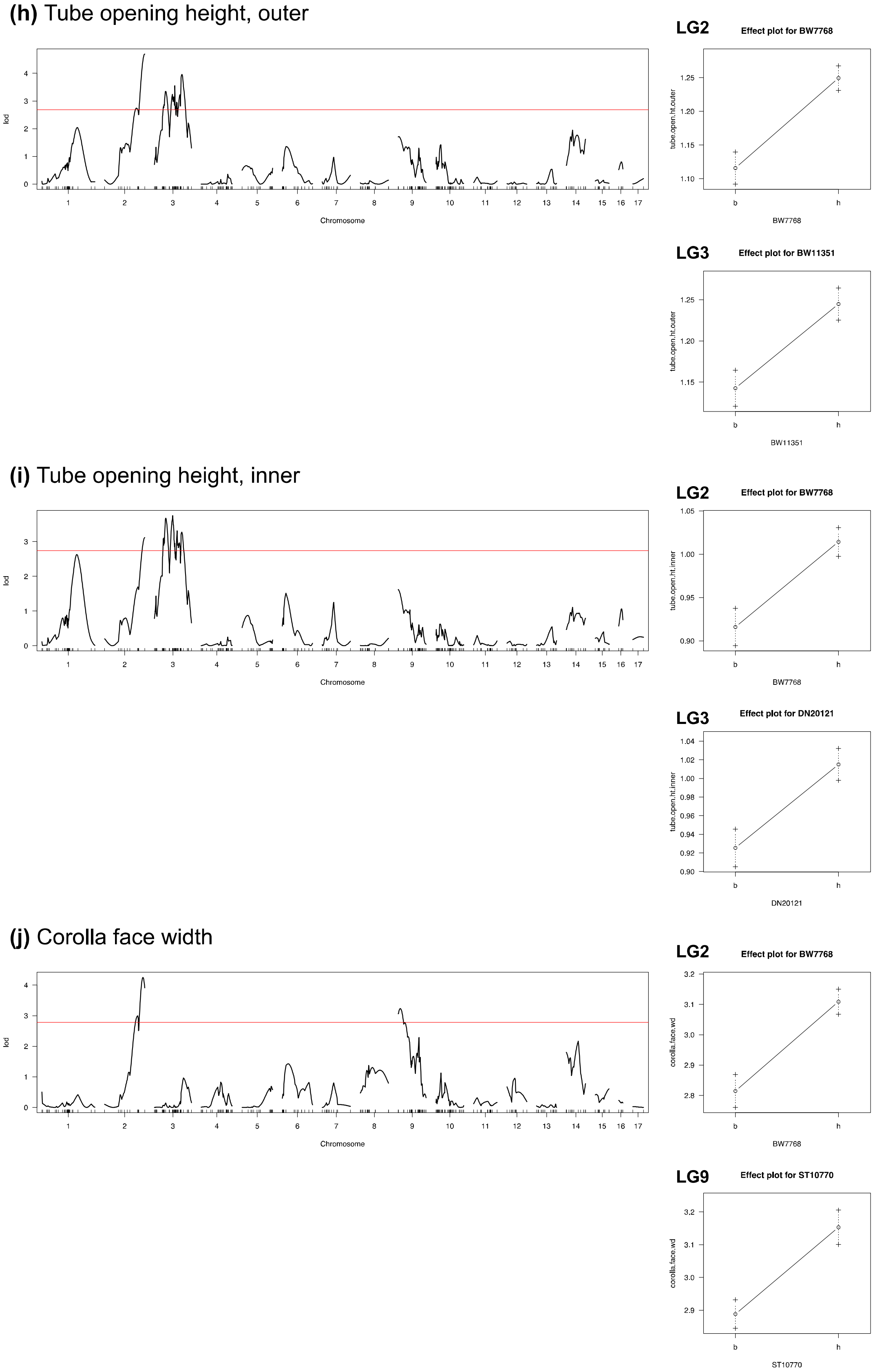


(f) Undilated tube width

(g) Dilated tube width

(g) Corolla face height

Figure 6.19 LOD curves and effect plots of the standard interval mapping on MapB-1. Full legend given on page 250.

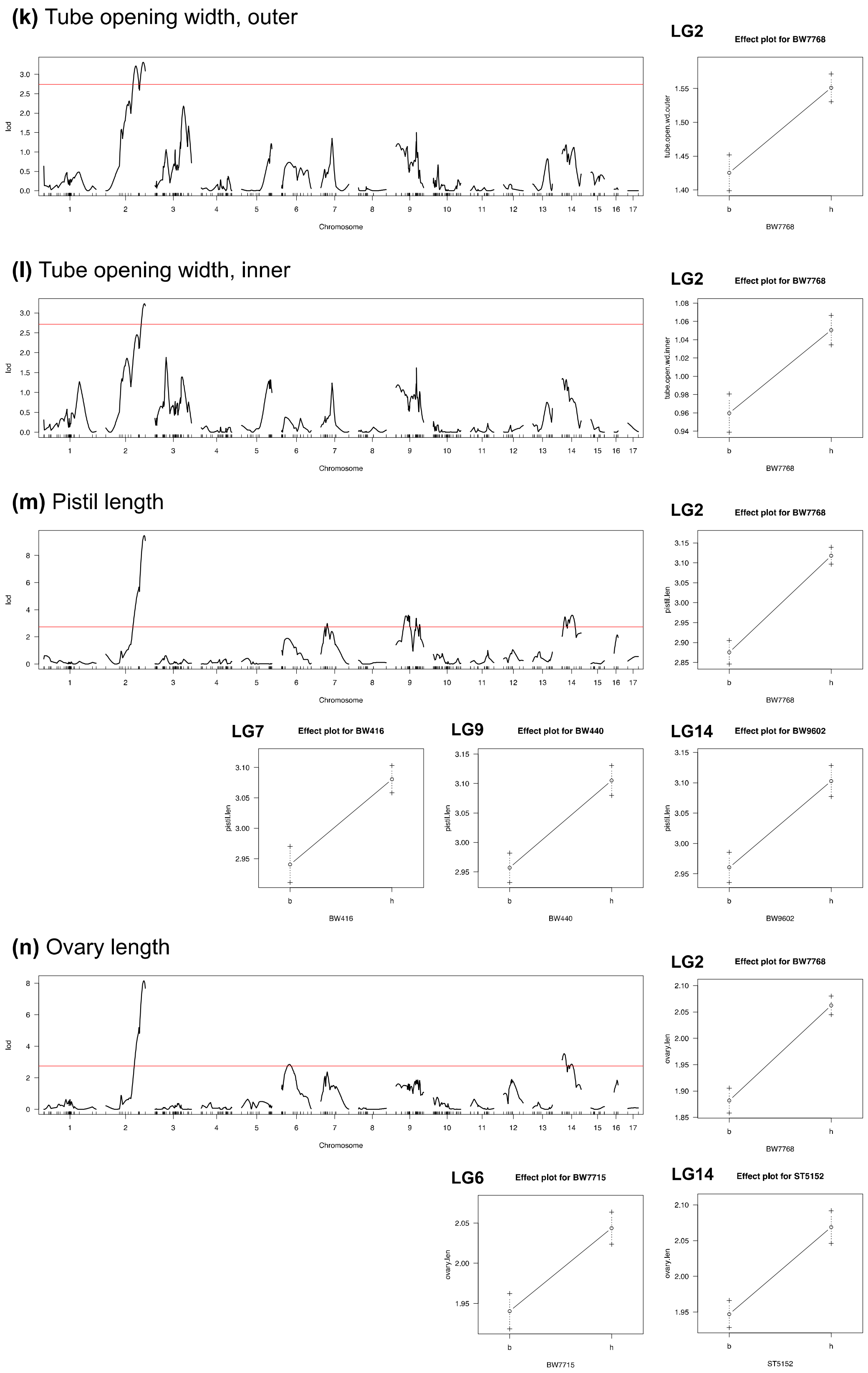


(h) Tube opening height, outer

(i) Tube opening height, inner

(j) Corolla face width

Figure 6.19 LOD curves and effect plots of the standard interval mapping on MapB-1. Full legend given on page 250.



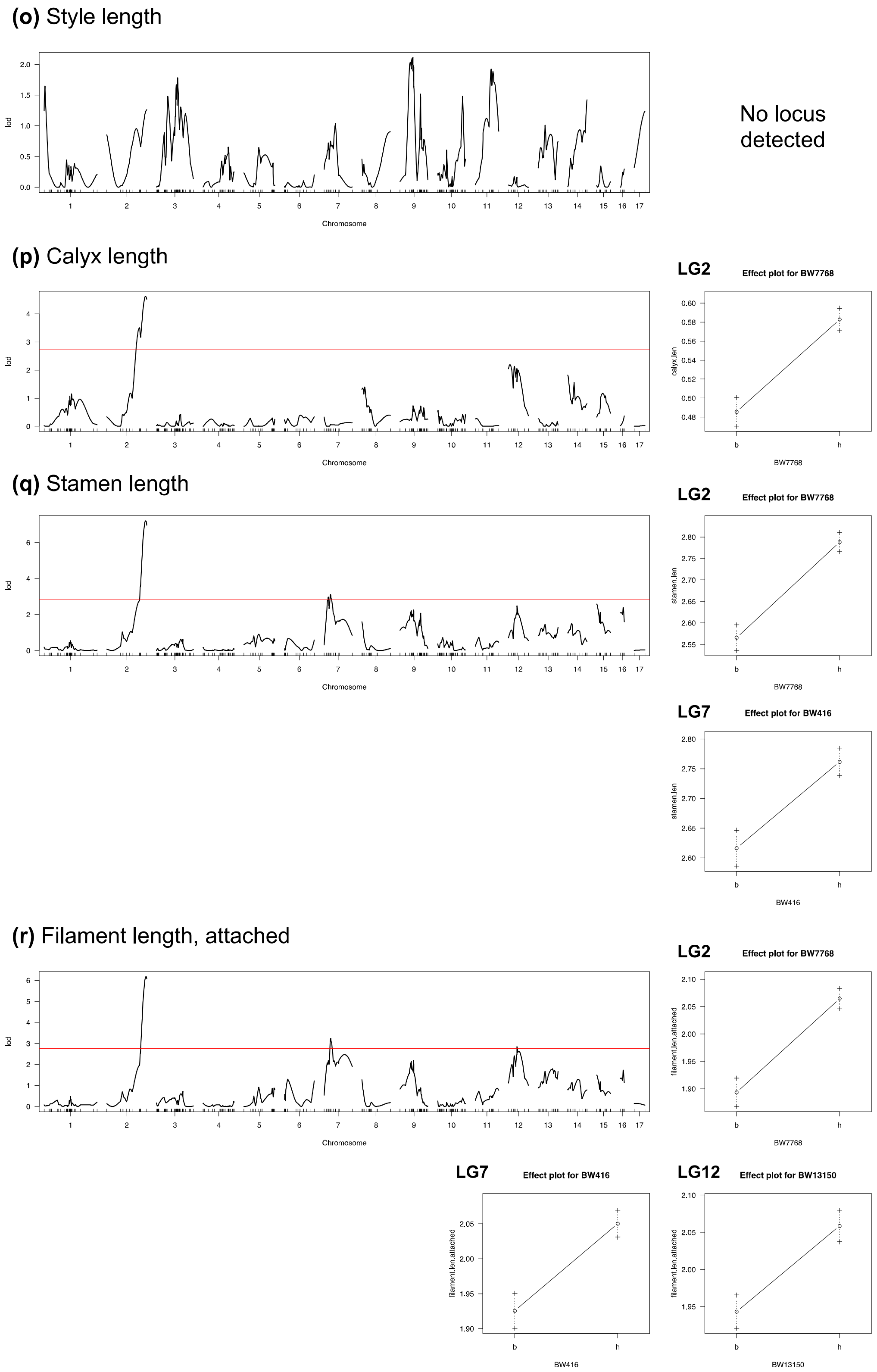
(k) Tube opening width, outer

(l) Tube opening width, inner

(m) Pistil length

(n) Ovary length

Figure 6.19 LOD curves and effect plots of the standard interval mapping on MapB-1. Full legend given on page 250.


(o) Style length

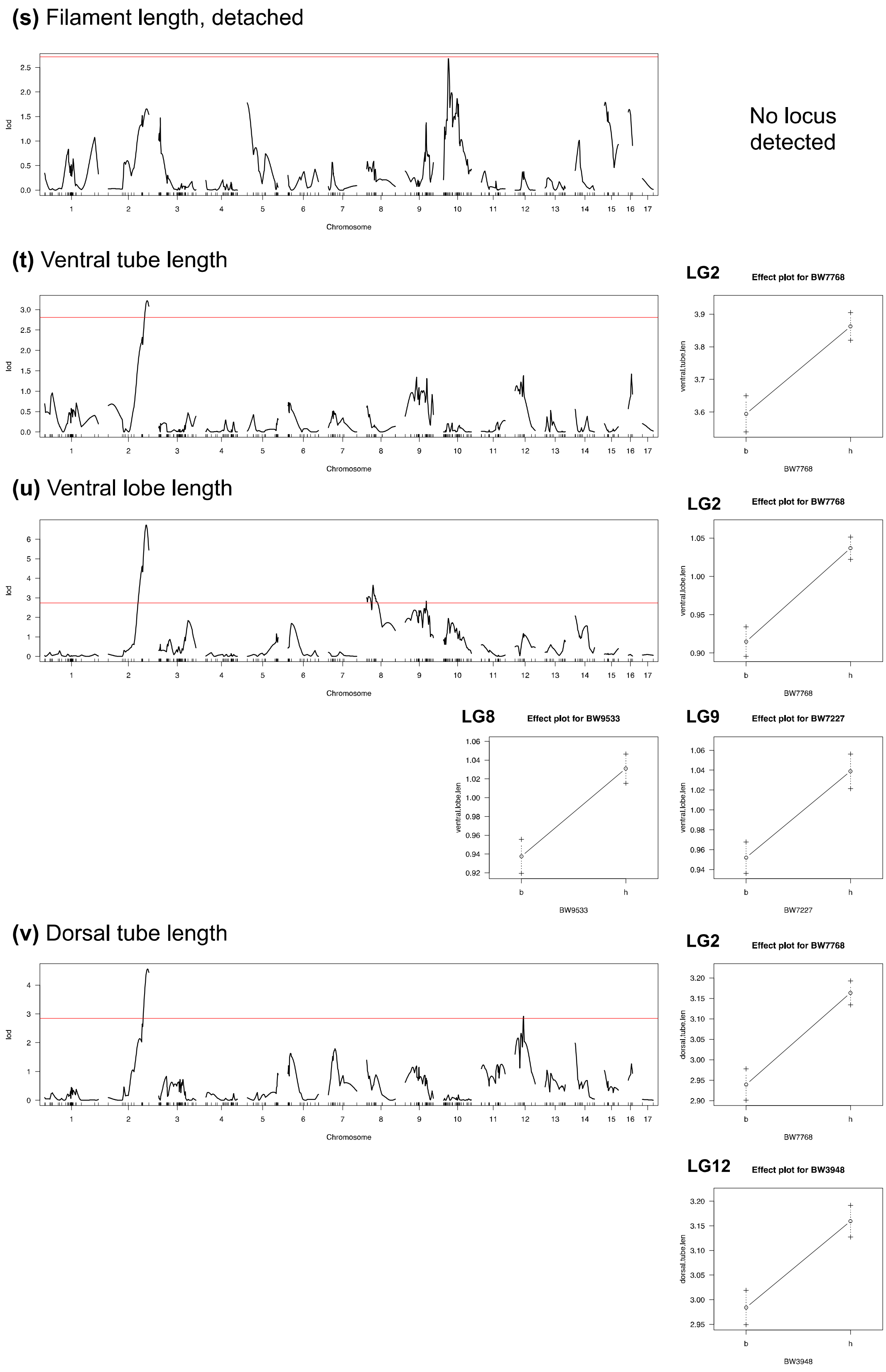
No locus detected

(p) Calyx length

(q) Stamen length

(r) Filament length, attached

Figure 6.19 LOD curves and effect plots of the standard interval mapping on MapB-1. Full legend given on page 250.



(s) Filament length, detached

(t) Ventral tube length

(u) Ventral lobe length

(v) Dorsal tube length

Figure 6.19 LOD curves and effect plots of the standard interval mapping on MapB-1. Full legend given on page 250.

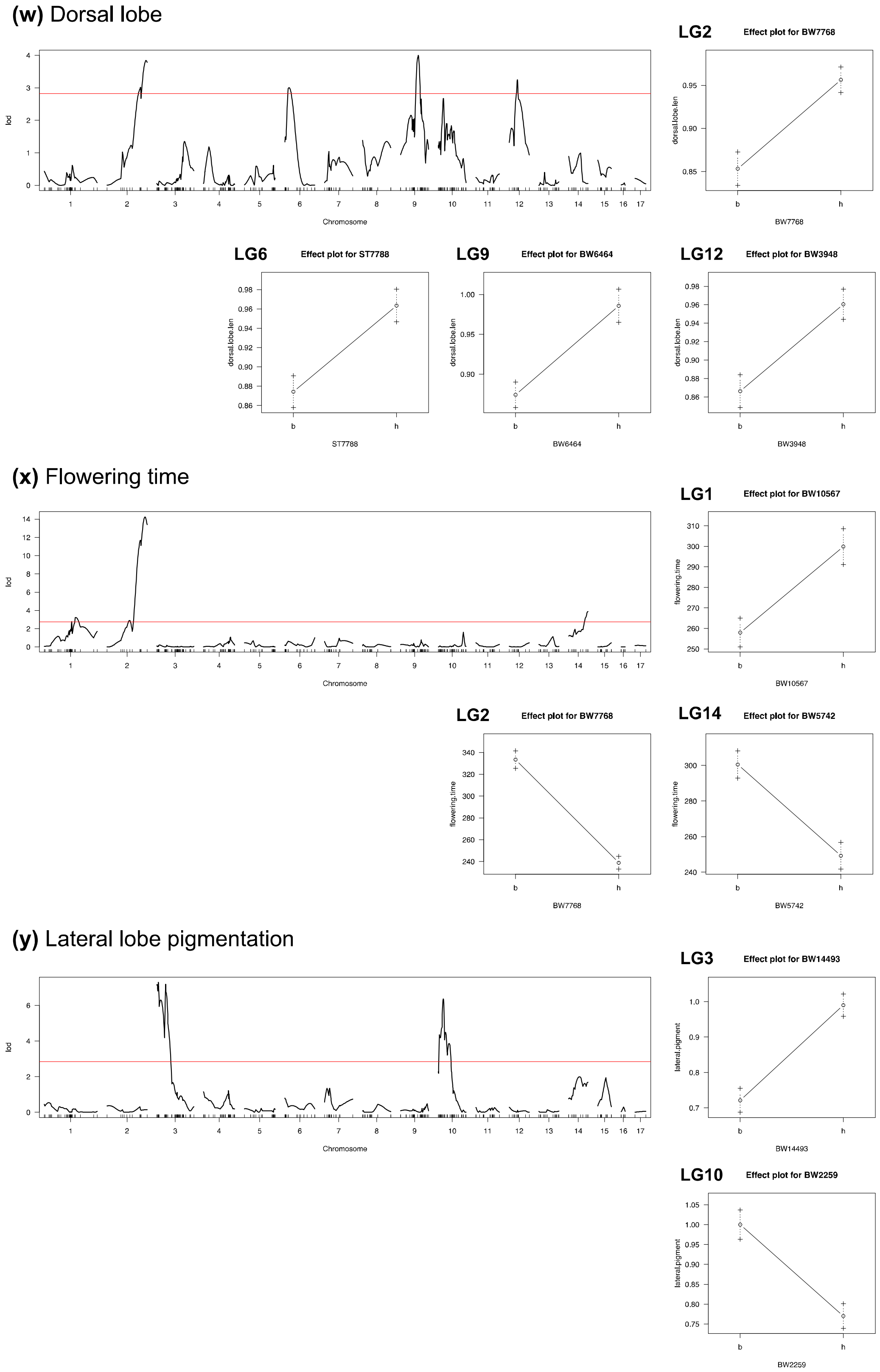
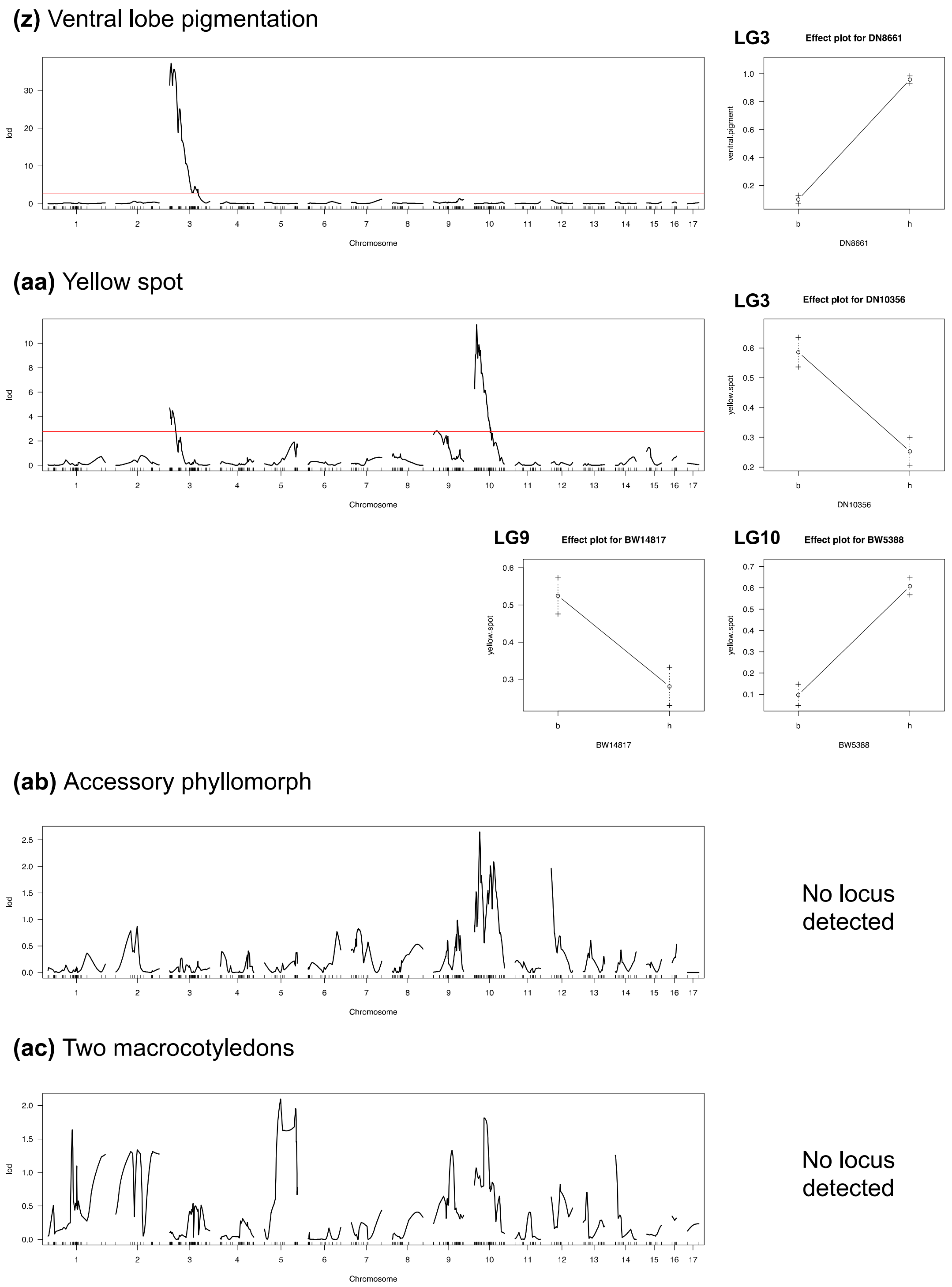


Figure 6.19 LOD curves and effect plots of the standard interval mapping on MapB-1. Full legend given on page 250.



(z) Ventral lobe pigmentation

(aa) Yellow spot

(ab) Accessory phyllomorph

(ac) Two macrocotyledons

Figure 6.19 LOD curves and effect plots of the standard interval mapping on MapB-1. Red horizontal lines indicates the LOD threshold.

Table 6.14 Standard interval mapping results of morphological traits on MapB-1

Trait	QTL marker	LG	QTL position (cM)	LOD threshold	Max LOD score	Bayes CI* (cM)	CI size (cM)	Variance explained† (%)
Corolla length	C2.loc98	2	98.00	2.76	5.44	87.22 – 102.78	15.56	12.30
Undilated tube length	BW7768	2	102.78	2.89	5.38	87.22 – 102.78	15.56	12.47
Dilated tube length	C2.loc96	2	96.00	2.72	4.09	65.12 – 102.78	37.66	9.29
Undilated tube height	BW5993	1	0.00	2.64	2.89	0.00 – 53.10	53.10	7.7
Dilated tube height	N/A	N/A	N/A	2.78	2.74	N/A	N/A	N/A
Undilated tube width	BW5993	1	0.00	2.64	3.25	0.00 – 34.36	34.36	19.36
	BW7768	2	102.78	2.64	2.85	65.12 – 102.78	37.66	
	ST11037	14	0.02	2.64	2.79	0.00 – 30.87	30.87	
Dilated tube width	C2.loc102	2	102.00	2.67	4.14	87.22 – 102.78	15.56	12.45
	BW15489	14	0.44	2.67	3.85	0.00 – 19.19	19.19	
Corolla face height	C2.loc97	2	97.00	2.66	4.67	65.12 – 102.78	37.66	17.86
	C6.loc13	6	13.00	2.66	3.00	3.70 – 37.82	34.12	
Tube opening height (outer)	BW7768	2	102.78	2.71	4.69	87.22 – 102.78	15.56	16.26
	C3.loc70	3	70.00	2.71	3.96	21.56 – 84.35	62.79	
Tube opening height (inner)	BW7768	2	102.78	2.65	3.12	87.22 – 102.78	15.56	12.59
	DN20121	3	46.50	2.65	3.74	21.56 – 84.35	62.79	
Corolla face width	C2.loc98	2	98.00	2.72	4.25	65.12 – 102.78	37.66	18.53
	C9.loc5	9	5.00	2.72	3.23	0.00 – 53.33	53.33	
Tube opening width (outer)	C2.loc97	2	97.00	2.64	3.30	58.33 – 102.78	44.45	6.50
Tube opening width (inner)	C2.loc100	2	100.00	2.66	3.22	50.70 – 102.78	52.08	5.47
Pistil length	C2.loc100	2	100.00	2.64	9.46	87.22 – 102.78	15.56	39.85
	BW416	7	16.41	2.64	2.99	6.19 – 37.55	31.36	
	BW440	9	32.36	2.64	3.60	14.46 – 63.89	49.43	
	C14.loc26	14	26.00	2.64	3.60	0.44 – 36.68	36.24	

Table 6.14 continued

Trait	QTL marker	LG	QTL position (cM)	LOD threshold	Max LOD score	Bayes CI* (cM)	CI* size (cM)	Variance explained† (%)
Ovary length	C2.loc99	2	99.00	2.74	8.15	87.22 – 102.78	15.56	26.94
	C6.loc20	6	20.00	2.74	2.84	3.70 – 37.82	34.12	
	C14.loc6	14	6.00	2.74	3.51	0.00 – 30.87	30.87	
Style length	N/A	N/A	N/A	2.72	2.11	N/A	N/A	N/A
Calyx length	C2.loc100	2	100.00	2.70	4.61	65.12 – 102.78	37.66	12.19
Stamen length	C2.loc100	2	100.00	2.77	7.21	87.22 – 102.78	15.56	25.13
	BW416	7	16.41	2.77	3.11	0.00 – 72.48	72.48	
Filament length (attached)	C2.loc101	2	101.00	2.75	6.17	87.22 – 102.78	15.56	26.76
	C7.loc17	7	17.00	2.75	3.24	6.19 – 72.48	66.29	
	BW13150	12	22.06	2.75	2.84	0.00 – 41.77	41.77	
Filament length (free)	N/A	N/A	N/A	2.75	2.67	N/A	N/A	N/A
Ventral tube length	C2.loc99	2	99.00	2.74	3.21	65.12 – 102.78	37.66	21.56
Ventral lobe length	C2.loc96	2	96.00	2.70	6.72	87.22 – 102.78	15.56	22.29
	C8.loc16	8	16.00	2.70	3.64	0.00 – 38.66	38.66	
	BW7227	9	53.31	2.70	2.83	0.25 – 62.61	62.36	
Dorsal tube length	C2.loc100	2	100.00	2.78	4.56	87.22 –102.78	15.56	17.04
	BW3948	12	21.54	2.78	2.91	0.00 – 41.77	41.77	
Dorsal lobe length	C2.loc100	2	100.00	2.71	3.84	65.12 – 102.78	37.66	30.79
	C6.loc11	6	11.00	2.71	3.00	3.70 – 30.00	26.3	
	C9.loc45	9	45.00	2.71	4.00	35.74 – 53.33	17.59	
	C12.loc21	12	21.00	2.71	3.24	0.00 – 41.77	41.77	
Flowering time	C1.loc81	1	81.00	2.76	3.21	60.62 – 136.05	75.43	50.88
	C2.loc97	2	97.00	2.76	14.24	87.22 – 102.78	15.56	
	C14.loc49	14	49.00	2.76	3.88	36.68 – 49.45	13.23	
Lateral lobe pigmentation	BW14493	3	4.17	2.78	7.29	0.00 – 28.56	28.56	29.92
	BW2259	10	12.74	2.78	6.36	9.62 – 15.77	6.15	
Ventral lobe pigmentation	C3.loc3	3	3.00	2.79	37.21	1.58 – 19.86	18.28	59.94

Table 6.14 continued

Trait	QTL marker	LG	QTL position (cM)	LOD threshold	Max LOD score	Bayes CI* (cM)	CI* size (cM)	Variance explained† (%)
Yellow spot	DN10356	3	0.00	2.66	4.70	0.00 – 19.86	19.86	45.58
	C9.loc8	9	8.00	2.66	2.82	0.00 – 34.74	34.74	
	BW5388	10	5.54	2.66	11.53	4.31 – 12.71	8.4	
Accessory phyllomorph	N/A	N/A	N/A	2.84	2.65	N/A	N/A	N/A
Two macrocotyledons	N/A	N/A	N/A	2.44	2.09	N/A	N/A	N/A

* CI: confidence interval

† Showing the combined variance explained of all the loci and inter-loci interactions

Table 6.15 Summary of the QTLs / BTLs by linkage group

Linkage group	Trait	Bayes CI (cM)
LG1	Undilated tube height	0.00 – 53.10
	Undilated tube width	0.00 – 34.36
	Flowering time	60.62 – 136.05
LG2	Corolla length	87.22 – 102.78
	Undilated tube length	87.22 – 102.78
	Dilated tube length	65.12 – 102.78
	Undilated tube width	65.12 – 102.78
	Dilated tube width	87.22 – 102.78
	Corolla face height	65.12 – 102.78
	Tube opening height (outer)	87.22 – 102.78
	Tube opening height (inner)	87.22 – 102.78
	Corolla face width	65.12 – 102.78
	Tube opening width (outer)	58.33 – 102.78
	Tube opening width (inner)	50.70 – 102.78
	Pistil length	87.22 – 102.78
	Ovary length	87.22 – 102.78
	Calyx length	65.12 – 102.78
	Stamen length	87.22 – 102.78
	Filament length (attached)	87.22 – 102.78
	Ventral tube length	65.12 – 102.78
	Ventral lobe length	87.22 – 102.78
	Dorsal tube length	87.22 – 102.78
	Dorsal lobe length	65.12 – 102.78
	Flowering time	87.22 – 102.78
	Rosulate / unifoliate (Method 4)	65.12 – 102.78
LG3	Tube opening height (outer)	21.56 – 84.35
	Tube opening height (inner)	21.56 – 84.35
	Lateral lobe pigmentation	0.00 – 28.56
	Ventral lobe pigmentation	1.58 – 19.86
	Yellow spot	0.00 – 19.86
LG4	Rosulate / unifoliate (Method 4)	28.66 – 70.23
LG5	N/A	N/A
LG6	Corolla face height	3.70 – 37.82
	Ovary length	3.70 – 37.82
	Dorsal lobe length	3.70 – 30.00
LG7	Pistil length	6.19 – 37.55
	Stamen length	0.00 – 72.48
	Filament length (attached)	6.19 – 72.48
LG8	Ventral lobe length	0.00 – 38.66
LG9	Corolla face width	0.00 – 53.33
	Pistil length	14.46 – 63.89
	Ventral lobe length	0.25 – 62.61
	Dorsal lobe length	35.74 – 53.33
	Yellow spot	0.00 – 34.74

Table 6.15 continued

Linkage group	Trait	Bayes CI (cM)
LG10	Lateral lobe pigmentation	9.62 – 15.77
	Yellow spot	4.31 – 12.71
	Rosulate / unifoliate (Method 4)	60.66 – 69.32
LG11	N/A	N/A
LG12	Filament length (attached)	0.00 – 41.77
	Dorsal tube length	0.00 – 41.77
	Dorsal lobe length	0.00 – 41.77
LG13	N/A	N/A
LG14	Undilated tube width	0.00 – 30.87
	Dilated tube width	0.00 – 19.19
	Pistil length	0.44 – 36.68
	Ovary length	0.00 – 30.87
	Flowering time	36.68 – 49.45
	Rosulate / unifoliate (Method 1)	19.19 – 49.45
	Rosulate / unifoliate (Method 2)	30.87 – 49.45
	Rosulate / unifoliate (Method 3)	10.99 – 49.45
	Rosulate / unifoliate (Method 4)	0.44 – 49.01
LG15	N/A	N/A
LG16	N/A	N/A
LG17	N/A	N/A

6.3.5 Genome annotation for the rosulate / unifoliate genetic regions using three genetic maps

In the mapping results of MapA, three genetic regions were associated with the rosulate / unifoliate trait (Table 6.16; LG1, LG7 and LG9). Among the identified regions, the confidence intervals (CI) on LG1 and LG7 mapped by scoring Method 4 and the LG9 locus mapped by scoring Method 2 were the most specific (Table 6.16). These three regions were chosen for genome annotation. 5, 7, and 23 markers fell within the confidence intervals for LG1, LG9, and LG7 respectively (Table 6.17). These corresponded to 3, 18, and 6 scaffolds from the preliminary *S. rexii* genome assembly respectively, with a total size of 1,341,715 bp (Table 6.17). The functional annotation of these scaffolds is summarised in Table 6.18.

Table 6.16 BTL regions identified for rosulate / unifoliate trait on MapA

LG	Scoring method	Bayes CI (cM)	CI size (cM)	Used for genome annotation
1	Method 2	0.00 – 148.04	148.04	
1	Method 4	126.43 – 139.33	12.90	✓
7	Method 4	61.75 – 113.64	51.89	✓
9	Method 1	77.12 – 104.94	27.82	
9	Method 2	80.56 – 101.20	20.64	✓
9	Method 3	63.54 – 91.18	27.64	
9	Method 4	21.89 – 104.94	83.05	

Table 6.17 List of genome scaffolds within the rosulate / unifoliate BTL regions identified on MapA. CI: confidence interval.

LG	Bayes CI (cM)	No. markers in CI	Corresponding genome scaffolds	Total length of scaffold (bp)
1	126.43 – 139.33	5	scaffold2920, scaffold22461, scaffold1363	164,489 bp
7	61.75 – 113.64	23	scaffold25327, scaffold14550, scaffold24553, scaffold81701, scaffold110, scaffold21844, scaffold29891, scaffold6967, scaffold3166, scaffold12762, scaffold6352, scaffold13454, scaffold7066, scaffold11456, scaffold17737, scaffold27909, scaffold11568, scaffold11935	815,681 bp
9	80.56 – 101.20	7	scaffold16151, scaffold96243, scaffold4937, scaffold28029, scaffold14085, scaffold19363	361,545 bp
Total	85.43 cM	35	27 scaffolds	1,341,715 bp

Table 6.18 Summary of functional annotation results of the genome scaffolds identified on MapA

LG	Genome scaffold	Scaffold length	Functional genes annotated*
1	scaffold2920	32,519 bp	Cytokinin riboside 5'-monophosphate phosphoribohydrolase Additional 5 hypothetical proteins
1	scaffold22461	18,084 bp	Hydroxyproline-rich glycoprotein family protein putative Auxin-responsive protein IAA8 Additional 2 hypothetical proteins
1	scaffold1363	113,886 bp	Mutant phytoene synthase RNA polymerase II transcription factor B subunit 2 Acyl-activating enzyme 11 Polyamine oxidase Microsomal glutathione S-transferase 3 Harpin inducing protein Yellow stripe-like protein 5 Protein GRIP Plant UBX domain-containing protein 11 Prolyl 4-hydroxylase 1 Peptide methionine sulfoxide reductase A5 Additional 8 hypothetical proteins
7	scaffold25327	58,538 bp	Casein kinase II subunit beta 60S ribosomal protein L37a Formyltetrahydrofolate deformylase RING-H2 finger protein ATL46 Additional 4 hypothetical proteins
7	scaffold14550	81,686 bp	Cysteine synthase Additional 5 hypothetical proteins

Table 6.18 continued

LG	Genome scaffold	Scaffold length	Functional genes annotated*
7	scaffold24553	46,317 bp	7 hypothetical proteins
7	scaffold81701	32,980 bp	UPF0176 protein Additional 3 hypothetical proteins
7	scaffold110	10,516 bp	Ethylene response factor 1
7	scaffold21844	43,263 bp	Chlororespiratory reduction 21 Calmodulin binding heat shock protein Additional 2 hypothetical proteins
7	scaffold29891	43,786 bp	Ferric reductase Additional 3 hypothetical proteins
7	scaffold6967	28,045 bp	Aspartic proteinase-like protein 2 Additional 1 hypothetical protein
7	scaffold3166	245,361 bp	Mevalonate kinase U11/U12 small nuclear ribonucleoprotein 35 kDa protein Mitochondrial Rho GTPase E3 ubiquitin-protein ligase Phosphatidylinositol 4-kinase gamma 3 Ankyrin repeat/KH domain protein (DUF1442) Terpene cyclase/mutase family member Pentatricopeptide repeat-containing protein mitochondrial Beclin 1 protein Helicase with zinc finger protein Additional 15 hypothetical proteins
7	scaffold12762	55,147 bp	Cation calcium exchanger 5-like Ubiquitin-fold modifier 1 Probable E3 ubiquitin-protein ligase LUL4 E3 ubiquitin-protein ligase ATL6 Additional 6 hypothetical proteins
7	scaffold6352	92,231 bp	DNA-directed RNA polymerase subunit Additional 5 hypothetical proteins
7	scaffold13454	9,419 bp	N/A
7	scaffold7066	68,392 bp	DNA mismatch repair protein MutS Additional 2 hypothetical proteins
7	scaffold11456	65,480 bp	Regulatory protein NPR1 Additional 6 hypothetical proteins
7	scaffold17737	44,571 bp	AR781, similar to yeast pheromone receptor Pentatricopeptide repeat protein Additional 3 hypothetical proteins
7	scaffold27909	32,519 bp	Cytokinin riboside 5'-monophosphate phosphoribohydrolase Additional 5 hypothetical proteins
7	scaffold11568	43,716 bp	Plasminogen activator inhibitor 1 RNA-binding protein Additional 2 hypothetical proteins
7	scaffold11935	17,008 bp	2 hypothetical proteins

Table 6.18 continued

LG	Genome scaffold	Scaffold length	Functional genes annotated*
9	scaffold16151	73,297 bp	Hydroxyproline O-galactosyltransferase HPGT2 Disease resistance protein-like Reticulon-like protein B8 Suppressor of G2 allele of SKP1 Putative septum site-determining protein MinD Additional 7 hypothetical proteins
9	scaffold96243	1,395 bp	N/A
9	scaffold4937	35,130 bp	DExH-box ATP-dependent RNA helicase DExH10 Auxin response factor
9	scaffold28029	118,704 bp	Transmembrane 9 superfamily member Metal-dependent phosphohydrolase Receptor-like protein kinase HERK 1 Additional 12 hypothetical proteins
9	scaffold14085	98,260 bp	Cytokinin oxidase 3 Eukaryotic translation initiation factor 3 subunit E Mitochondrial glycoprotein Zeatin O-glucosyltransferase-like Kinesin KP1 Insulin-degrading enzyme-like 1 peroxisomal Additional 5 hypothetical proteins
9	scaffold19363	34,759 bp	Receptor-like protein kinase HSL1 Calmodulin-binding receptor-like cytoplasmic kinase 3 Proline-rich receptor-like protein kinase PERK1 Protein SPEAR3 Additional 3 hypothetical proteins

* Hypothetical genes are uncharacterised genes and with unknown functions

The same mapping procedure on MapB-1 and MapB-3 showed that for MapB-1, four genetic regions were identified on LG2, LG4, LG10 and LG14 for the rosulate / unifoliate trait (Table 6.19). These regions were used for marker check and genome annotation, and the linkage groups corresponded to 4, 14, 3, 4 scaffolds respectively, with a total of 1,840,643 bp of sequence (Table 6.20). On the other hand, four genetic regions were identified on MapB-3, LG1, LG2, LG4, and LG14 (Table 6.21). In particular, the 6.93 cM region on LG1 detected by CIM spanned 39 markers and 12 genome scaffolds (Table 6.22). For the other three regions, nine of their corresponding genome scaffolds had been found in the results of MapB-1 already (Table 6.22; scaffolds marked with an asterisk were identified in MapB-1). Finally, the scaffolds identified in both MapB-1 and MapB-3 were used for functional annotation (Table 6.23). The relationship between all the retrieved scaffolds are summarised in Table 6.24.

Table 6.19 BTL regions identified for rosulate / unifoliate trait on MapB-1

LG	Scoring method	Bayes CI (cM)	CI size (cM)	Used for genome annotation
2	Method 4	65.12 – 102.78	37.66	✓
4	Method 4	28.66 – 70.23	41.57	✓
10	Method 4	60.66 – 69.32	8.65	✓
14	Method 1	19.19 – 49.45	30.25	
14	Method 2	30.87 – 49.45	18.58	✓
14	Method 3	10.99 – 49.45	38.46	
14	Method 4	0.44 – 49.01	49.01	

Table 6.20 List of genome scaffolds fall within the rosulate / unifoliate BTL regions identified on MapB-1

LG	Bayes CI (cM)	No. markers in CI	Corresponding genome scaffolds	Total length of scaffold (bp)
2	65.12 – 102.78	7	4621390, 4602510, 4628119, 4619676	437,759 bp
4	28.66 – 70.23	23	4621038, 4628495, 4606711, 4596587, 4602628, 4626269, 4626131, 4625029, 4598249, 4303143, 4626813, 4628153, 4624979, 4626266	750,999 bp
10	60.66 – 69.32	4	4628071, 4621448, 4624923	353,203 bp
14	30.87 – 49.45	6	4621635, 4583468, 4628222, 4609125	298,682 bp
Total	106.46 cM	40	25 scaffolds	1,840,643 bp

Table 6.21 BTLs identified for rosulate / unifoliate trait on MapB-3

LG	Scoring method	Bayes CI (cM)	CI size (cM)	Used for genome annotation
1	Method 1	64.64 – 71.57	6.93	✓
2	Method 2	81.82 – 97.88	16.06	✓
2	Method 4	67.47 – 97.88	30.41	
4	Method 4	34.54 – 53.00	18.45	✓
14	Method 1	57.70 – 78.20	20.49	
14	Method 2	59.99 – 78.20	18.20	✓
14	Method 3	39.45 – 78.20	38.74	
14	Method 4	25.11 – 78.20	53.09	

Table 6.22 Candidate genome scaffolds identified in MapB-3

LG	Bayes CI (cM)	No. markers in CI	Corresponding genome scaffolds	Total length of scaffold (bp)
1	64.64 – 71.57	39	4628601, 4618086, 4629712, 4605369, 4621557, 4619330, 4628439, 4605598, 4627533, 4620874, 4592339, 4626926, 4586234, 4603630, 4621018, 4605598, 4628027, 4622699, 4618924, 4596325, 4628211, 4625786, 4621691, 4618583	1,666,522 bp
2	81.82 – 97.88	5	4628119*, 4619231, 4582452, 4619676*	340,454 bp
4	34.54 – 53.00	7	4621038*, 4628495*, 4627304, 4606711*, 4596587*	361,184 bp
14	59.99 – 78.20	9	4583468*, 4620652, 4628222*, 4609125*, 4602532	399,859 bp
Total	59.64 cM	60	38 scaffolds	2,768,019 bp

* Repeated genome scaffolds that were also identified in the result of MapB-1 (Table 6.S)

Table 6.23 Summary of functional annotation results of the genome scaffolds identified on MapB-1 and MapB-3

LG	Genome scaffold	Scaffold length	Functional genes annotated*
1	4628601	129,642 bp	COMPASS-like H3K4 histone methylase component WDR5A Chaperone protein ClpB1 Additional 13 hypothetical proteins
1	4618086	62,244 bp	Reverse transcriptase-related family protein DUF4228 domain protein Additional 6 hypothetical proteins
1	4629712	140,813 bp	Non-specific lipid transfer protein GPI-anchored 2-like isoform Mitochondrial transcription termination factor family protein Additional 19 hypothetical proteins
1	4605369	52,719 bp	2-oxoglutarate (2OG) and Fe(II)-dependent oxygenase Cholinephosphate cytidylyltransferase Additional 2 hypothetical proteins
1	4621557	32,784 bp	Integral membrane protein like Additional 5 hypothetical proteins
1	4619330	86,919 bp	MA3 domain-containing protein Tetratricopeptide repeat (TPR)-containing protein Additional 11 hypothetical proteins
1	4628439	121,713 bp	Adenosine 5'-phosphosulfate reductase 8 Mitochondrial dicarboxylate/tricarboxylate transporter DTC ARM repeat superfamily protein Additional 12 hypothetical proteins
1	4605598	68,158 bp	Transcription factor Pur-alpha 1 Pentatricopeptide repeat-containing protein At5g55840 Additional 7 hypothetical proteins
1	4627533	54,234 bp	DNA replication complex GINS protein SLD5 Additional 11 hypothetical proteins
1	4620874	128,869 bp	Leucine-rich repeat (LRR) family protein Prolyl carboxypeptidase like protein Kinesin-like protein NACK1 WAT1-related protein At5g64700 Additional 11 hypothetical proteins
1	4592339	15,962 bp	1 hypothetical protein
1	4626926	56,617 bp	Mevalonate kinase U11/U12 small nuclear ribonucleoprotein 35 kDa protein Protein BRASSINOSTEROID INSENSITIVE 1 Additional 8 hypothetical proteins
1	4586234	44,456 bp	tRNA pseudouridine synthase A
1	4603630	18,303 bp	2 hypothetical proteins
1	4621018	69,478 bp	Putative E3 ubiquitin-protein ligase HERC1 Additional 6 hypothetical proteins
1	4605598	68,158 bp	Transcription factor Pur-alpha 1 Pentatricopeptide repeat-containing protein At5g55840 Additional 7 hypothetical proteins

Table 6.23 continued

LG	Genome scaffold	Scaffold length	Functional genes annotated*
1	4628027	189,799 bp	Ankyrin repeat domain-containing protein EMB506 Protein NRT1/ PTR FAMILY 8.1-like Alpha-mannosidase At3g26720 Type IV inositol polyphosphate 5-phosphatase 9 Probable mitochondrial adenine nucleotide transporter BTL3 50S ribosomal protein L29 LRR receptor-like kinase BZIP transcription factor, putative (DUF630 and DUF632) Tyrosine transaminase Additional 15 hypothetical proteins
1	4622699	57,786 bp	TCP20 protein Peptidylprolyl isomerase Additional 8 hypothetical proteins
1	4618924	37,328 bp	Polyprotein Heavy metal-associated isoprenylated protein 1 Protease Do-like 9 Additional 3 hypothetical proteins
1	4596325	27,567 bp	4 hypothetical proteins
1	4628211	84,038 bp	Subtilisin-like serine protease Bifunctional inhibitor/lipid-transfer protein/seed storage 2S-albumin superfamily protein AT-hook motif nuclear-localized protein 15 Additional 14 hypothetical proteins
1	4625786	53,980 bp	Polyprotein 40S ribosomal protein SA Microtubule associated protein Additional 9 hypothetical proteins
1	4621691	24,664 bp	7 hypothetical proteins
1	4618583	40,291 bp	BTB/POZ domain-containing protein At1g03010 isoform X1 Additional 4 hypothetical proteins
2	4621390	58,262 bp	Zinc finger BED domain-containing protein RICESLEEPER 1 Alkaline alpha-galactosidase Additional 9 hypothetical proteins

Table 6.23 continued

LG	Genome scaffold	Scaffold length	Functional genes annotated*
2	4602510	165,194 bp	C2 calcium/lipid-binding plant phosphoribosyltransferase family R2R3 MYB Lim domain protein Bifunctional protein FolD 4, chloroplastic Digalactosyldiacylglycerol synthase 2, chloroplastic RRNA adenine N(6)-methyltransferase SET domain-containing protein LysM domain containing protein Peroxisome biogenesis protein 16 UTP:RNA uridylyltransferase 1 Probable carboxylesterase 9 Protein NRT1/ PTR FAMILY 8.1-like Pentatricopeptide repeat (PPR) superfamily protein Cysteine-rich repeat secretory protein 15 Protein YABBY 5 Scarecrow-like protein 6 isoform X1 Phosphatidate cytidylyltransferase SPX domain-containing protein 3 BHLH transcription factor Additional 8 hypothetical proteins
2	4619676	52,194 bp	Putative serine/threonine-protein kinase Rad53 ABC transporter D family member 1 Additional 7 hypothetical proteins
2	4619231	107,848 bp	Zinc finger (C3HC4-type RING finger) family protein Protein RMD5 homolog Phosphoenolpyruvate carboxykinase (ATP) EPIDERMAL PATTERNING FACTOR-like protein 2 Thioredoxin superfamily protein ACT domain-containing protein ACR4 NADH-ubiquinone oxidoreductase 18 kDa subunit Telomere repeat-binding factor 1 Remorin family protein Mitogen-activated protein kinase kinase kinase npk1 Nicotinate phosphoribosyltransferase 2 Transcription factor APETALA2 Pollen receptor-like kinase 3 Transmembrane protein, putative (DUF247) Additional 8 hypothetical proteins
2	4582452	18,303 bp	Auxin-responsive protein IAA8 Hydroxyproline-rich glycoprotein family protein, putative Enolase (DUF1399) Additional 2 hypothetical proteins

Table 6.23 continued

LG	Genome scaffold	Scaffold length	Functional genes annotated*
4	4621038	112,522 bp	UDP-glycosyltransferase 79B30 Ribosomal protein L11 methyltransferase BnMAP4K alpha1 Serine/threonine-protein kinase WNK-like protein Protein DETOXIFICATION 16 Additional 8 hypothetical proteins
4	4628495	81,446 bp	Ankyrin repeat protein SKIP35 Cysteine synthase COMPASS-like H3K4 histone methylase component WDR5A BAG-associated GRAM protein 1 Additional 5 hypothetical proteins
4	4606711	43,402 bp	E3 ubiquitin-protein ligase UPL6 Inositol transporter 4 Additional 9 hypothetical proteins
4	4596587	33,990 bp	UPF0176 protein UDP-glucose 4-epimerase GEPI48 SNARE-interacting protein KEULE Additional 2 hypothetical proteins
4	4602628	10,463 bp	1 hypothetical protein
4	4626269	97,206 bp	Mechanosensitive ion channel protein 1, mitochondrial UPF0496 protein 4 Probable serine/threonine-protein kinase At1g01540 Ferric reductase Protein phosphatase 2C 16 Calmodulin binding protein Additional 8 hypothetical proteins
4	4626131	67,802 bp	Cation/calcium exchanger 4 Signal peptidase complex subunit 3B Ubiquitin-fold modifier 1 Probable E3 ubiquitin-protein ligase LUL4 E3 ubiquitin-protein ligase ATL6 Additional 9 hypothetical proteins
4	4625029	75,423 bp	Aspartic peptidase A1 family, Aspartic peptidase domain protein 4-coumarate:coenzyme A ligase Polyketide cyclase/dehydrase and lipid transport superfamily Truncated xanthoxin dehydrogenase Additional 4 hypothetical proteins
4	4598249	39,938 bp	DNA mismatch repair protein MSH2 Additional 2 hypothetical proteins
4	4303143	31,880 bp	Cytokinin riboside 5'-monophosphate phosphoribohydrolase Additional 5 hypothetical proteins
4	4626813	17,364 bp	1 hypothetical protein
4	4628153	20,308 bp	2 hypothetical proteins

Table 6.23 continued

LG	Genome scaffold	Scaffold length	Functional genes annotated*
4	4624979	38,803 bp	Pentatricopeptide repeat protein AR781, similar to yeast pheromone receptor Serine carboxypeptidase II-3 Additional 3 hypothetical proteins
4	4626266	80,452 bp	Proline-, glutamic acid/leucine-rich protein IMP dehydrogenase Glucuronoxylan methyltransferase Additional 7 hypothetical proteins
4	4627304	89,824 bp	Haloacid dehalogenase-like hydrolase domain-containing protein Dynamin-related protein 1E Malate dehydrogenase (oxaloacetate-decarboxylating) Chloroplast chaperonin 10 E3 ubiquitin-protein ligase ATL4 Formyltetrahydrofolate deformylase Eukaryotic translation initiation factor 4G-like isoform X1 60S ribosomal protein L37a Histone H3 K4-specific methyltransferase SET7/9 family protein Casein kinase II subunit beta Additional 5 hypothetical proteins
10	4628071	95,141 bp	Glycine hydroxymethyltransferase Adenosylhomocysteinase Rapid alkalinisation factor 1 Selenium binding protein Pentatricopeptide repeat-containing protein Putative SPINDLY protein MazG nucleotide pyrophosphohydrolase domain protein Transcription factor RF2a Electron transfer flavoprotein-ubiquinone oxidoreductase Strictosidine synthase Protein ENHANCED DISEASE RESISTANCE 2 Additional 11 hypothetical proteins

Table 6.23 continued

LG	Genome scaffold	Scaffold length	Functional genes annotated*
10	4621448	188,400 bp	Mannose-1-phosphate guanylyltransferase Calmodulin binding protein Succinate dehydrogenase subunit 7A, mitochondrial Nucleobase-ascorbate transporter 6 Pollen receptor-like kinase 3 NHL domain protein Pyruvate decarboxylase 2 Deleted in split hand/splt foot protein 1 Bidirectional sugar transporter SWEET Serine/threonine-protein kinase HT1 S locus glycoprotein E3 ubiquitin-protein ligase SIS3 DUF1645 family protein Serine/threonine-protein kinase ATM isoform X1 Histone deacetylase Chloride channel protein Pollen receptor-like kinase 3 Plastid division protein CDP1, chloroplastic Additional 9 hypothetical proteins
10	4624923	69,662 bp	Protein LIGHT-DEPENDENT SHORT HYPOCOTYLS 3 Malate dehydrogenase Cryptochrome 2 Inorganic phosphate transporter Ethylene-responsive transcription factor 1B RING-H2 finger protein ATL16 AT-hook motif nuclear-localized protein 28 Ethylene receptor Additional 6 hypothetical proteins
14	4621635	80,310 bp	MYB transcription factor 77 Outer membrane OMP85 family protein Probable inactive receptor-like protein kinase At3g56050 Phosphoribosylanthranilate isomerase CTD small phosphatase-like protein Serine/threonine-protein kinase-like protein CCR1 Additional 5 hypothetical proteins
14	4583468	26,938 bp	Keratin-associated protein, putative (DUF819) Putative septum site-determining protein MinD Suppressor of G2 allele of SKP1 Reticulon-like protein B8 Additional 3 hypothetical proteins

Table 6.23 continued

LG	Genome scaffold	Scaffold length	Functional genes annotated*
14	4628222	163,556 bp	Protein FLC EXPRESSOR Receptor-like protein kinase HERK 1 Metal-dependent phosphohydrolase Putative clathrin assembly protein Transmembrane 9 superfamily member Histone-lysine N-methyltransferase ASHR3 ATP-dependent DNA helicase Eukaryotic translation initiation factor isoform 4G-1 Protein MULTIPOLAR SPINDLE 1 Signal recognition particle receptor subunit alpha-like protein Transmembrane 9 superfamily member Eukaryotic translation initiation factor 5A Leucine-rich repeat receptor protein kinase EMS1 Photosynthetic NDH subunit of lumenal location 5, chloroplastic Glycerol 3-phosphate permease Additional 19 hypothetical proteins
14	4609125	27,878 bp	Auxin response factor Additional 2 hypothetical proteins
14	4620652	90,616 bp	Lipase class 3-like Polypyrimidine tract-binding protein 1 Mitochondrial carrier protein MTM1 Inorganic phosphate transporter Hydroxyproline O-galactosyltransferase HPGT2 Disease resistance protein-like Additional 8 hypothetical proteins
14	4602532	90,871 bp	Cytokinin dehydrogenase Eukaryotic translation initiation factor 3 subunit E Mitochondrial glycoprotein Zeatin O-glucosyltransferase-like Kinesin KP1 Additional 6 hypothetical proteins

* Hypothetical genes are uncharacterised genes and with unknown functions

Table 6.24 Genome-to-genome alignment between identified scaffolds and *S. grandis* genome

Scaffolds identified in MapB-1 and MapB-3	Corresponding scaffolds identified in MapA	Corresponding scaffolds in *S. grandis* genome assembly
4628601	N/A	4294169
4618086	N/A	4315545
4629712	N/A	4321228
4605369	N/A	4283118
4621557	N/A	4356665
4619330	N/A	4350554
4628439	N/A	4284933
4605598	N/A	4345973
4627533	N/A	4355442
4620874	N/A	4277864
4592339	N/A	4352733
4626926	N/A	4353756
4586234	N/A	4278727
4603630	N/A	4350841
4621018	N/A	4294905
4605598	N/A	4345973
4628027	N/A	4349035
4622699	N/A	4352460
4618924	N/A	4343558
4596325	N/A	4361608
4628211	N/A	4289226
4625786	N/A	4348323
4621691	N/A	4354299
4618583	N/A	4310861
4621390	N/A	4350128
4602510	N/A	4356806
4628119	N/A	4355104
4619676	N/A	4348241
4619231	Scaffold2920	4349104
4582452	Scaffold22461	4349104
4621038	N/A	4353991
4628495	Scaffold14550	4349065
4606711	Scaffold24553	4319817
4596587	Scaffold81701	4358164
4602628	Scaffold110	4284635, 4287149
4626269	Scaffold29891, Scaffold11568	4348201
4626131	Scaffold16762	4303565
4625029	Scaffold6967	4349078
4598249	Scaffold7066	4355834
4303143	Scaffold27909	4357159
4626813	Scaffold13454	4361635
4628153	Scaffold11935, Scaffold11456	4323079
4624979	Scaffold17713	4356839
4626266	Scaffold6352	4287335
4627304	Scaffold25327	4319817
4628071	N/A	4320697

Table 6.24 continued

Scaffolds identified in MapB-1 and MapB-3	Corresponding scaffolds identified in MapA	Corresponding scaffolds in *S. grandis* genome assembly
4621448	N/A	4359247
4624923	N/A	4352617
4621635	N/A	4318672
4583468	Scaffold16151	4287149
4628222	Scaffold28029, Scaffold3166, Scaffold1363	4359040
4609125	Scaffold4937	4352832
4620652	N/A	4352732
4602532	Scaffold21844, Scaffold14085	4345871
N/A	Scaffold19363	4352075
N/A	Scaffold96243	4352667

6.4 Discussion

6.4.1 Genetic architecture of the rosulate and unifoliate growth form

Genetic mapping of the rosulate / unifoliate loci was carried out using four different scoring methods on three genetic maps. The result identified up to five genetic loci, on LG1, LG2, LG4, LG10, LG14 of MapB-1 and MapB-3, and on LG1, LG7, LG9 of MapA (corresponds to MapB LG2, LG4 and LG14, respectively). In particular, the loci on MapB LG14 and LG2 were consistently detected in most of the mapping results. These identified loci contributed small to medium proportion of variance (7.72% - 38.66%), indicating that a substantial amount of phenotype was not explained. Overall, these results suggest that the determination of rosulate and unifoliate trait might be controlled by several genes. However, several key issues remained to be addressed, including the ambiguous growth forms observed, the skewed non-Mendelian segregation ratio, the multiple loci found and differences between QTL mapping results of various scoring methods, and the verification of the early and late loci hypothesis. For ease of following the discussion, the linkage group numbering below follows the MapB-1 system.

Since there were some ambiguous growth forms in our BC mapping population, the distinction between rosulate and unifoliate phenotype was not always clear (Figure 6.4). Some of the observed phenotypes did not represent *S. rexii* or *S. grandis*, thus it was difficult to distinguish between rosulate and unifoliate categories. Similar complications were encountered by Harrison (2002) who described ‘a spectrum‘ of rosulate / unifoliate phenotypes in the crosses between *S. rexii* and *S. dunnii* (unifoliate) and *S. rexii* and *S. wittei* (unifoliate). It is common for inter- and intraspecific hybrids to exhibit novel phenotypes that differ from both parental lineages (Rieseberg et al., 1999). They could be a result of additive allele effects or epistasis interactions of the novel allele combinations obtained through hybridisation, which leads to unusual phenotypes than observed in the parents (Dittrich-Reed and Fitzpatrick, 2012). It is possible that the diverse growth forms observed in the BC population were novel phenotypes originating from the combination of *S. rexii* and *S. grandis* genetic backgrounds.

The diverse growth forms complicated the scoring process, and either a Mendelian 3:1 ratio or non-Mendelian ratio was obtained depending on the scoring scheme chosen (Table 6.9). Deviations from the expected 3:1 ratio for a trait inherited by two dominant loci has previously being reported by Harrison (2002), who recorded a rosulate:unifoliate ratio of 7.92:1 in the BC population of (*S. rexii* × *S. wittei*) × *S. wittei* (N = 116). The segregation ratio may be greatly affected by the decision whether or not to score the presence of accessory phyllomorph as rosulate (Harrison, 2002). *S. grandis* is capable of producing accessory phyllomorphs (Jong, 1970; Nishii et al., 2012a), and it is possible for the BC individuals to inherit this trait from *S. grandis* and falsely be scored as rosulate. In this study, four different scoring methods were employed that differed in the placement of plants with accessory phyllomorphs and other ambiguous phenotypes. When scoring accessory phyllomorphs as rosulate (i.e. Method 2, and Method 4 if the accessory phyllomorph is originated from the groove meristem), the expected 3:1 ratio (for 2 dominant loci) of rosulate:unifoliate was obtained; when scoring accessory phyllomorphs as unknown or unifoliate (i.e. Method 1 and Method 3) the ratio of rosulate to unifoliate = 1:1 (for 1 dominant locus) and 5:1 ratio (non-Mendelian) was obtained, respectively (Table 6.9). Thus, the decision on whether the accessory phyllomorph is scored as rosulate or not affected the resulting segregation ratios.

However, it remained unknown whether the accessory phyllomorphs found in the BC plants represented the 'late developing rosulates' described by Oehlkers (1938; 1942). Oehlkers (1938; 1942) recorded a rosulate to unifoliate ratio of 3:1 at the flowering season of the BC plants. This is partly overlapped with the time the accessory phyllomorphs emerged, i.e. after flowering (Jong, 1978; Nishii et al., 2012a). The development of the accessory phyllomorph in *S. grandis* was first properly documented by Jong (1970) who named it 'subtending phyllomorph' in his PhD thesis. Later, Nishii et al. (2012) published this distinctive development for several unifoliates. It is possible that Oehlkers (1938; 1942) overlooked the capability of *S. grandis* to produce additional phyllomorphs, and subsequently scored all BC plants with accessory phyllomorphs as rosulate for their similar appearance to rosulates in his study. This hypothesis can be partly supported by our Method 2 scoring result, where plants with accessory phyllomorphs were scored as rosulate and a 3:1 Mendelian segregation ratio was obtained, suggesting 2 dominant loci. The fact that scoring Method 1, where only truly rosulates were scored as rosulates and the remaining as unifoliates, resulted in a 1:1 segregation ratio, indicative of a single dominant locus (for true and early rosulateness), further suggests that it is possible to separate the two loci and that the accessory phyllomorphs may represent the late acting rosulate locus.

Different rosulate / unifoliate loci were mapped among the 12 mapping attempts (i.e. four scoring methods × three genetic maps). In total, five genetic loci were found associated to the growth form variation and the two loci on LG14 and LG2 were consistently detected. The LG14 locus was detected in all 12 mapping results; the LG2 locus was found in the mapping using Method 2 and Method 4 scoring methods (Table 6.11, Table 6.12 and Table 6.13). This result suggests that the two loci may have major effects on determining the rosulate phenotype and are thus frequently detected. In particular, the locus on LG14 is probably associated with the production of true phyllomorphs (type 1 phyllomorphs in Figure 6.4 a), as this locus was detected in both the strictest scoring (Method 1; only counts true rosulates) as well as the less-strict scorings (Method 2 and Method 4; counting true rosulates, accessory phyllomorphs, and other additional phyllomorphs associated with flowering). On the other hand, the locus on LG2 is possibly associated with accessory phyllomorphs (type 2 phyllomorphs in Figure 6.4 b), as it was only detected when using the less-strict scoring method (Method 2 and Method 4) but not the strict scoring Method 1.

Mapping using Method 4 scoring always detected 1 to 2 loci in addition to the loci on LG2 and LG 14 described above, and it detected the highest number of loci compared to other scoring methods (Table 6.11, Table 6.12 and Table 6.13). One possible explanation is that the multiple loci detected using Method 4 scoring are additional genes related to the novel phenotypes of phyllomorph formation. For example, as Method 4 scoring considered any phyllomorph related to the groove meristem as rosulate, it is possible that it detected additionally linked phenotypes, such as modifier loci involving in regulating length of petiolode, a common feature of accessory phyllomorphs (Figure 6.4 b and c). On the other hand, mapping using Method 3 scoring always gave the wider confidence intervals, and was only able to detect the LG14 locus. This is probably due to the high proportion of 'unknown' phenotype scored in Method 3 (52 unknowns) that were excluded from analysis and thus a major part of the genetic information was therefore discarded. It is known that scoring errors and missing data in phenotyping can reduce the detection of genetic associations (Edwards et al., 2005; Broman and Sen, 2009). In addition, one locus on LG1 was uniquely mapped in the CIM result on MapB-3 but was not detected in other mapping analyses (Figure 6.17 a). The validity of this locus may require further study as it was not consistently found in other mapping scenarios, and maybe therefore an artefact. Also, MapB-3 was built using the least stringent marker-filtering strategy, that can increase the chances of incorrect marker order and distance information of the map (Hackett and Broadfoot, 2003).

In his original work Oehlkers (1938, 1942) speculated that the rosulate / unifoliate trait is regulated by an early and a late acting dominant genetic locus, where the rosulate to unifoliate ratio of a BC population is expected to be 1:1 at six months after sowing, and 3:1 at around nine months after sowing. There may be great variation in the timing at which point the rosulate trait appears, and it may take a longer time to observe the 3:1 ratio in a backcross population, though exactly how long was not specified in the original study (Oehlkers, 1942). However, our mapping results suggested that more than two loci may be associated with the rosulate / unifoliate trait. Thus, it is possible that the number of phyllomorphs is regulated by a more complicated genetic mechanism than previously hypothesised. However, there may be two particular loci that are associated with time of development. As previously discussed, the LG14 locus could be associated with the regulation of type 1 (Figure 6.4 a) phyllomorph development, which is usually observed prior to plant flowering and may represent phyllomorph production at earlier developmental stages. The LG2 locus are possibly associated with type 2 (Figure 6.4 b) phyllomorph development, which is always produced together with the inflorescence and may represent the phyllomorphs produced at later developmental stages.

Strangely, the effect plots of almost all the detected loci suggest that the homozygous genotype (b; carrying only *S. grandis* alleles) was associated with a rosulate phenotype while the heterozygous genotype (h; carrying one *S. rexii* allele and one *S. grandis* allele) was associated with a unifoliate phenotype (Appendix 6.8). The exceptions were the LG10 locus identified in MapB-1 Method 4 mapping (Appendix 6.8 b), and the LG1 locus identified in MapB-3 Method 1 mapping (Appendix 6.8 c). At the same time, most of the effect plots suggested that the difference of the average phenotypic values between the two genotypes were quite small, usually between 0.3 – 0.4 (Appendix 6.8). One observation was that the genetic regions identified for LG4, LG10 and LG14 were all quite poorly resolved and had relatively lower marker density (Figure 6.18). It is possible that the available markers in these regions were still quite distant from the actual causative locus. This may have affected the calculation of the conditional genotype probabilities (i.e. the calc.genoprob function in r/qtl), that the genotype of the causative loci may have been incorrectly estimated (Broman and Sen, 2009). As can be observed in the LG10 and LG1 loci, the two BTL regions identified had a higher marker density and a narrower confidence interval was obtained. Thus, increasing the marker density of the LG2, LG4 and LG14 BTL regions and reanalysis of the effect plots may give a more accurate picture of the effect of the potentially causative loci.

6.4.2 Candidate genes identified for the rosulate and unifoliate growth form

Several plant developmental and hormone-related genes were identified in the annotated genome scaffolds and the results are summarised by linkage groups (Table 6.25). In terms of developmental proteins and genes, the COMPASS-like H3K4 histone methylase component WDR5A protein (LG1 and LG4) regulates the methylation of histone and is related to the suppression of *FLOWERING LOCUS T* gene, where the knockout mutant causes accelerated floral transition in *A. thaliana* (Jiang et al., 2011). The ankyrin repeat domain-containing *EMB506* gene (LG1) is essential for embryogenesis and vegetative development, especially for the transition of radial symmetry to bilateral symmetry at the early heart stage (Despres et al., 2001). Mosaic *emb506 Arabidopsis* plants exhibit defect leaf morphologies, including elimination of one cotyledon, altered shaped cotyledon, addition of cotyledon number (possibly related to a complete bifurcation of one of the cotyledons), and similar phenotypes can be observed in the true leaves (Latvala-Kilby and Kilby, 2006). The SPEAR3 protein (LG14), or TIE1, are associated with transcription factors TOPLESS protein and the mutation results in abnormal leaf growth in *Arabidopsis* (Tao et al., 2013).

Some of the developmental proteins and genes are related to the genes previously studied in *Streptocarpus*. For instance, SET domain-containing proteins (LG2) regulate gene expression through histone modification (Ng et al., 2007), and some of the gene members such as the *CURLY LEAF* (*CLF*) functions to repress meristem identity genes including *AGAMOUS* (*AG*) and *SHOOTMERISTEMLESS* (*STM*) (Goodrich et al., 1997; Ng et al., 2007). In *Streptocarpus*, the orthologs of *STM* are expressed in groove and basal meristems and are related to meristem activity (Harrison et al., 2005; Mantegazza et al., 2009). YABBY proteins (LG2) regulate the development of adaxial and abaxial polarity and the expression can be detected from embryo stage (Siegfried et al., 1999; Stahle et al., 2009). In *Streptocarpus*, the orthologous gene *SrGRAMILIFOLIA* (*SrGRAM*) is expressed in the basal meristem (Tononi et al., 2010). Finally, the *LIGHT-DEPENDENT SHORT HYPOCOTYL* genes (LG10; identical to *ORGAN BOUNDARY* (*OBO*) genes) are expressed at the junction between shoot apical meristem and lateral organs. It may act as the transcription factor for several meristem related genes such as *CUP-SHAPED COTYLEDON* (*CUC*), *LATERAL ORGAN BOUNDARIES* (*LOB*), or *ASYMMETRIC LEAVES* (*AS*) (Cho and Zambryski, 2011). Overexpression of *OBO1* gene leads to disrupt phyllotaxy and multiple shoot apex (Cho and Zambryski, 2011). Interestingly, one member of the gene family *LSH1* is involved in light sensing which the knockout mutation exhibits hypersensitive to red, far-red and blue lights, resulted in shorter hypocotyl (Zhao et al., 2004). In *S. rexii* light was also found to be an important factor for early seedling and anisocotylous development (Nishii et al., 2012b). Seedling grown under blue light condition shows normal anisocotyly development, while seedling grown under red light show no basal meristem activity and remained with two microcotyledons, with a small proportion (32%) of plants showing leaf with elongated petiole emerged between the two cotyledons (Nishii et al., 2012b).

In *Streptocarpus*, gibberellin and cytokinin are related to the establishment of anisocotyly and production of additional phyllomorphs (Rosenblum and Basile, 1984; Mantegazza et al., 2009; Nishii et al., 2012a; 2014; Chen et al., 2017). Gibberellin and cytokinin metabolic proteins and genes were also found in the annotated genome scaffolds. For gibberellin, the scarecrow-like proteins (LG1) promotes gibberellin signalling by counteract the signalling repressor DELLA protein (Zhang et al., 2011), while the SPINDLY protein negatively regulates the signalling pathway (Silverstone et al., 2006). The Scarecrow-like protein 6 and its homologs, also known as the *LOST MERISTEM* genes, are crucial for meristem maintenance in both *Arabidopsis* and *Petunia* (Stuurman et al., 2002; Engstrom et al., 2011). For cytokinin, the *LONELY GUY* (*LOG*; LG2 and LG4) gene encodes a cytokinin activating enzyme for the biosynthesis of biologically active cytokinin molecules, and is directly involved in shoot apical meristem maintenance (Kuroha et al., 2009). On the other hand, cytokinin oxidase (CKX; LG14) degrades biologically active cytokinin and overexpression leads to reduced shoot apical meristem size and reduced leaf number in *Arabidopsis* (Schmülling et al., 2003). The zeatin-O-glucosyltransferase protein (ZOG; LG14) does not degrade cytokinin but instead converts it into a non-active form for storage, which can be reactivated by other enzymes (Martin et al., 2001). The protease Do-like 9 (LG1) is a ATP-independent serine protease involved in cytokinin and light-signalling pathway through degrading the ARABIDOPSIS RESPONSE REGULATOR 4 (ARR4) protein. It is also involved in seedling development that the mutation *deg9* shows the phenotype of elongated hypocotyl under red light (Chi et al., 2016).

The plant hormone auxin are involved in shoot apical dominance and lateral organ differentiation (Azizi et al., 2015). The identified genes *AUXIN/INDOLE-3-ACETIC ACID* (*IAA*; LG2) are involved in auxin signalling pathway by encoding for short-lived transcriptional repressor, which is inhibit in the presence of auxin (Overvoorde et al., 2005). The Auxin Response Factor proteins (ARF; LG14) are transcription factors that target auxin-related downstream genes (Liscum and Reed, 2002; Li et al., 2016). Finally, TCP20 (LG1) is a transcription factor that induces expression of *LIPOXYGENASE2* (*LOX2*), a gene involved in jasmonate signalling pathway for leaf development and is down regulated by the miRNA *JAGGED AND WAVY* (JAW) (Danisman et al., 2012).

However, the currently detected BTL regions found in the present study still ranged around 10 cM to 30 cM, and the corresponding genome assemblies are still fragmented that all markers were traced back to different scaffolds (i.e. only partial and fragmented genome sequences are available within the BTL regions). Thus, it is likely that more candidate genes can be found in the gaps between scaffolds, and the identity of the causative rosulate / unifoliate gene still remains elusive.

Table 6.25 List of developmental and hormone-related genes identified in genome annotation

Linkage group	**Protein name / gene name (italic)**	**Description**	**Reference**
LG1 (MapA LG3)	COMPASS-like H3K4 histone methylase component WDR5A	Regulates histone methylation that is related to floral transition through repressing *FLOWERING LOCUS T*	Jiang et al., 2011
	Ankyrin repeat domain-containing protein EMB506	Related to the embryogenesis, chlorophyll biogenesis and leaf development of cotyledon, true leaf, and cauline leaves	Despres et al., 2001; Latvala-Kilby and Kilby, 2006
	TCP20 protein	Involved in jasmonate signalling pathway and leaf development	Danisman et al., 2012
	Protease Do-like 9	Involved in cytokinin and light-signalling pathways	Chi et al., 2016
LG2 (MapA LG1)	Cytokinin riboside 5'-monophosphate phosphoribohydrolase (*LONELY GUY*)	Involved in cytokinin biosynthesis	Kuroha et al., 2009
	Auxin-responsive protein IAA8	Involved in auxin signalling	Overvoorde et al., 2005; Liscum and Reed, 2002
	SET domain-containing protein	Protein family that consist of genes such as *CURLY LEAF* that regulates meristem related gene	Ng et al., 2007
	Protein YABBY 5	Promotes adaxial cell identity and regulates the initiation of embryonic shoot apical meristem (SAM) development	Siegfried et al., 1999; Stahle et al., 2009
	Scarecrow-like protein 6 isoform X1	Involved in gibberellin signalling	Zhang et al., 2011
LG4 (MapA LG7)	Cytokinin riboside 5'-monophosphate phosphoribohydrolase (*LONELY GUY*)	Cytokinin-activating enzyme that hydrolise cytokinin riboside 5'-monophosphate and release bioactive cytokinin	Kuroha et al., 2009
	COMPASS-like H3K4 histone methylase component WDR5A	Regulates histone methylation that is related to floral transition through repressing *FLOWERING LOCUS T*	Jiang et al., 2011
	Ankyrin repeat protein SKIP35	Related to the embryogenesis, chlorophyll biogenesis and leaf development of cotyledon, true leaf, and cauline leaves	Despres et al., 2001; Latvala-Kilby and Kilby, 2006

Table 6.25 continued

Linkage group	Protein name / gene name (italic)	Description	Reference
LG10 (MapA LG6)	SPINDLY protein	Involved in gibberellin signalling	Silverstone et al., 2006
	LIGHT-DEPENDENT SHORT HYPOCOTYLS 3 (*LSH3*)	Expressed between SAM and lateral organs and may act as transcription factor for several meristem related genes	Cho and Zambryski, 2011
LG14 (MapA LG9)	Auxin response factor (ARF)	Involved in auxin signalling	Li et al., 2016
	Cytokinin oxidase 3	Involved in cytokinin degredation	Schmülling et al., 2003
	Zeatin O-glucosyltransferase (ZOG)	Inactivation of cytokinin molecule	Martin et al., 2001
	Protein SPEAR3 (TIE1)	Transcription factors related to leaf development	Tao et al., 2013

6.4.3 Genetic architectures of the floral traits and other vegetative traits

The genetics of floral dimensions, floral pigmentation, and flowering time traits were studied in *Streptocarpus* here. Most of the traits measured between the two *S. grandis* lineages (*S. grandis*^{F1} and *S. grandis*^{BC}) show little to no differences, yet the flowering time recorded in *S. grandis*^{F1} (265 DAS) is considerably shorter than that of *S. grandis*^{BC} (377 DAS). This is possibly due to the different sowing time of the two lineages, which the *S. grandis*^{F1} was sown in July 2015 and *S. grandis*^{BC} in January 2015 (Table 6.1). The difference in growing season may resulted in the variations in flowering time observed (M Möller personal communication).

Between *S. rexii* and *S. grandis*, dominance effects were observed for several traits, including corolla length, dilated tube length, corolla face height, tube opening height (outer), tube opening height (inner), style length, ventral lobe length, and dorsal tube length. In these traits the average phenotypic value of F1 was statistically similar to that *S. rexii*, implying that the *S. rexii* carries the dominant alleles for the QTL of these traits (Appendix 6.4). Dominance effects of the *S. grandis* alleles were observed in the trait ‘undilated tube width’ and ‘calyx length’, where the average phenotypic value of the F1 was more similar to the *S. grandis* value (Appendix 6.4; Trait 6 and 17). The three parental lineages show no statistical variation in the trait ‘undilated tube height’ (Appendix 6.4; Trait 4).

In the BC population, the segregation of the floral dimension traits were found to deviate from normal distributions (Table 6.10 and Appendix 6.6). Traits regulated by multiple QTLs with similar effect sizes usually segregated approximating a normal distribution (Lynch and Walsh, 1998). On the other hand, skewed, dichotomous, or even spike distributions of the trait can be the result of the presence of major-effect loci (Lynch and Walsh, 1998).

The segregation of floral pigmentation showed a gradual pattern, from near absence of stripes on the corolla tube floor to densely purple colour pigmentation (Figure 6.9). Both of the extreme phenotypes (Figure 6.9 a and d) were not observed in the parental lineages, and may be a result of transgressive segregation (Rieseberg et al., 1999).

In this study, the floral pigmentation was classified into binary traits (presence / absence), including lateral lobe pigmentation, ventral lobe pigmentation, and yellow spot (Figure 6.8). The segregation of the presence and absence of ventral lobe pigmentation was found to conform to a Mendelian 1:1 ratio, suggesting the presence of one major-effect genetic locus (Lawrence and Sturgess, 1957). However, the segregation of the yellow spot trait did not follow a 1:1 ratio as previously observed (Lawrence and Sturgess, 1957; Oehlkers, 1966). It is possible that the yellow spot trait cannot be correctly scored in plants where the flower exhibited a densely purple pigmented phenotype that masked the yellow spot (Figure 6.9 d). Thus, the number of 'yellow spot absence' individuals could be overestimated. In order to score the presence of yellow pigmentation correctly, a possible solution for future study is to use chromatography techniques to separate different pigments from petal extracts (Tatsuzaka and Hosokawa, 2015).

Evidence of significant phenotypic correlations was found for most of the measured floral traits (Figure 6.11). Strong correlations were found among the floral dimension traits (Figure 6.11, trait 1 - 24), suggesting that the overall flower size changes in a synchronised fashion, i.e. the flower were usually larger or smaller as a whole, rather than larger in some parts and smaller in others. Correlation results in the co-localisation of QTLs of more than 20 floral traits on LG2 (Figure 6.18) suggested the presence of a pleiotropic effect where a single gene is regulating multiple phenotypes (Lynch and Walsh, 1998). Pleiotropic effects are known for several genes regulating floral organ size, such as *AINTEGUMENTA* and the auxin-related *ARGOS* genes (Weiss et al., 2005). Almost all floral dimension traits were negatively correlated to flowering time (Figure 6.11, trait 25). In other words, the later a plant flowers the smaller the flower is. The negative correlation between the floral dimension traits and flowering time trait may also be explained by their co-localised QTL on LG2 (Figure 6.18), and the effect plot showing that homozygous (b) genotype at the flowering time locus resulted in later flowering (Appendix 6.9 x), in contradiction to other traits with the homozygous genotype (thus genetically more similar to *S. grandis* which has smaller flowers) leading to smaller floral organs (Appendix 6.9).

The rest of the traits, including floral pigmentation, rosulate / unifoliate, accessory phyllomorph, and two macrocotyledons, showed less apparent correlation patterns (Figure 6.11, trait 26 - 34). The three pigmentation traits were all found correlated to the tube opening height (Figure 6.10, trait 26 - 28). In particular, lateral lobe and ventral lobe pigmentation were significantly positively correlated to the tube opening height and (though not significantly) to tube opening width (Figure 6.11, trait 12 - 13). This implies that the wider the corolla tube opening is, the more likely the flower is pigmented. This can be explained by the co-localisation of their effective loci on LG3, which do not overlap but are genetically linked (Figure 6.18). Variations in corolla tube opening are usually associated with sizes of different pollinators (Hilliard and Burtt, 1971), and pigmentation patterns on flowers such as stripes are often considered as nectar guides for the pollinators (Leonard and Papaj, 2011). The co-localisation of these two effective loci on the same linkage group suggest that the traits are more likely to cosegregate, which may contribute to the pollination syndrome of larger flowers with more distinct nectar guides (Lynch and Walsh, 1998). In addition, flowering time was found to be negatively correlated to all four rosulate / unifoliate scoring results (Figure 6.11; traits 29 - 32), suggesting that the later the plant flowers the more likely that the plant is rosulate. Their co-localised effective loci on LG14 may have contributed to the phenotypic correlation observed (Figure 6.18).

Small to medium-sized loci were detected for most of the traits, which explained about 10% to 25% of the phenotype variance (Table 6.14). On the other hand, loci with major-effect that explain more than 30% of the phenotype variance were found for pistil length, dorsal lobe length, flowering time, and the three pigmentation traits (Table 6.14). In particular, the loci identified for pistil length, flowering time, ventral lobe pigmentation and yellow spot contributed to 39.85%, 50.88%, 59.94%, and 45.58% of the variance, respectively. Very high LOD scores were obtained in the LOD curves of flowering time, ventral lobe pigmentation and yellow spot, with the LOD value of 14.24, 37.21, and 11.53, respectively (Table 6.14). The identification of one major effect locus and the 1:1 segregation ratio of the ventral lobe pigmentation trait supports the previous Mendelian inheritance observation (Lawrence and Sturgess, 1957). Overall, these results suggests that major effect loci of these traits are tightly linked to our genetic markers, and further study of the corresponding genome regions may help identify the causative genes.

Interestingly, no genetic regions were found associated with the accessory phyllomorph and two macrocotyledons traits (Table 6.14). In particular, the presence of two macrocotyledons was only recorded in six of the BC individuals and is unlikely to provide sufficient linkage information to identify the effective loci. One possibility is that these two traits have low heritability, which a large proportion of the phenotype is not determined by genetic variance but instead by environmental factors or gene × environment interaction (Lynch and Walsh, 1998). The efficiency of QTL mapping is strongly influenced by the heritability of the trait studied, and the power to detect the QTL was found proportional to heritability, sample size and marker density (Li et al., 2010; Viana et al., 2016). These two traits are not commonly observed in the parental lineages and their genetic inheritance has not been documented before, i.e. whether a two macrocotyledons parents will lead to two macrocotyledon offspring, and whether selfing of *S. grandis* with accessory phyllomorph can and produce offspring with accessory phyllomorphs. Further study of these two traits is required to understand their genetic inheritance.

6.4.4 Conclusion

Five loci were identified to be associated with the rosulate / unifoliate growth forms. The loci on LG14 and LG2 were consistently found in most of the mapping analysis and may represent major loci regulating the formation of additional phyllomorphs. The LG14 locus could be associated with true phyllomorph development at earlier developmental stages, and the LG2 locus could be related to accessory phyllomorph development at later stages when plant is flowering. On the other hand, the identification of multiple loci and the novel phyllomorph morphologies observed suggested that the regulation of the growth form may be more complicated than previously hypothesised with more than two genes being involved. While the identity of the candidate 'rosulate' gene remained inconclusive, this study narrowed down the genetic regions for further investigation and several developmental related genes were annotated. In addition, the corresponding *S. grandis* genome scaffolds were retrieved. These resources provide the foundation for further fine mapping or resequencing study to pin down the exact location of the rosulate / unifoliate loci and to identify their sequences.

The genetic architecture of the floral traits was studied and several small to medium size QTLs were identified for most of the traits. Phenotypic correlations were found between many floral dimension traits, and were likely due to the co-localisation of QTL or pleiotropic effects; for example the LG2 locus was found associated with more than 20 floral dimension traits. Major effect loci were identified for pistil length, dorsal lobe length, flowering time, and the pigmentation traits. In particular, the ventral lobe pigmentation trait was found to follow a Mendelian 1:1 segregation ratio and a single major effect locus was identified on LG3.

Overall, the results add to our knowledge towards the genetic basis of *Streptocarpus* morphological characters. Their inheritance pattern, how the phenotypic traits were correlated, and the approximate location of the regulatory loci were uncovered. Further studies on the identified genetic regions to identify the causative genes would greatly enhance our understanding about the molecular regulation of these characters.

Chapter 7 Discussion and conclusions

This study revisits the classic genetic inheritance observation of the rosulate / unifoliate growth forms as well as other floral characters in the genus *Streptocarpus* using modern NGS technologies. The main outcomes of this study includes establishing the procedures for generating next generation sequencing data, and constructing the genomic and genetic resources for this non-model *Streptocarpus*. Analysis pipelines were set up and documented in detail for future reference and downstream applications. The results, including nucleic acid extraction protocols, genome and transcriptome assemblies, genetic maps, and QTL / BTL mapping approaches, revealed the genetic architectures of several morphological traits and offer the basis for future genomic research using *Streptocarpus* as study material.

7.1 *Streptocarpus* as a model system for developmental studies

The regulation of SAM development is a fundamental research topic for plant developmental biology as it concerns the formation of new lateral organs and self-perpetuation of the meristem (Laux et al., 1996; Lenhard et al., 2002; Nishii et al., 2010b). Developmental studies of model plant species greatly expanded our knowledge on how plant growth is regulated and what genetic network or hormones are involved. For example, previous work in *A. thaliana* revealed that the SAM activity is regulated by meristem identity genes including *WUSCHEL* (*WUS*), *CLAVATA* (*CLV*), and *SHOOTMERISTEMLESS* (*STM*), and the establishment of hormone gradients such as auxin, gibberellin, and cytokinin (reviewed in Soyars et al., 2016). On the other hand, studies using non-model species provide important insight into special biological features that are not present in model systems. Some of these species, referred to as non-model model organisms, possess unconventional and often unique properties that can be utilised to answer critical biological questions, including how major evolutionary processes were achieved (Russell et al., 2017).

The genus *Streptocarpus* is a highly suitable system to be developed for the studying of unconventional meristem regulation in plants for several reasons. First, species in the genus exhibit a highly unconventional development lacking a SAM in the embryo stage, which greatly deviates from most other angiosperm species (Imaichi et al., 2000; Mantegazza et al., 2007; Nishii et al., 2010b; 2016). Second, the genus consists of at least three distinct basic growth forms: caulescent, rosulate, and unifoliate (for latest classification see Nishii et al., 2015). By utilising the opportunity that viable hybrids between the growth forms can be produced between rosulates and unifoliates, it is possible to carry out genetic studies and identify developmental genes related to differences in rosulate and unifoliate vegetative habit (Chen et al. 2017). Third, an extensive and well-resolved phylogeny exists for *Streptocarpus* as well as chromosome counts (e.g. Jong and Möller, 2000; Nishii et al., 2015). Fourth, developmental processes have been studied in at least 8 species of different growth forms (e.g. Jong, 1970; Jong and Burtt, 1975; Imaichi et al., 2007; Nishii and Nagata, 2007; Nishii et al., 2017). This knowledge provided the opportunity for choosing appropriate study material. Fifth, the cultivation method of *Streptocarpus* is well established during its development as an ornamental horticultural plant, and they can be mass propagated sexually and asexually in conventional temperature-controlled glasshouses and it is straight forward to produce inbred lines for genetic studies or genome sequencing. Sixth, the wet lab molecular techniques are well established, especially for *S. rexii*. This includes protocols for tissue sectioning, SEM, DNA and RNA extraction, RNA *in situ* hybridisation, and RT-PCR (see example in Nishii et al., 2017). These techniques have been applied to study several meristem related gene expression, such as *STM* (Harrison et al., 2005; Mantegazza et al., 2009; Nishii et al., 2017), *WUS* (Mantegazza et al., 2009), and *ASSYMETRIC LEAVES1 / ROUGH SHEATH 2 / PHANTASTICA* (ARP; Nishii et al., 2010a). All this invaluable knowledge helped establishing *Streptocarpus* as a study system by enabling the relatively rapid setting up of investigations using this material.

As for this study, it presents optimised NGS workflows and provides genomic, transcriptomic, and genetic resources of *Streptocarpus*. The NGS workflow, including the preparation of nucleic acid samples and detail documentation of bioinformatics analysis pipelines and parameters, will be beneficial for future NGS works of *Streptocarpus*. The draft genomes of *S. rexii* and *S. grandis* can serve as the reference sequence for future sequencing experiments, designing new markers, or for gene identification. The transcriptomes of the two species provides gene sequence information which is useful for candidate gene isolation. The genetic maps and the QTL mapping revealed the genetic architectures of several morphological traits, including floral dimension, floral pigmentation, flowering time, and growth form variations. These traits are important in terms of the evolution of pollination syndromes or vegetative habits, and may also be interesting for their ornamental values that may potentially be utilised through marker assisted selection to facilitate the generation of new cultivars (Kole and Abbott, 2008). The mapping results narrowed down the genetic region to be screened for morphological trait-related genes, which can ultimately help resolving the molecular mechanisms that shape the morphological diversities exhibited in this genus.

In conclusion, this work will be a key step for establishing *Streptocarpus* as a model system for studying plant meristem regulation and growth form evolution, and in a broader sense provide useful genomic resources for the Gesneriaceae family. This study can also be an example on the methodological approaches for establishing genomic resources for non-model plant organisms.

7.2 Directions for future studies

In this study, the genetic loci identified for the rosulate / unifoliate growth form trait were not specific enough to generate a short list of candidate genes for functional verification. One possible reason is the limitation of the population size, which in current study a modest number of 200 BC individuals were used, and by increasing the population size the sensitivity and accuracy of QTL mapping can potentially be improved (Vales et al., 2005; Li et al., 2006; Raghavan and Collard, 2012). However, the resolution in QTL mapping may still be dependent on the species and mapping population used even when similar number of individual and mapping algorithm were taken. For example, QTL mapping in *Primulina* sp. used 201 F2 individuals for composite interval mapping (similar to the 200 BC individuals in current study) and achieved to narrow down the confidence interval to 0.5 cM to 2 cM (Feng et al., 2018). On the other hand, QTL mapping in *Rhytidophyllum* sp. used 177 F2 individuals for standard interval mapping, and obtained QTL confidence interval ranging from 10 cM to 120 cM (Alexandre et al,. 2015). It is therefore important to consider the limitation on our current mapping population and QTL mapping strategy, and whether increasing the population size can greatly enhance the mapping result or not. This is partly related to the intrinsic limitations of the QTL / BTL mapping methodology: firstly, only the genetic variation segregating between the two parents can be tested (Borevitz and Nordborg, 2003), and secondly, the resolution of the mapping relies on the recombination events that occurred within the mapping population (Balasubramanian et al., 2009). Both factors limited the number of recombination events recorded, thus thus reduce the resolution of the map. To overcome these limitations, an alternative approach is perhaps through fine-mapping method using a different mapping population and develop new markers to increase the marker density within the current BTL loci (Cockram et al., 2015; Calderon et al., 2016). In this approach advanced inbred lines (AILs) are typically used, such as recombinant inbred lines (RILs), near-isogenic lines (NILs) or Multiparent Advanced Generation Inter-Cross (Cavanagh et al., 2008; Gonzales and Palmer, 2014; Schneeberger, 2014). So far, we have generated a BC2 progeny by backcrossing BC individuals to the *S. grandis* parent, and in the future we can construct NILs by continuously backcrossing the progenies. Another alternative is the NGS-based bulk-segregant analysis (Schneeberger, 2014): by pooling the genomic DNA of individuals with the same phenotype (either rosulate or unifoliate) and performing whole genome resequencing on the pooled DNA, the output data are expected to cover the genome regions outside the original RAD-Seq markers, thus more genetic variations can be observed and the SNPs genotype frequency can be compared between the two phenotypes to narrow down the candidate regions (Schneeberger, 2014). This method was used to identify the *Hairy* gene responsible for trichome development in *A. majus* using a NIL (9th backcross generation) population (Tan, 2018).

Another important aspect for future study is to improve the contiguity of the genome assembly. A high contiguity genome is crucial for candidate gene isolation, as the mapped loci on the genetic linkage map will eventually be integrated with the physical map (genome sequence) so that the genome sequence corresponding to the loci can be examined for differences at SNPs level (Yang et al., 2004; Zhou et al., 2015). In the current *Streptocarpus* genome assembly, multiple genome scaffolds were found inside the rosulate / unifoliate loci and the sequences between these scaffolds (gaps) remained unknown. This can be problematic as the physical distance (bp) of the gaps can be too large for candidate gene screening, or the actual causative genes may be located in these gaps and are yet to be discovered. To improve the genome contiguity we may incorporate new genome sequencing data based on mate-pair library or long read sequencing such as PacBio and Nanopore (Jiao and Schneeberger, 2017; Li and Harkess, 2018). In particular, the latest Nanopore device PromethION is expected to generate 50 Gbp of data per flow cell and produce the longest read among currently available long-range sequencing technologies, and with modified protocols up to 882 Kbp reads can be achieved (Loose, 2017; Jain et al., 2018). In the most ideal case this would suggest an approximately 50× depth of coverage of long read data for the *Streptocarpus* genome (~1 Gbp). While these technologies are mostly been tested in human at current stage, applications on plant materials has been proven successful (Michael et al., 2018). All sequencing technologies mentioned above have just been made available at the Edinburgh Genomics facility, which can be beneficial for the near future work. In addition, a genome can be further improved by anchoring the assembled scaffolds to the genetic map to achieve chromosome-level assembly (Fierst, 2015; Jiao and Schneeberger, 2017). Tools such as Chromonomers (Small et al., 2016) are designed for increasing genome contiguity based on RAD-Seq derived genetic maps, which may be suitable for our data.

The annotation of the *Streptocarpus* genome should be improved for searching the candidate genes. The current annotation results were partially based on aligning the genes of distantly related model species (i.e. *N. tabacum*) to the *Streptocarpus* genome, and the pipeline had a stringent cut-off threshold for the alignment, which must show > 90% identity and coverage when aligning (Numa and Itoh, 2014). This suggests that if a *Nicotiana* gene provided in the pipeline is sharing low sequence homology to the *Streptocarpus* gene, it may not be aligned and hence the gene in our genome will not be annotated. This can be resolved by mapping the *Streptocarpus* RNA-Seq data to the genome assembly, which will help predicting correct intron-exon structures and potentially identify more functioning genes (Bolger et al., 2017b; Dominguez Del Angel et al., 2018). Tools such as BRAKER were developed for this purpose and have been widely applied to plant genome annotation (Hoff et al., 2016).

On the other hand, molecules such as small RNAs are important regulator for plant growth, and are involved in developmental processes such as meristem regulation, leaf development, flower development, and floral pigmentation patterning (D'Ario et al., 2017; Bradley et al., 2017). Yet, the prediction of small RNAs in genome assemblies is not as straightforward as predicting protein coding genes due to their small size and poorly conserved sequence homology (20 to 24 nucleotides; D'Ario et al., 2017). Current bioinformatics tools are limited to small RNA predictions in bacterial genomes (Lindgreen et al., 2014; Li and Kwan, 2014; Dominguez Del Angel et al., 2018). The recently released Rfam 13.0, a genome-centric resource, greatly expanded the collection of small RNA sequences (Kalvari et al., 2018), and incorporation of the RNA alignment tools such as 'Infernal' may be applicable for small RNA prediction in the near future (Nawrocki and Eddy, 2013; Barquist et al., 2016).

In addition to the rosulate / unifoliate growth forms, several major effect loci were found for the floral traits examined in this study that could be interesting targets for further genetic fine mapping. The photo records of all plant materials, particularly the floral photos, may potentially be used for different phenotyping approaches such as geometric morphometrics for more precise quantification of shape and pattern variations (e.g. Hsu et al., 2015; Sun et al., 2017; Hsu et al., 2018). Establishment of transgenic systems for the *Streptocarpus* materials would be another important method to establish for functional verification of candidate genes. *Agrobacterium*-mediated transformation and particle bombardment systems were developed for *Saintpaulia* (now *Streptocarpus*) materials (Mercuri et al., 2000; Kushikawa et al., 2001; Ghorbanzade and Ahmadabadi, 2015) and were used in *glucanase*-*chitinase* and *AtIPT5* (*ISOPENTENYLTRANSFERASE*) transgenic studies (Ram and Mohandas, 2003; Ye et al., 2014). Genome editing methods, such as the clustered regularly interspaced short palindromic repeats (CRISPR)-Cas9, has yet to be applied to Gesneriaceae species, but could potentially be used to construct knockout mutant or alter transcription levels to study candidate gene function (Bortesi and Fischer, 2015). Incorporation of the new technologies mentioned above would enable the determination of the molecular mechanisms underlying the diverse morphologies present in *Streptocarpus* species, and ultimately improve our understanding on how this diversity has evolved.

References

Adams, M.D., Celniker, S.E., Holt, R.A., Evans, C.A., Gocayne, J.D., Amanatides, P.G., Scherer, S.E., Li, P.W., Hoskins, R.A., Galle, R.F., George, R.A., Lewis, S.E., Richards, S., Ashburner, M., Henderson, S.N., Sutton, G.G., Wortman, J.R., Yandell, M.D., Zhang, Q., Chen, L.X., Brandon, R.C., Rogers, Y.H., Blazej, R.G., Champe, M., Pfeiffer, B.D., Wan, K.H., Doyle, C., Baxter, E.G., Helt, G., Nelson, C.R., Gabor, G.L., Abril, J.F., Agbayani, A., An, H.J., Andrews-Pfannkoch, C., Baldwin, D., Ballew, R.M., Basu, A., Baxendale, J., Bayraktaroglu, L., Beasley, E.M., Beeson, K.Y., Benos, P.V., Berman, B.P., Bhandari, D., Bolshakov, S., Borkova, D., Botchan, M.R., Bouck, J., Brokstein, P., Brottier, P., Burtis, K.C., Busam, D.A., Butler, H., Cadieu, E., Center, A., Chandra, I., Cherry, J.M., Cawley, S., Dahlke, C., Davenport, L.B., Davies, P., de Pablos, B., Delcher, A., Deng, Z., Mays, A.D., Dew, I., Dietz, S.M., Dodson, K., Doup, L.E., Downes, M., Dugan-Rocha, S., Dunkov, B.C., Dunn, P., Durbin, K.J., Evangelista, C.C., Ferraz, C., Ferriera, S., Fleischmann, W., Fosler, C., Gabrielian, A.E., Garg, N.S., Gelbart, W.M., Glasser, K., Glodek, A., Gong, F., Gorrell, J.H., Gu, Z., Guan, P., Harris, M., Harris, N.L., Harvey, D., Heiman, T.J., Hernandez, J.R., Houck, J., Hostin, D., Houston, K.A., Howland, T.J., Wei, M.H., Ibegwam, C., Jalali, M., Kalush, F., Karpen, G.H., Ke, Z., Kennison, J.A., Ketchum, K.A., Kimmel, B.E., Kodira, C.D., Kraft, C., Kravitz, S., Kulp, D., Lai, Z., Lasko, P., Lei, Y., Levitsky, A.A., Li, J., Li, Z., Liang, Y., Lin, X., Liu, X., Mattei, B., McIntosh, T.C., McLeod, M.P., McPherson, D., Merkulov, G., Milshina, N.V., Mobarry, C., Morris, J., Moshrefi, A., Mount, S.M., Moy, M., Murphy, B., Murphy, L., Muzny, D.M., Nelson, D.L., Nelson, D.R., Nelson, K.A., Nixon, K., Nusskern, D.R., Pacleb, J.M., Palazzolo, M., Pittman, G.S., Pan, S., Pollard, J., Puri, V., Reese, M.G., Reinert, K., Remington, K., Saunders, R.D., Scheeler, F., Shen, H., Shue, B.C., Sidén-Kiamos, I., Simpson, M., Skupski, M.P., Smith, T., Spier, E., Spradling, A.C., Stapleton, M., Strong, R., Sun, E., Svirskas, R., Tector, C., Turner, R., Venter, E., Wang, A.H., Wang, X., Wang, Z.Y., Wassarman, D.A., Weinstock, G.M., Weissenbach, J., Williams, S.M., WoodageT, Worley, K.C., Wu, D., Yang, S., Yao, Q.A., Ye, J., Yeh, R.F., Zaveri, J.S., Zhan, M., Zhang, G., Zhao, Q., Zheng, L., Zheng, X.H., Zhong, F.N., Zhong, W., Zhou, X., Zhu, S., Zhu, X., Smith, H.O., Gibbs, R.A., Myers, E.W., Rubin, G.M., and Venter, J.C. (2000). The genome sequence of *Drosophila melanogaster*. Science 287: 2185–2195.

Ai, B., Gao, Y., Zhang, X., Tao, J., Kang, M., and Huang, H. (2014). Comparative transcriptome resources of eleven *Primulina* species, a group of "stone plants" from a biodiversity hot spot. Mol Ecol Resour 15: 619–632.

Alexandre, H., Vrignaud, J., Mangin, B., and Joly, S. (2015). Genetic architecture of pollination syndrome transition between hummingbird-specialist and generalist species in the genus *Rhytidophyllum* (Gesneriaceae). PeerJ 3: e1028.

Allen, G.C., Flores-Vergara, M.A., Krasynanski, S., Kumar, S., and Thompson, W.F. (2006). A modified protocol for rapid DNA isolation from plant tissues using cetyltrimethylammonium bromide. Nat Protoc 1: 2320–2325.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. J Mol Biol 215: 403–410.

Amores, A., Catchen, J., Ferrara, A., Fontenot, Q., and Postlethwait, J.H. (2011). Genome evolution and meiotic maps by massively parallel DNA sequencing: spotted gar, an outgroup for the teleost genome duplication. Genetics 188: 799–808.

Andrews, S. (2010). FastQC: a quality control tool for high throughput sequence data. Available at: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Apweiler, R., Bairoch, A., Wu, C.H., Barker, W.C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., Martin, M.J., Natale, D.A., O'Donovan, C., Redaschi, N., and Yeh, L.-S.L. (2004). UniProt: the Universal Protein knowledgebase. Nucleic Acids Res 32: D115–D119.

Arabidopsis Genome Initiative (2000). Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature 408: 796–815.

Aranda, R., Dineen, S.M., Craig, R.L., Guerrieri, R.A., and Robertson, J.M. (2009). Comparison and evaluation of RNA quantification methods using viral, prokaryotic, and eukaryotic RNA over a 104 concentration range. Anal Biochem 387: 122–127.

Azizi, P., Rafii, M.Y., Maziah, M., Abdullah, S.N.A., Hanafi, M.M., Latif, M.A., Rashid, A.A., and Sahebi, M. (2015). Understanding the shoot apical meristem regulation: A study of the phytohormones, auxin and cytokinin, in rice. Mech Dev 135: 1–15.

Bai, B., Wang, L., Zhang, Y.J., Lee, M., Rahmadsyah, R., Alfiko, Y., Ye, B.Q., Purwantomo, S., Suwanto, A., Chua, N.-H., and Yue, G.H. (2018). Developing genome-wide SNPs and constructing an ultrahigh-density linkage map in oil palm. Sci Rep 8: 691-698.

Baird, N.A., Etter, P.D., Atwood, T.S., Currey, M.C., Shiver, A.L., Lewis, Z.A., Selker, E.U., Cresko, W.A., and Johnson, E.A. (2008). Rapid SNP discovery and genetic mapping using sequenced RAD markers. PLoS ONE 3: e3376.

Baker, M. (2012). *De novo* genome assembly: what every biologist should know. Nat Meth 9: 333–337.

Balasubramanian, S., Schwartz, C., Singh, A., Warthmann, N., Kim, M.C., Maloof, J.N., Loudet, O., Trainer, G.T., Dabi, T., Borevitz, J.O., Chory, J., and Weigel, D. (2009). QTL mapping in new *Arabidopsis thaliana* advanced intercross-recombinant inbred lines. PLoS ONE 4: e4318.

Banks, J.A. (2015). The evolution of the shoot apical meristem from a gene expression perspective. New Phytol. 207: 486–487.

Barquist, L., Burge, S.W., and Gardner, P.P. (2016). Studying RNA homology and conservation with Infernal: from single sequences to RNA families. Curr Protoc Bioinformatics 54: 12.13.1–12.13.25.

Barton, M.K. and Poethig, R.S. (1993). Formation of the shoot apical meristem in *Arabidopsis thaliana*: an analysis of development in the wild type and in the *shoot meristemless* mutant. Development 119: 823–831.

Bauer, D.F. (1972). Constructing confidence sets using rank statistics. J Am Stat Assoc 67: 687–690.

Beavis, W.D. and Grant, D. (1991). A linkage map based on information from four F2 populations of maize (*Zea mays* L.). Theor Appl Genet 82: 636–644.

Best, D.J. and Roberts, D.E. (1975). Algorithm AS 89: the upper tail probabilities of Spearman's rho. J R Stat Soc Series C 24: 377–379.

Bolger, A.M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30: 2114–2120.

Bolger, M., Schwacke, R., Gundlach, H., Schmutzer, T., Chen, J., Arend, D., Oppermann, M., Weise, S., Lange, M., Fiorani, F., Spannagl, M., Scholz, U., Mayer, K., and Usadel, B. (2017a). From plant genomes to phenotypes. J Biotechnol 261: 46–52.

Bolger, M.E., Arsova, B., and Usadel, B. (2017b). Plant genome and transcriptome annotations: from misconceptions to simple solutions. Brief Bioinformatics 19: 437–449.

Borevitz, J.O. and Nordborg, M. (2003). The impact of genomics on the study of natural variation in *Arabidopsis*. Plant Physio 132: 718–725.

Bortesi, L. and Fischer, R. (2015). The CRISPR/Cas9 system for plant genome editing and beyond. Biotechnol Adv 33: 41–52.

Bradley, D., Xu, P., Mohorianu, I.-I., Whibley, A., Field, D., Tavares, H., Couchman, M., Copsey, L., Carpenter, R., Li, M., Li, Q., Xue, Y., Dalmay, T., and Coen, E. (2017). Evolution of flower color pattern through selection on regulatory small RNAs. Science 358: 925–928.

Bradshaw, H.D., Otto, K.G., Frewen, B.E., McKay, J.K., and Schemske, D.W. (1998). Quantitative trait loci affecting differences in floral morphology between two species of monkeyflower (*Mimulus*). Genetics 149: 367–382.

Broman, K.W. (2003). Mapping quantitative trait loci in the case of a spike in the phenotype distribution. Genetics 163: 1169–1175.

Brunt, A.A. (1971). Some hosts and properties of dahlia mosaic virus. Ann Appl Biol 67: 357–368.

Buckingham, L. and Flaws, M.L. (2007). Molecular Diagnostics: Fundamentals, Methods, & Clinical Applications. F.A. Davis Company. Philadelphia. USA.

Burtt, B.L. (1939). Notes on *Streptocarpus*. Bull Misc Inform Kew 1939: 68–84.

Buta, E., Cantor, M., Buta, M., and Zaharia, A. (2010). The effect of rooting substrates on the development of leaf cuttings of *Saintpaulia*. Analele Univ Craiova 15: 110–119.

Cabanettes, F. and Klopp, C. (2018). D-GENIES: dot plot large genomes in an interactive, efficient and simple way. PeerJ 6: e4958.

Calderon, C.I., Yandell, B.S., and Doebley, J.F. (2016). Fine mapping of a QTL associated with kernel row number on chromosome 1 of maize. PLoS ONE 11: e0150276.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., and Madden, T.L. (2009). BLAST+: architecture and applications. BMC Bioinformatics 10: 421-426.

Campbell, N.R., LaPatra, S.E., Overturf, K., Towner, R., and Narum, S.R. (2014). Association mapping of disease resistance traits in rainbow trout using restriction site associated DNA sequencing. G3 4: 2473–2481.

Caspary, R. (1858). Über die Anisokotylie von *Streptocarpus polyanthus* Hook. und *Streptocarpus rexii*. Lindl.. Verh Naturhist Vereines Preuss Rheinl Westphalens 15: 74–75.

Catchen, J., Hohenlohe, P.A., Bassham, S., Amores, A., and Cresko, W.A. (2013). Stacks: an analysis tool set for population genomics. Mol Ecol 22: 3124–3140.

Catchen, J.M., Amores, A., Hohenlohe, P., Cresko, W., and Postlethwait, J.H. (2011). Stacks: building and genotyping loci *de novo* from short-read sequences. G3 1: 171–182.

Catchen, J.M., Hohenlohe, P.A., Bernatchez, L., Funk, W.C., Andrews, K.R., and Allendorf, F.W. (2017). Unbroken: RADseq remains a powerful tool for understanding the genetics of adaptation in natural populations. Mol Ecol Resour 17: 362–365.

Causse, M., Saliba-Colombani, V., Lecomte, L., Duffé, P., Rousselle, P., and Buret, M. (2002). QTL analysis of fruit quality in fresh market tomato: a few chromosome regions control the variation of sensory and instrumental traits. J Exp Bot 53: 2089–2098.

Cavaller-Smith, T. (1985). The evolution of genome size. John Wiley and Sons Inc., New York. USA.

Cavanagh, C., Morell, M., Mackay, I., and Powell, W. (2008). From mutations to MAGIC: resources for gene discovery, validation and delivery in crop plants. Curr Opin Plant Biol 11: 215–221.

Chen, Y.-Y., Nishii, K., Barber, S., Hackett, C., Kidner, C.A., Gharbi, K., Nagano, A.J., Iwamoto, A., and Möller, M. (2018). A first genetic map in the genus *Streptocarpus* generated with RAD sequencing based SNP markers. S Afr J Bot 117: 158–168.

Chen, Y.-Y., Nishii, K., Spada, A., Wang, C.-N., Sakakibara, H., Kojima, M., Wright, F., MacKenzie, K., and Möller, M. (2017). Cytokinin biosynthesis *ISOPENTENYLTRANSFERASE* genes are differentially expressed during phyllomorph development in the acaulescent *Streptocarpus rexii* (Gesneriaceae). S Afr J Bot 109: 96–111.

Chi, W., Li, J., He, B., Chai, X., Xu, X., Sun, X., Jiang, J., Feng, P., Zuo, J., Lin, R., Rochaix, J.-D., and Zhang, L. (2016). DEG9, a serine protease, modulates cytokinin and light signaling by regulating the level of *ARABIDOPSIS* RESPONSE REGULATOR 4. Proc Natl Acad Sci USA 113: E3568–E3576.

Chiara, M., Horner, D.S., and Spada, A. (2013). *De Novo* assembly of the transcriptome of the non-model plant *Streptocarpus rexii* employing a novel heuristic to recover locus-specific transcript clusters. PLoS ONE 8: e80961.

Cho, E. and Zambryski, P.C. (2011). *ORGAN BOUNDARY1* defines a gene expressed at the junction between the shoot apical meristem and lateral organs. Proc Natl Acad Sci USA 108: 2154–2159.

Chomczynski, P. and Sacchi, N. (1987). Single-step method of RNA isolation by acid guanidinium thiocyanate-phenol-chloroform extraction. Anal Biochem 162: 156–159.

Chou, S.-P. (2008). Genetic analysis of pollination-related floral shape in *Streptocarpus* (Gesneriaceae). Master thesis. National Taiwan University. Taipei. Taiwan.

Christenhusz, M.J.M. (2015). On African violets and Cape primroses—towards a monophyletic *Streptocarpus* (Gesneriaceae). Phytotaxa 46: 3–8.

Christou, A., Georgiadou, E.C., Filippou, P., Manganaris, G.A., and Fotopoulos, V. (2014). Establishment of a rapid, inexpensive protocol for extraction of high quality RNA from small amounts of strawberry plant tissues and other recalcitrant fruit crops. Gene 537: 169–173.

Chu, T.-C., Lu, C.-H., Liu, T., Lee, G.C., Li, W.-H., and Shih, A.C.-C. (2013). Assembler for *de novo* assembly of large genomes. Proc Natl Acad Sci USA 110: E3417–E3424.

Chutimanitsakun, Y., Nipper, R.W., Cuesta-Marcos, A., Cistué, L., Corey, A., Filichkina, T., Johnson, E.A., and Hayes, P.M. (2011). Construction and application for QTL analysis of a Restriction Site Associated DNA (RAD) linkage map in barley. BMC Genomics 12: 4-17.

Claros, M.G., Bautista, R., Guerrero-Fernández, D., Benzerki, H., Seoane, P., and Fernández-Pozo, N. (2012). Why assembling plant genome sequences is so challenging. Biology 1: 439–459.

Cockram, J., Scuderi, A., Barber, T., Furuki, E., Gardner, K.A., Gosman, N., Kowalczyk, R., Phan, H.P., Rose, G.A., Tan, K.-C., Oliver, R.P., and Mackay, I.J. (2015). Fine-mapping the wheat *Snn1* locus conferring sensitivity to the *Parastagonospora nodorum* necrotrophic effector *SnTox1* using an eight founder Multiparent Advanced Generation Inter-Cross population. G3 5: 2257–2266.

Coffman, C.J., Doerge, R.W., Simonsen, K.L., Nichols, K.M., Duarte, C.K., Wolfinger, R.D., and McIntyre, L.M. (2005). Model selection in binary trait locus mapping. Genetics 170: 1281–1297.

Compeau, P.E.C., Pevzner, P.A., and Tesler, G. (2017). Why are de Bruijn graphs useful for genome assembly? Nature Biotechnol 29: 987–991.

Conesa, A., Gotz, S., Garcia-Gomez, J.M., Terol, J., Talon, M., and Robles, M. (2005). Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics 21: 3674–3676.

Currey, C.J. and Flax, N.J. (2015). Ethephon foliar sprays prevent premature flowering of tissue culture-propagated *Streptocarpus* hybrids. Horttechnology 25: 635–638.

D'Ario, M., Griffiths-Jones, S., and Kim, M. (2017). Small RNAs: big impact on plant development. Trends Plant Sci 22: 1056–1068.

Da Silva, A.C.R., Ferro, J.A., Reinach, F.C., Farah, C.S., Furlan, L.R., Quaggio, R.B., Monteiro-Vitorello, C.B., Van Sluys, M.A., Almeida, N.F., Alves, L.M.C., do Amaral, A.M., Bertolini, M.C., Camargo, L.E.A., Camarotte, G., Cannavan, F., Cardozo, J., Chambergo, F., Ciapina, L.P., Cicarelli, R.M.B., Coutinho, L.L., Cursino-Santos, J.R., El-Dorry, H., Faria, J.B., Ferreira, A.J.S., Ferreira, R.C.C., Ferro, M.I.T., Formighieri, E.F., Franco, M.C., Greggio, C.C., Gruber, A., Katsuyama, A.M., Kishi, L.T., Leite, R.P., Lemos, E.G.M., Lemos, M.V.F., Locali, E.C., Machado, M.A., Madeira, A.M.B.N., Martinez-Rossi, N.M., Martins, E.C., Meidanis, J., Menck, C.F.M., Miyaki, C.Y., Moon, D.H., Moreira, L.M., Novo, M.T.M., Okura, V.K., Oliveira, M.C., Oliveira, V.R., Pereira, H.A., Rossi, A., Sena, J.A.D., Silva, C., de Souza, R.F., Spinola, L.A.F., Takita, M.A., Tamura, R.E., Teixeira, E.C., Tezza, R.I.D., Trindade dos Santos, M., Truffi, D., Tsai, S.M., White, F.F., Setubal, J.C., and Kitajima, J.P. (2002). Comparison of the genomes of two *Xanthomonas* pathogens with differing host specificities. Nature 417: 459–463.

Danisman, S., van der Wal, F., Dhondt, S., Waites, R., de Folter, S., Bimbo, A., van Dijk, A.D.J., Muino, J.M., Cutri, L., Dornelas, M.C., Angenent, G.C., and Immink, R.G.H. (2012). *Arabidopsis* class I and class II TCP transcription factors regulate jasmonic acid metabolism and leaf development antagonistically. Plant Physio 159: 1511–1523.

Darbyshire, I. and Massingue, A.O. (2014). Two new species of *Streptocarpus* (Gesneriaceae) from tropical Africa. Edinb J Bot 71: 3–13.

Darling, A.E., Mau, B., and Perna, N.T. (2010). progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. PLoS ONE 5: e11147.

Darling, A.E., Mau, B., Blattner, F.R., and Perna, N.T. (2004). Mauve: multiple alignment of conserved genomic sequence with rearrangements. Genome Res 14: 1394–1403.

Davey, J.L. and Blaxter, M.W. (2010). RADSeq: next-generation population genetics. Brief Funct Genomics 9: 416–423.

Davey, J.W., Cezard, T., Fuentes-Utrilla, P., Eland, C., Gharbi, K., and Blaxter, M.L. (2012). Special features of RAD Sequencing data: implications for genotyping. Mol Ecol 22: 3151–3164.

Davey, J.W., Hohenlohe, P.A., Etter, P.D., Boone, J.Q., Catchen, J.M., and Blaxter, M.L. (2011). Genome-wide genetic marker discovery and genotyping using next-generation sequencing. Nat Rev Genet 12: 499–510.

De Candolle, A. (1845). 'Cyrtandraceae'. in *Prodromus Systematis Naturalis Regni Vegetabilis* (ed. By De Candolle). Treittel et Würtz. Paris. France. pp. 277-286.

Dees, M.W., Lysøe, E., Rossmann, S., Perminow, J., and Brurberg, M.B. (2017). *Pectobacterium polaris* sp. nov., isolated from potato (*Solanum tuberosum*). Int J Syst Evol Microbiol 67: 5222–5229.

Delcher, A.L., Phillippy, A., Carlton, J., and Salzberg, S.L. (2002). Fast algorithms for large-scale genome alignment and comparison. Nucleic Acids Res 30: 2478–2483.

Despres, B., Delseny, M., and Devic, M. (2001). Partial complementation of embryo defective mutations: a general strategy to elucidate gene function. Plant J. 27: 149–159.

Dierckxsens, N., Mardulyn, P., and Smits, G. (2017). NOVOPlasty: *de novo* assembly of organelle genomes from whole genome data. Nucleic Acids Res 45: e18.

Dittrich-Reed, D.R. and Fitzpatrick, B.M. (2012). Transgressive hybrids as hopeful monsters. Evol Biol 40: 310–315.

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29: 15–21.

Dominguez Del Angel, V., Hjerde, E., Sterck, L., Capella-Gutierrez, S., Notredame, C., Vinnere Pettersson, O., Amselem, J., Bouri, L., Bocs, S., Klopp, C., Gibrat, J.-F., Vlasova, A., Leskosek, B.L., Soler, L., Binzer-Panchal, M., and Lantz, H. (2018). Ten steps to get started in genome assembly and annotation. F1000Res pii: ELIXER–148.

Dubuc-Lebreux, M.A. (1978). Modification on the unifoliate habit of *Streptocarpus wendlandii* and *Streptocarpus michelmorei* by some growth regulators. Phytomorphology 28: 224–238.

Dunn, O.J. (1964). Multiple comparisons using rank sums. Technometrics 6: 241–252.

Eckshtain-Levi, N., Shkedy, D., Gershovits, M., Da Silva, G.M., Tamir-Ariel, D., Walcott, R., Pupko, T., and Burdman, S. (2016). Insights from the genome sequence of *Acidovorax citrulli* M6, a group I strain of the causal agent of bacterial fruit blotch of cucurbits. Front Microbiol 7: 969–1012.

Edwards, B.J., Haynes, C., Levenstien, M.A., Finch, S.J., and Gordon, D. (2005). Power and sample size calculations in the presence of phenotype errors for case/control genetic association studies. BMC Genet 6: 18.

Ekblom, R. and Wolf, J.B.W. (2014). A field guide to whole-genome sequencing, assembly and annotation. Evol Appl 7: 1026–1042.

Ellegren, H. (2014). Genome sequencing and population genomics in non-model organisms. Trends Ecol Evol 29: 51–63.

Elshire, R.J., Glaubitz, J.C., Sun, Q., Poland, J.A., Kawamoto, K., Buckler, E.S., and Mitchell, S.E. (2011). A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS ONE 6: e19379.

Emms, D.M. and Kelly, S. (2015). OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. Genome Biol 16: 157–171.

Endrullat, C., Glökler, J., Franke, P., and Frohme, M. (2016). Standardization and quality management in next-generation sequencing. Appl Transl Genom 10: 2–9.

Engstrom, E.M., Andersen, C.M., Gumulak-Smith, J., Hu, J., Orlova, E., Sozzani, R., and Bowman, J.L. (2011). *Arabidopsis* homologs of the petunia hairy meristem gene are required for maintenance of shoot and root indeterminacy. Plant Physio 155: 735–750.

Etter, P.D., Bassham, S., Hohenlohe, P.A., Johnson, E.A., and Cresko, W.A. (2011). SNP discovery and genotyping for evolutionary genetics using RAD sequencing. Methods Mol Biol 772: 157–178.

Ewels, P., Magnusson, M., Lundin, S., and Käller, M. (2016). MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics 32: 3047–3048.

Feenstra, B., Skovgaard, I.M., and Broman, K.W. (2006). Mapping quantitative trait loci by an extension of the Haley-Knott regression method using estimating equations. Genetics 173: 2269–2282.

Feng, C., Feng, C., and Kang, M. (2016). The first genetic linkage map of *Primulina eburnea* (Gesneriaceae) based on EST-derived SNP markers. J Genet 95: 377–382.

Feng, C., Feng, C., Yang, L., Kang, M., and Rausher, M.D. (2018). Genetic architecture of quantitative flower and leaf traits in a pair of sympatric sister species of *Primulina*. Heredity 11: 36.

Fierst, J.L. (2015). Using linkage maps to correct and scaffold *de novo* genome assemblies: methods, challenges, and computational tools. Front Genet 6: 220.

Fischer, S., Hoffmann, P., Herms, S., and Pfeifer-Sancar, K. (2016). QIAxpert®—a powerful system for nucleic acid quality control. Qiagen Application Note. Qiagen. Hilden. Germany.

Fletcher, J.C. (2002). Shoot and floral meristem maintenance in *Arabidopsis*. Annu Rev Plant Biol. 53: 45–66.

Florea, L.D. and Salzberg, S.L. (2013). Genome-guided transcriptome assembly in the age of next-generation sequencing. IEEE/ACM Trans Comput Biol Bioinform 10: 1234–1240.

Fountain, E.D., Pauli, J.N., Reid, B.N., Palsbøll, P.J., and Peery, M.Z. (2016). Finding the right coverage: the impact of coverage and sequence quality on single nucleotide polymorphism genotyping error rates. Mol Ecol Resour 16: 966–978.

Frary, A., Xu, Y., Liu, J., Mitchell, S., Tedeschi, E., and Tanksley, S. (2005). Development of a set of PCR-based anchor markers encompassing the tomato genome and evaluation of their usefulness for genetics and breeding experiments. Theor Appl Genet 111: 291–312.

Fritsch, K. (1893–1894). 'Gesneriaceae'. in *Die Natürlichen Pflanzenfamilien* (ed. By Engler, A. and Prantl, K.). Wilhelm Engelmann. Leipzig. Germany. pp. 183-375.

Fritsch, K. (1904). Die Keimpflanzen der Gesneriaceen. Nature 70: 453–453.

Fujie, M., Kuroiwa, H., Kawano, S., Mutoh, S., and Kuroiwa, T. (1994). Behavior of organelles and their nucleoids in the shoot apical meristem during leaf development in *Arabidopsis thaliana* L. Planta 194: 395–405.

Gailing, O. (2008). QTL analysis of leaf morphological characters in a *Quercus* robur full-sib family (*Q. robur* x *Q. robur* ssp. slavonica). Plant Biol 10: 624–634.

Garcia-Elias, A., Alloza, L., Puigdecanet, E.X.L., Nonell, L., Tajes, M., Curado, J., Enjuanes, C., az, O.D.X., Bruguera, J., Almor, J.M.X., n-Colet, J.C.X., and Benito, B.X.A. (2017). Defining quantification methods and optimizing protocols for microarray hybridization of circulating microRNAs. Sci Rep 7: 7725.

Genome 10K Community of Scientists (2009). Genome 10K: a proposal to obtain whole-genome sequence for 10,000 vertebrate species. J Hered 100: 659–674.

Ghorbanzade, Z. and Ahmadabadi, M. (2015). Stable transformation of the *Saintpaulia ionantha* by particle bombardment. Iran J Biotechnol 13: 11–16.

Glenn, T.C. (2011). Field guide to next-generation DNA sequencers. Mol Ecol Resour 11: 759–769.

Gonen, S., Lowe, N.R., Cezard, T., Gharbi, K., Bishop, S.C., and Houston, R.D. (2014). Linkage maps of the Atlantic salmon (*Salmo salar*) genome derived from RAD sequencing. BMC Genomics 15: 166.

Gonzales, N.M. and Palmer, A.A. (2014). Fine-mapping QTLs in advanced intercross lines and other outbred populations. Mamm Genome 25: 271–292.

Goodrich, J., Puangsomlee, P., Martin, M., Long, D., Meyerowitz, E.M., and Coupland, G. (1997). A Polycomb-group gene regulates homeotic gene expression in *Arabidopsis*. Nature 386: 44–51.

Goodwin, S., McPherson, J.D., and McCombie, W.R. (2016). Coming of age: ten years of next- generation sequencing technologies. Nat Rev Genet 17: 333–351.

Graham, C.F., Glenn, T.C., Mcarthur, A.G., Boreham, D.R., Kieran, T., Lance, S., Manzon, R.G., Martino, J.A., Pierson, T., Rogers, S.M., Wilson, J.Y., and Somers, C.M. (2015). Impacts of degraded DNA on restriction enzyme associated DNA sequencing (RADSeq). Mol Ecol Resour 15: 1304–1315.

Gualberto, J.M., Mileshina, D., Wallet, C., Niazi, A.K., Weber-Lotfi, F., and Dietrich, A. (2014). The plant mitochondrial genome: dynamics and maintenance. Biochimie 100: 107–120.

Gurevich, A., Saveliev, V., Vyahhi, N., and Tesler, G. (2013). QUAST: quality assessment tool for genome assemblies. Bioinformatics 29: 1072–1075.

Haas, B.J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P.D., Bowden, J., Couger, M.B., Eccles, D., Li, B., Lieber, M., Macmanes, M.D., Ott, M., Orvis, J., Pochet, N., Strozzi, F., Weeks, N., Westerman, R., William, T., Dewey, C.N., Henschel, R., Leduc, R.D., Friedman, and N., Regev, A. (2013). *De novo* transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nat Protoc 8: 1494–1512.

Haberer, G., Young, S., Bharti, A.K., Gundlach, H., Raymond, C., Fuks, G., Butler, E., Wing, R.A., Rounsley, S., Birren, B., Nusbaum, C., Mayer, K.F.X., and Messing, J. (2005). Structure and architecture of the maize genome. Plant Physio 139: 1612–1624.

Hackett, C.A. and Broadfoot, L.B. (2003). Effects of genotyping errors, missing values and segregation distortion in molecular marker data on the construction of linkage maps. Heredity 90: 33–38.

Haldane, J.S. (1919). The combination of linkage values, and the calculation of distances between the loci of linked factors. J Genet: 299–309.

Hanson, L., McMachon, K.A., Johnson, M.A.T., and Bennett, M.D. (2001). First nuclear DNA C-values for another 25 angiosperm families. Ann Bot 88: 851–858.

Harrison, J. (2002). Developmental genetics and evolution of plant form in *Streptocarpus*. PhD thesis. University of Edinburgh. Edinburgh. UK.

Harrison, J., Möller, M., and Cronk, Q.C.B. (1999). Evolution and development of floral diversity in *Streptocarpus* and *Saintpaulia*. Ann Bot 84: 49–60.

Harrison, J., Möller, M., Langdale, J., Cronk, Q., and Hudson, A. (2005). The role of *KNOX* genes in the evolution of morphological novelty in *Streptocarpus*. Plant Cell 17: 430–443.

Hart, M.L., Forrest, L.L., Nicholls, J.A., and Kidner, C.A. (2016). Retrieval of hundreds of nuclear loci from herbarium specimens. Taxon 65: 1081–1092.

Harushima, Y., Yano, M., Shomura, A., Sato, M., Shimano, T., Kuboki, Y., Yamamoto, T., Lin, S.Y., Antonio, B.A., Parco, A., Kajiya, H., Huang, N., Yamamoto, K., Nagamura, Y., Kurata, N., Khush, G.S., and Sasaki, T. (1998). A high-density rice genetic linkage map with 2275 markers using a single F2 population. Genetics 148: 479–494.

Haston, E. and Ronse De Craene, L.P. (2007). Inflorescence and floral development in *Streptocarpus* and *Saintpaulia* (Gesneriaceae) with particular reference to the impact of bracteole suppression. Plant Syst Evol 265: 13–25.

Healey, A., Furtado, A., Cooper, T., and Henry, R.J. (2014). Protocol: a simple method for extracting next-generation sequencing quality genomic DNA from recalcitrant plant species. Plant Methods 10: 21.

Hellsten, U., Wright, K.M., Jenkins, J., Shu, S., Yuan, Y., Wessler, S.R., Schmutz, J., Willis, J.H., and Rokhsar, D.S. (2013). Fine-scale variation in meiotic recombination in *Mimulus* inferred from population shotgun sequencing. Proc Natl Acad Sci USA 110: 19478–19482.

Hilliard, O.M. and Burtt, B.L. (1971). *Streptocarpus*: an African Plant Study. University of Kwazulu Natal Press. Pietermaritzburg. South Africa.

Hirsch, C.N., Hirsch, C.D., Brohammer, A.B., Bowman, M.J., Soifer, I., Barad, O., Shem-Tov, D., Baruch, K., Lu, F., Hernandez, A.G., Fields, C.J., Wright, C.L., Koehler, K., Springer, N.M., Buckler, E., Buell, C.R., de Leon, N., Kaeppler, S.M., Childs, K.L., and Mikel, M.A. (2016). Draft assembly of elite inbred line PH207 provides insights into genomic and transcriptome diversity in maize. Plant Cell 28: 2700–2714.

Hoff, K.J., Lange, S., Lomsadze, A., Borodovsky, M., and Stanke, M. (2016). BRAKER1: unsupervised RNA-seq-based genome annotation with GeneMark-ET and AUGUSTUS. Bioinformatics 32: 767–769.

Hollander, M. and Wolfe, D.A. (1973). Nonparametric Statistical Methods. John Wiley & Sons. New York. USA. pp. 791-797.

Hook, I., Mills, C., and Sheridan, H. (2014). 'Bioactive Naphthoquinones from Higher Plants'. in *Studies in Natural Products Chemistry* (ed. by Atta-ur-Rahman). Elsevier B. V.. Amsterdam. Netherlands. pp. 119-160.

Hsu, H.-C., Chen, C.-Y., Lee, T.-K., Weng, L.-K., Yeh, D.-M., Lin, T.-T., Wang, C.-N., and Kuo, Y.-F. (2015). Quantitative analysis of floral symmetry and tube dilation in an F2 cross of *Sinningia speciosa*. Sci Hortic 188: 71–77.

Hsu, H.-C., Hsu, K.-L., Chan, C.-Y., Wang, C.-N., and Kuo, Y.-F. (2018). Quantifying colour and spot characteristics for the ventral petals in *Sinningia speciosa*. Biosyst Eng 167: 40–50.

Hsu, H.-C., Wang, C.-N., Liang, C.-H., Wang, C.-C., and Kuo, Y.-F. (2017). Association between petal form variation and *CYC2*-like genotype in a hybrid line of *Sinningia speciosa*. Front Plant Sci 8: 5–13.

Hughes, M., Möller, M., Bellstedt, D.U., Edwards, T.J., and VILLIERS, M. (2005). Refugia, dispersal and divergence in a forest archipelago: a study of *Streptocarpus* in eastern South Africa. Mol Ecol 14: 4415–4426.

Hughes, M., Möller, M., Bellstedt, D.U., Edwards, T.J., and Woodhead, M. (2004). EST and random genomic nuclear microsatellite markers for *Streptocarpus*. Mol Ecol Notes 4: 36–38.

Hughes, M., Möller, M., Edwards, T.J., Bellstedt, D.U., and Villiers, M. de (2007). The impact of pollination syndrome and habitat on gene flow: a comparative study of two *Streptocarpus* (Gesneriaceae) species. Am J Bot 94: 1688–1695.

Huijser, P. and Schmid, M. (2011). The control of developmental phase transitions in plants. Development 138: 4117–4129.

Humbert, H. (1971). Gesnèriaceès. In *Flore de Madagascar et des Comores : plantes vasculaires* (ed. Humbert, H. and Jean-François, L.). Museum National d'Histoire Naturelle, Paris. France. fam. 180.

Idury, R.M. and Waterman, M.S. (1995). A new algorithm for DNA sequence assembly. J Comput Biol 2: 291–306.

Imaichi, R., Nagumo, S., and Kato, M. (2000). Ontogenetic anatomy of *Streptocarpus grandis* (Gesneriaceae) with implications for evolution of monophylly. Ann Bot 86: 37–46.

Imaichi, R., Omura-Shimadate, M., Ayano, M., and Kato, M. (2007). Developmental morphology of the caulescent species *Streptocarpus pallidiflorus* (Gesneriaceae), with implications for evolution of monophylly. Int J Plant Sci 168: 251–260.

Inoue, K., Ueda, S., Nayeshiro, H., and Inouye, H. (1982). Structures of unusually prenylated naphthoquinones of *Streptocarpus dunii* and its cell cultures. Chem Pharm Bull 30: 2265–2268.

International Human Genome Sequencing Consortium (2001). Initial sequencing and analysis of the human genome. Nature 409: 860–921.

International Rice Genome Sequencing Project (2005). The map-based sequence of the rice genome. Nature 436: 793–800.

International Wheat Genome Sequencing Consortium (IWGSC) (2014). A chromosome-based draft sequence of the hexaploid bread wheat (*Triticum aestivum*) genome. Science 345: 1251788.

Ivanova, Z., Sablok, G., Daskalova, E., Zahmanova, G., Apostolova, E., Yahubyan, G., and Baev, V. (2017). Chloroplast genome analysis of resurrection tertiary relict *Haberlea rhodopensis* highlights genes important for desiccation stress response. Front Plant Sci 8: 69–16.

Jackman, S.D., Vandervalk, B.P., Mohamadi, H., Chu, J., Yeo, S., Hammond, S.A., Jahesh, G., Khan, H., Coombe, L., Warren, R.L., and Birol, I. (2017). ABySS 2.0: resource-efficient assembly of large genomes using a Bloom filter. Genome Res 27: 768–777.

Jackson, B.C. (2011). Recombination-suppression: how many mechanisms for chromosomal speciation? Genetica 139: 393–402.

Jain, M., Koren, S., Miga, K.H., Quick, J., Rand, A.C., Sasani, T.A., Tyson, J.R., Beggs, A.D., Dilthey, A.T., Fiddes, I.T., Malla, S., Marriott, H., Nieto, T., O'Grady, J., Olsen, H.E., Pedersen, B.S., Rhie, A., Richardson, H., Quinlan, A.R., Snutch, T.P., Tee, L., Paten, B., Phillippy, A.M., Simpson, J.T., Loman, N.J., and Loose, M. (2018). Nanopore sequencing and assembly of a human genome with ultra-long reads. Nature Biotechnol 36: 338–345.

Jansen, R.K., Kaittanis, C., Saski, C., Lee, S.-B., Tomkins, J., Alverson, A.J., and Daniell, H. (2006). Phylogenetic analyses of *Vitis* (Vitaceae) based on complete chloroplast genome sequences: effects of taxon sampling and phylogenetic methods on resolving relationships among rosids. BMC Evol Biol 6: 32.

Jiang, D., Kong, N.C., Gu, X., Li, Z., and He, Y. (2011). *Arabidopsis* COMPASS-like complexes mediate histone H3 lysine-4 trimethylation to control floral transition and plant development. PLoS Genet 7: e1001330.

Jiao, W.-B. and Schneeberger, K. (2017). The impact of third generation genomic technologies on plant genome assembly. Curr Opin Plant Biol 36: 64–70.

Jiao, W.-B., Accinelli, G.G., Hartwig, B., Kiefer, C., Baker, D., Severing, E., Willing, E.-M., Piednoel, M., Woetzel, S., Madrid-Herrero, E., Huettel, B., Hümann, U., Reinhard, R., Koch, M.A., Swan, D., Clavijo, B., Coupland, G., and Schneeberger, K. (2017). Improving and correcting the contiguity of long-read genome assemblies of three plant species using optical mapping and chromosome conformation capture data. Genome Res 27: 778–786.

Jong, K. (1970). Developmental aspects of vegetative morphology in *Streptocarpus*. PhD thesis. University of Edinburgh. Edinburgh. UK.

Jong, K. (1978). Phyllomorphic organisation in rosulate *Streptocarpus*. Notes R Bot Gard Edinb 36: 369–396.

Jong, K. and Burtt, B. (1975). The evolution of morphological novelty exemplified in the growth patterns of some Gesneriaceae. New Phytol 75: 297–311.

Jong, K. and Möller, M. (2000). New chromosome counts in *Streptocarpus* (Gesneriaceae) from Madagascar and the Comoro Islands and their taxonomic significance. Plant Syst Evol 224: 173–182.

Jong, K., Christie, F., Paik, J.H., Scott, S.M., and Möller, M. (2012). Unusual morphological and anatomical features of two woody Madagascan endemics, *Streptocarpus papangae* and *S. suffruticosus* (Gesneriaceae), and their potential taxonomic value. S Afr J Bot 80: 44–56.

Kahlau, S., Aspinall, S., Gray, J.C., and Bock, R. (2006). Sequence of the tomato chloroplast DNA and evolutionary comparison of Solanaceous plastid genomes. J Mol Evol 63: 194–207.

Kakioka, R., Kokita, T., Kumada, H., Watanabe, K., and Okuda, N. (2013). A RAD-based linkage map and comparative genomics in the gudgeons (genus *Gnathopogon*, Cyprinidae). BMC Genomics 14: 32.

Kalvari, I., Argasinska, J., Quinones-Olvera, N., Nawrocki, E.P., Rivas, E., Eddy, S.R., Bateman, A., Finn, R.D., and Petrov, A.I. (2018). Rfam 13.0: shifting to a genome-centric resource for non-coding RNA families. Nucleic Acids Res 46: D335–D342.

Kanehisa, M., Sato, Y., and Morishima, K. (2016a). BlastKOALA and GhostKOALA: KEGG tools for functional characterization of genome and metagenome sequences. J Mol Biol 428: 726–731.

Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M., and Tanabe, M. (2016b). KEGG as a reference resource for gene and protein annotation. Nucleic Acids Res 44: D457–D462.

Kay, K.M., Whittall, J.B., and Hodges, S.A. (2006) A survey of nuclear ribosomal internal transcribed spacer substitution rates across angiosperms: an approximate molecular clock with life history effects. BMC Evol Biol 6:36.

Keats, J.J., Cuyugan, L., Adkins, J., and Liang, W.S. (2018). Whole genome library construction for next generation sequencing. Methods Mol Biol 1706: 151-161.

Kent, W.J. (2002). BLAT - the BLAST-like alignment tool. Genome Res 12: 656–664.

Kiehn, M., Hellmayr, E., and Weber, A. (1998). Chromosome numbers of Malayan and other paleotropical Gesneriaceae: 1 Tribe Didymocarpeae. Beitrage zur Biologie der Pflanzen 70: 407-444.

Kim, C.S., Lee, C.H., Shin, J.S., Chung, Y.S., and Hyung, N.I. (1997). A simple and rapid method for isolation of high quality genomic DNA from fruit trees and conifers using PVP. Nucleic Acids Res 25: 1085–1086.

Kinoshita, A. and Tsukaya, H. (2018). One-leaf plants in the Gesneriaceae: natural mutants of the typical shoot system. Develop Growth Differ 118: 99–109.

Klahre, U., Gurba, A., Hermann, K., Saxenhofer, M., Bossolini, E., Guerin, P.M., and Kuhlemeier, C. (2011). Pollinator choice in *Petunia* depends on two major genetic loci for floral scent production. Curr Biol 21: 730–739.

Kole, C. and Abbott, A.G. (2008). ‘Fundamentals of plant genome mapping’. in *Principles and Practices of Plant Genomics Vol 1 Genome Mapping* (ed. by Kole, C. and Abbott, A.G.). CRC Press. Florida, USA.

Koornneef, M., Van Eden, J., of, C.H.J., 1983 (1983). Linkage map of *Arabidopsis thaliana*. J Hered 74: 265–272.

Korpelainen, E., Tuimala, J., Somervuo, P., Huss, M., and Wong, G. (2014). RNA-seq Data Analysis. CRC Press. Florida. USA.

Koutsovoulos, G., Kumar, S., Laetsch, D.R., Stevens, L., Daub, J., Conlon, C., Maroon, H., Thomas, F., Aboobaker, A.A., and Blaxter, M. (2016). No evidence for extensive horizontal gene transfer in the genome of the tardigrade Hypsibius dujardini. Proc Natl Acad Sci USA 113: 5053–5058.

Križman, M., Jakše, J., Baričevič, D., Agriculturae, B.J.A., 2006 (2006). Robust CTAB-activated charcoal protocol for plant DNA extraction. Acta agriculturae Slovenica 87: 427–433.

Kruglyak, L. and Lander, E.S. (1995). A nonparametric approach for mapping quantitative trait loci. Genetics 139: 1421–1428.

Kulski, J.K. (2016). 'Next-Generation Sequencing - An Overview of the History, Tools, and "Omic" Applications'. in *Next Generation Sequencing - Advances, Applications and Challenges* (ed. by J. Kulski). IntechOpen. London. UK. pp. 3–60.

Kumar, S., Jones, M., Koutsovoulos, G., Clarke, M., and Blaxter, M. (2013). Blobology: exploring raw genome data for contaminants, symbionts and parasites using taxon-annotated GC-coverage plots. Front Genet 4: 237.

Kuroha, T., Tokunaga, H., Kojima, M., Ueda, N., Ishida, T., Nagawa, S., Fukuda, H., Sugimoto, K., and Sakakibara, H. (2009). Functional analyses of *LONELY GUY* cytokinin-activating enzymes reveal the importance of the direct activation pathway in *Arabidopsis*. Plant Cell 21: 3152–3169.

Kushikawa, S., Hoshino, Y., and Mii, M. (2001). *Agrobacterium*-mediated transformation of *Saintpaulia ionantha* Wendl. Plant Sci 161: 953–960.

Kyalo, C.M., Gichira, A.W., Li, Z.-Z., Saina, J.K., Malombe, I., Hu, G.-W., and Wang, Q.-F. (2018). Characterization and comparative analysis of the complete chloroplast genome of the critically endangered species *Streptocarpus teitensis* (Gesneriaceae). Biomed Res Int 2018: 1–11.

Laetsch, D.R. and Blaxter, M.L. (2017). BlobTools: interrogation of genome assemblies. F1000Res 6: 1287.

Lanceras, J.C., Pantuwan, G., Jongdee, B., and Toojinda, T. (2004). Quantitative trait loci associated with drought tolerance at reproductive stage in rice. Plant Physio 135: 384–399.

Lander, E.S. and Botstein, D. (1989). Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. Genetics 121: 185–199.

Langmead, B. and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat Meth 9: 357–359.

Lapitanz, N.L.V. (1992). Organization and evolution of higher plant nuclear genomes. Genome 35: 171–181.

Lassmann, T., Hayashizaki, Y., and Daub, C.O. (2011). SAMStat: monitoring biases in next generation sequencing data. Bioinformatics 27: 130–131.

Latvala-Kilby, S.M.H. and Kilby, N.J. (2006). Uncovering the post-embryonic role of embryo essential genes in *Arabidopsis* using the controlled induction of visibly marked genetic mosaics: EMB506, an illustration. Plant Mol Biol 61: 179–194.

Laurence, M., Hatzis, C., and Brash, D.E. (2014). Common contaminants in next-generation sequencing that hinder discovery of low-abundance microbes. PLoS ONE 9: e97876–8.

Laux, T., Mayer, K.F., Berger, J., and Jürgens, G. (1996). The *WUSCHEL* gene is required for shoot and floral meristem integrity in *Arabidopsis*. Development 122: 87–96.

Lawrence, W.J.C. (1940). The Genus *Streptocarpus*. J R Hortic Soc 65: 17–26.

Lawrence, W.J.C. (1947). Studies on *Streptocarpus* II. Complementary sublethal genes. J Genet 48: 16–30.

Lawrence, W.J.C. (1957). Studies on *Streptocarpus* IV. Genetics of flower colour patterns. Heredity 11: 337–357.

Lawrence, W.J.C. (1958). Studies on *Streptocarpus* Lindl. V. Speciation and gene systems. Heredity 12: 333–356.

Lawrence, W.J.C. and Sturgess, V.C. (1957). Studies on *Streptocarpus* III Genetics and chemistry of flower colour in the garden forms. Heredity 11: 303–336.

Lawrence, W.J.C., Scott-Moncrieff, R., and Sturgess, V.C. (1939). Studies on *Streptocarpus* I. Genetics and chemistry of flower colour in the garden strains. J Genet 38: 299–306.

Leggett, R.M. and Maclean, D. (2014). Reference-free SNP detection: dealing with the data deluge. BMC Genomics 15: S10.

Lenhard, M., Jürgens, G., and Laux, T. (2002). The *WUSCHEL* and *SHOOTMERISTEMLESS* genes fulfil complementary roles in *Arabidopsis* shoot meristem regulation. Development 129: 3195–3206.

Leonard, A.S. and Papaj, D.R. (2011). "X" marks the spot: The possible benefits of nectar guides to bees and plants. Funct Ecol 25: 1293–1301.

Lévesque, C.A., Brouwer, H., Cano, L., Hamilton, J.P., Holt, C., Huitema, E., Raffaele, S., Robideau, G.P., Thines, M., Win, J., Zerillo, M.M., Beakes, G.W., Boore, J.L., Busam, D., Dumas, B., Ferriera, S., Fuerstenberg, S.I., Gachon, C.M.M., Gaulin, E., Govers, F., Grenville-Briggs, L., Horner, N., Hostetler, J., Jiang, R.H.Y., Johnson, J., Krajaejun, T., Lin, H., Meijer, H.J.G., Moore, B., Morris, P., Phuntmart, V., Puiu, D., Shetty, J., Stajich, J.E., Tripathy, S., Wawra, S., van West, P., Whitty, B.R., Coutinho, P.M., Henrissat, B., Martin, F., Thomas, P.D., Tyler, B.M., De Vries, R.P., Kamoun, S., Yandell, M., Tisserat, N., and Buell, C.R. (2010). Genome sequence of the necrotrophic plant pathogen *Pythium ultimum* reveals original pathogenicity mechanisms and effector repertoire. Genome Biol 11: R73.

Li, B.J., Li, H.L., Shi, Y.X., and Xie, X.W. (2014). First report of *Pseudomonas cichorii* causing leaf spot of vegetable sponge gourd in China. Plant Dis 98: 153.

Li, F.-W. and Harkess, A. (2018). A guide to sequence your favorite plant genomes. Appl Plant Sci 6: e1030–7.

Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics 2018: 1–7.

Li, H. and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25: 1754–1760.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics 25: 2078–2079.

Li, H., Hearne, S., nziger, M.B.A., Li, Z., and Wang, J. (2010). Statistical properties of QTL linkage mapping in biparental genetic populations. Heredity 105: 257–267.

Li, L. and Kwan, H.S. (2014). A novel computational approach for genome-wide prediction of small RNAs in bacteria. bioRxiv. doi: https://doi.org/10.1101/011668.

Li, S.-B., Xie, Z.-Z., Hu, C.-G., and Zhang, J.-Z. (2016). A review of auxin response factors (ARFs) in plants. Front Plant Sci 7: 137–144.

Li, X., Quigg, R.J., Zhou, J., Xu, S., Masinde, G., Mohan, S., and Baylink, D.J. (2006). A critical evaluation of the effect of population size and phenotypic measurement on QTL detection and localization using a large F2 murine mapping population. Genet Mol Biol 29: 166–173.

Li, Z., Chen, Y., Mu, D., Yuan, J., Shi, Y., Zhang, H., Gan, J., Li, N., Hu, X., Liu, B., Yang, B., and Fan, W. (2012). Comparison of the two major classes of assembly algorithms: overlap-layout-consensus and de-bruijn-graph. Brief Funct Genomics 11: 25–37.

Lincoln, S.P., Fermor, T.R., and Tindall, B.J. (1999). *Janthinobacterium agaricidamnosum* sp. nov., a soft rot pathogen of *Agaricus bisporus*. Int J Syst Bacteriol 49: 1577–1589.

Lindgreen, S., Umu, S.U., Lai, A.S.-W., Eldai, H., Liu, W., McGimpsey, S., Wheeler, N.E., Biggs, P.J., Thomson, N.R., Barquist, L., Poole, A.M., and Gardner, P.P. (2014). Robust identification of noncoding RNA from transcriptomes requires phylogenetically-informed sampling. PLoS Comput Biol 10: e1003907.

Lindley, J. (1828). *Streptocarpus rexii*. Cape *Streptocarpus*. Botanical Register: pl. 1173.

Liscum, E. and Reed, J.W. (2002). Genetics of Aux/IAA and ARF action in plant growth and development. Plant Mol Biol 49: 387–400.

Lohse, M., Drechsel, O., and Bock, R. (2007). OrganellarGenomeDRAW (OGDRAW): a tool for the easy generation of high-quality custom graphical maps of plastid and mitochondrial genomes. Curr Genet 52: 267–274.

Lohse, M., Drechsel, O., Kahlau, S., and Bock, R. (2013). OrganellarGenomeDRAW—a suite of tools for generating physical maps of plastid and mitochondrial genomes and visualizing expression data sets. Nucleic Acids Res 41: W575–W581.

Lohse, M., Nagel, A., Herter, T., May, P., Schroda, M., Zrenner, R., Tohge, T., Fernie, A.R., Stitt, M., and Usadel, B. (2013). Mercator: a fast and simple web server for genome scale functional annotation of plant sequence data. Plant Cell Environ 37: 1250–1258.

Loose, M.W. (2017). The potential impact of nanopore sequencing on human genetics. Hum Mol Genet 26: R202–R207.

Loureiro, J., Rodriguez, E., Dolezel, J., and Santos, C. (2007). Two new nuclear isolation buffers for plant DNA flow cytometry: a test with 37 species. Ann Bot 100: 875–888.

Lowe, T. (1997). tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res 25: 955–964.

Lowe, T.M. and Chan, P.P. (2016). tRNAscan-SE On-line: integrating search and context for analysis of transfer RNA genes. Nucleic Acids Res 44: W54–W57.

Lowry, D.B., Hoban, S., Kelley, J.L., Lotterhos, K.E., Reed, L.K., Antolin, M.F., and Storfer, A. (2016). Breaking RAD: An evaluation of the utility of restriction site associated DNA sequencing for genome scans of adaptation. Mol Ecol Resour 17: 142–152.

Lowry, D.B., Hoban, S., Kelley, J.L., Lotterhos, K.E., Reed, L.K., Antolin, M.F., and Storfer, A. (2017). Responsible RAD: Striving for best practices in population genomic studies of adaptation. Mol Ecol Resour 17: 366–369.

Lu, B., Zeng, Z., and Shi, T. (2013). Comparative study of *de novo* assembly and genome-guided assembly strategies for transcriptome reconstruction based on RNA-Seq. Sci China Life Sci 56: 143–155.

Lunter, G. and Goodson, M. (2011). Stampy: A statistical algorithm for sensitive and fast mapping of Illumina sequence reads. Genome Res 21: 936–939.

Luo, R., Liu, B., Xie, Y., Li, Z., Huang, W., Yuan, J., He, G., Chen, Y., Pan, Q., Liu, Y., Tang, J., Wu, G., Zhang, H., Shi, Y., Liu, Y., Yu, C., Wang, B., Lu, Y., Han, C., Cheung, D.W., Yiu, S.-M., Peng, S., Xiaoqian, Z., Liu, G., Liao, X., Li, Y., Yang, H., Wang, J., Lam, T.-W., and Wang, J. (2012). SOAPdenovo2: an empirically improved memory-efficient short-read *de novo* assembler. Gigascience 1: 18.

Lynch, M. and Walsh, B. (1998). Genetics and Analysis of Quantitative Traits. Sinauer Associates Incorporated. Oxford University Press. Oxford. UK.

Manichaikul, A., Dupuis, J., Sen, Ś., and Broman, K.W. (2006). Poor performance of bootstrap confidence intervals for the location of a quantitative trait locus. Genetics 174: 481–489.

Mantegazza, R., Möller, M., Harrison, J., Fior, S., De Luca, C., and Spada, A. (2007). Anisocotyly and meristem initiation in an unorthodox plant, *Streptocarpus rexii* (Gesneriaceae). Planta 225: 653–663.

Mantegazza, R., Tononi, P., Möller, M., and Spada, A. (2009). *WUS* and *STM* homologs are linked to the expression of lateral dominance in the acaulescent *Streptocarpus rexii* (Gesneriaceae). Planta 230: 529–542.

Mapleson, D., Garcia Accinelli, G., Kettleborough, G., Wright, J., and Clavijo, B.J. (2017). KAT: a K-mer analysis toolkit to quality control NGS datasets and genome assemblies. Bioinformatics 33: 574–576.

Marçais, G. and Kingsford, C. (2011). A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics 27: 764–770.

Marçais, G., Delcher, A.L., Phillippy, A.M., Coston, R., Salzberg, S.L., and Zimin, A. (2018). MUMmer4: A fast and versatile genome alignment system. PLoS Comput Biol 14: e1005944–14.

Maria, C., Stana, D., and Pop, I. (2004). *Streptocarpus* - flowering pot plant - propagation and culture. Not Bot Hort Agrobot Cluj 32: 15–19.

Martin, R.C., Mok, M.C., Habben, J.E., and Mok, D.W. (2001). A maize cytokinin gene encoding an O-glucosyltransferase specific to *cis*-zeatin. Proc Natl Acad Sci USA 98: 5922–5926.

Masson-Boivin, C., Giraud, E., Perret, X., and Batut, J. (2009). Establishing nitrogen-fixing symbiosis with legumes: how many rhizobium recipes? Trends Microbiol 17: 458–466.

Matasci, N., Hung, L.-H., Yan, Z., Carpenter, E.J., Wickett, N.J., Mirarab, S., Nguyen, N., Warnow, T., Ayyampalayam, S., Barker, M., Burleigh, J.G., Gitzendanner, M.A., Wafula, E., Der, J.P., dePamphilis, C.W., Roure, B., Philippe, H., Ruhfel, B.R., Miles, N.W., Graham, S.W., Mathews, S., Surek, B., Melkonian, M., Soltis, D.E., Soltis, P.S., Rothfels, C., Pokorny, L., Shaw, J.A., DeGironimo, L., Stevenson, D.W., Villarreal, J.C., Chen, T., Kutchan, T.M., Rolf, M., Baucom, R.S., Deyholos, M.K., Samudrala, R., Tian, Z., Wu, X., Sun, X., Zhang, Y., Wang, J., Leebens-Mack, J., and Wong, G.K.-S. (2014). Data access for the 1,000 Plants (1KP) project. Gigascience 3: 17.

Matsubara, K., Yamamoto, E., Kobayashi, N., Ishii, T., Tanaka, J., Tsunematsu, H., Yoshinaga, S., Matsumura, O., Yonemaru, J.-I., Mizobuchi, R., Yamamoto, T., Kato, H., and Yano, M. (2016). Improvement of rice biomass yield through QTL-based selection. PLoS ONE 11: e0151830.

Mauricio, R. (2001). Mapping quantitative trait loci in plants: uses and caveats for evolutionary biology. Nat Rev Genet 2: 370–381.

McKinney, G.J., Larson, W.A., Seeb, L.W., and Seeb, J.E. (2017). RADseq provides unprecedented insights into molecular ecology and evolutionary genetics: comment on Breaking RAD by Lowry et al. (2016). Mol Ecol Resour 17: 356–361.

McPherson, H., van der Merwe, M., Delaney, S.K., Edwards, M.A., Henry, R.J., McIntosh, E., Rymer, P.D., Milner, M.L., Siow, J., and Rossetto, M. (2013). Capturing chloroplast variation for molecular ecology studies: a simple next generation sequencing approach applied to a rainforest tree. BMC Ecol 13: 8.

Meinke, D.W., Meinke, L.K., Showalter, T.C., Schissel, A.M., Mueller, L.A., and Tzafrir, I. (2003). A sequence-based map of *Arabidopsis* genes with mutant phenotypes. Plant Physio 131: 409–418.

Merchant, S., Wood, D.E., and Salzberg, S.L. (2014). Unexpected cross-species contamination in genome sequencing projects. PeerJ 2: e675.

Merchuk-Ovnat, L., Barak, V., Fahima, T., Ordon, F., Lidzbarsky, G.A., Krugman, T., and Saranga, Y. (2016). Ancestral QTL alleles from wild emmer wheat improve drought resistance and productivity in modern wheat cultivars. Front Plant Sci 7: 452.

Mercuri, A., De Benedetti, L., Burchi, G., and Schiva, T. (2000). *Agrobacterium*-mediated transformation of African violet. Plant Cell Tissue Organ Cult 60: 39–46.

Michael, T.P., Jupe, F., Bemm, F., Motley, S.T., Sandoval, J.P., Lanz, C., Loudet, O., Weigel, D., and Ecker, J.R. (2018). High contiguity *Arabidopsis thaliana* genome assembly with a single nanopore flow cell. Nat Commun 9: 541.

Miller, M.R., Dunham, J.P., Amores, A., Cresko, W.A., and Johnson, E.A. (2007). Rapid and cost-effective polymorphism identification and genotyping using restriction site associated DNA (RAD) markers. Genome Res 17: 240–248.

Möller, M. (2018). Nuclear DNA C-values are correlated with pollen size at tetraploid but not diploid level and linked to phylogenetic descent in *Streptocarpus* (Gesneriaceae). S Afr J Bot 114: 323–344.

Möller, M. and Clark, J.L. (2013). The state of molecular studies in the family Gesneriaceae: a review. Selbyana 31: 95–125.

Möller, M. and Cronk, Q.C. (2001). Evolution of morphological novelty: a phylogenetic analysis of growth patterns in *Streptocarpus* (Gesneriaceae). Evolution 55: 918–929.

Möller, M. and Cronk, Q.C.B. (1997). Origin and relationships of *Saintpaulia* (Gesneriaceae) based on ribosomal DNA internal transcribed spacer (ITS) sequences. Am J Bot 84: 956–965.

Möller, M. and Pullan, M. (2015). RBGE WebCyte2 – An updated Gesneriaceae cytology database. Available at: http://elmer.rbge.org.uk/webcyte/webcyteintro.php.

Möller, M., Barber, S., Atkins, H. J. and Purvis, D. A. (2019) The living collection at Royal Botanic Garden Edinburgh illustrates the floral diversity in *Streptocarpus* (Gesneriaceae). Sibbaldia 17: 155.

Möller, M., Brooks, K.J., and Hughes, M. (2004). Plastic inheritance in *Streptocarpus* (Gesneriaceae) and an inferred hybrid origin for a population of *S.* aff. *primulifolius* from Igoda river, South Africa. Edinb J Bot 60: 389–408.

Möller, M., Pfosser, M., Jang, C.-G., Mayer, V., Clark, A., Hollingsworth, M.L., Barfuss, M.H.J., Wang, Y.-Z., Kiehn, M., and Weber, A. (2009). A preliminary phylogeny of the "didymocarpoid Gesneriaceae" based on three molecular data sets: incongruence with available tribal classifications. Am J Bot 96: 989–1010.

Mower, J.P., Case, A.L., Floro, E.R., and Willis, J.H. (2012). Evidence against equimolarity of large repeat arrangements and a predominant master circle structure of the mitochondrial genome from a monkeyflower (*Mimulus guttatus*) lineage with cryptic CMS. Genome Biol Evol 4: 670–686.

Mueller, L.A., Solow, T.H., Taylor, N., Skwarecki, B., Buels, R., Binns, J., Lin, C., Wright, M.H., Ahrens, R., Wang, Y., Herbst, E.V., Keyder, E.R., Menda, N., Zamir, D., and Tanksley, S.D. (2005). The SOL Genomics Network: a comparative resource for Solanaceae biology and beyond. Plant Physio 138: 1310–1317.

Nadiya, F., Anjali, N., Gangaprasad, A., and Sabu, K.K. (2015). High quality RNA extraction from small cardamom tissues rich in polysaccharides and polyphenols. Anal Biochem 485: 25–27.

Nawrocki, E.P. and Eddy, S.R. (2013). Infernal 1.1: 100-fold faster RNA homology searches. Bioinformatics 29: 2933–2935.

Ng, D.W.-K., Wang, T., Chandrasekharan, M.B., Aramayo, R., Kertbundit, S., and Hall, T.C. (2007). Plant SET domain-containing proteins: structure, function and regulation. Biochim Biophys Acta Gene Struct Expr 1769: 316–329.

Nielsen, R., Paul, J.S., Albrechtsen, A., and Song, Y.S. (2011). Genotype and SNP calling from next-generation sequencing data. Nat Rev Genet 12: 443–451.

Nishii, K. and Nagata, T. (2007). Developmental analyses of the phyllomorph formation in the rosulate species *Streptocarpus rexii* (Gesneriaceae). Plant Syst Evol 265: 135–145.

Nishii, K., Ho, M.-J., Chou, Y.-W., Gabotti, D., Wang, C.-N., Spada, A., and Möller, M. (2014). *GA2* and *GA20-oxidase* expressions are associated with the meristem position in *Streptocarpus rexii* (Gesneriaceae). Plant Growth Regul 72: 123–140.

Nishii, K., Huang, B.-H., Wang, C.-N., and Möller, M. (2017). From shoot to leaf: step-wise shifts in meristem and *KNOX1* activity correlate with the evolution of a unifoliate body plan in Gesneriaceae. Dev Genes Evol 227: 41–60.

Nishii, K., Hughes, M., Briggs, M., Haston, E., Christie, F., DeVilliers, M.J., Hanekom, T., Roos, W.G., Bellstedt, D.U., and Möller, M. (2015). *Streptocarpus* redefined to include all Afro-Malagasy Gesneriaceae: Molecular phylogenies prove congruent with geographical distribution and basic chromosome numbers and uncover remarkable morphological homoplasies. Taxon 64: 1243–1274.

Nishii, K., Kuwabara, A., and Nagata, T. (2004). Characterization of anisocotylous leaf formation in *Streptocarpus wendlandii* (Gesneriaceae): significance of plant growth regulators. Ann Bot 94: 457–467.

Nishii, K., Möller, M., Kidner, C., Spada, A., Mantegazza, R., Wang, C.-N., and Nagata, T. (2010a). A complex case of simple leaves: indeterminate leaves co-express *ARP* and *KNOX1* genes. Dev Genes Evol 220: 25–40.

Nishii, K., Nagata, T., and Wang, C.-N. (2010b). High morphological plasticity in Gesneriaceae meristems: Reversions in vegetative and floral development. Trends Dev Biol 4: 33–40.

Nishii, K., Wang, C.-N., Spada, A., Nagata, T., and Möller, M. (2012a). Gibberellin as a suppressor of lateral dominance and inducer of apical growth in the unifoliate *Streptocarpus wendlandii* (Gesneriaceae). New Zeal J Bot 50: 1–22.

Nishii, K., Nagata, T., Wang, C.-N., and Möller, M. (2012b). Light as environmental regulator for germination and macrocotyledon development in *Streptocarpus rexii* (Gesneriaceae). S Afr J Bot 81: 50–60.

Nitsch, J.P. (1967). *Streptocarpus nobilis*, plante de jours courts. Bulletin de la Société Botanique de France 114: 128–132.

Numa, H. and Itoh, T. (2014). MEGANTE: a web-based system for integrated plant genome annotation. Plant Cell Physiol 55: e2.

Oehlkers, F. (1938). Bastardierungsversuche in der gattung *Streptocarpus* Lindl. 1. Plasmatische vererbung und die geschlechtbestim- mung von zwitterpflanzen. Zeitschrift f Botanik 32: 305–393.

Oehlkers, F. (1942). Faktorenanalytische Ergebnisse an Artbastarden. Biol Zbl 62: 280–289.

Oehlkers, F. (1956). Veränderungen in der Blühbereitschaft vernalisierter Cotyledonen von *Streptocarpus*, kenntlich gemacht durch Blattstecklinge. Z Naturf 11B: 471–480.

Oehlkers, F. (1966). Der gelbe fleck in der blüte Der Gesneriaceae *Streptocarpus* Lindl. und seine vererbung. I Z Vererbungsl 98: 127–136.

Oehlkers, F. (1967). Der gelbe fleck in der blüte Der Gesneriaceae *Streptocarpus* Lindl. und seine vererbung.II. Molec Gen Genetics 99: 62–68.

O'Neill, M., McPartlin, J., Arthure, K., Riedel, S., and McMillan, N.D. (2011). Comparison of the TLDA with the Nanodrop and the reference Qubit system. J Phys Conf Ser 307: 012047.

Ouyang, S., Zhu, W., Hamilton, J., Lin, H., Campbell, M., Childs, K., Thibaud-Nissen, F., Malek, R.L., Lee, Y., Zheng, L., Orvis, J., Haas, B., Wortman, J., and Buell, C.R. (2007). The TIGR Rice Genome Annotation Resource: improvements and new features. Nucleic Acids Res 35: D883–D887.

Overvoorde, P.J., Okushima, Y., Alonso, J.M., Chan, A., Chang, C., Ecker, J.R., Hughes, B., Liu, A., Onodera, C., Quach, H., Smith, A., Yu, G., and Theologis, A. (2005). Functional genomic analysis of the *AUXIN/INDOLE-3-ACETIC ACID* gene family members in *Arabidopsis thaliana*. Plant Cell 17: 3282–3300.

Palaiokostas, C., Bekaert, M., Khan, M.G.Q., Taggart, J.B., Gharbi, K., Mcandrew, B.J., and Penman, D.J. (2013). Mapping and validation of the major sex-determining region in Nile Tilapia (*Oreochromis niloticus* L.) using RAD sequencing. PLoS ONE 8: e68389.

Parra, G., Bradnam, K., and Korf, I. (2007). CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. Bioinformatics 23: 1061–1067.

Peleg, Z., Fahima, T., Krugman, T., Abbo, S., Yakir, D., Korol, A.B., and Saranga, Y. (2009). Genomic dissection of drought resistance in durum wheat × wild emmer wheat recombinant inbreed line population. Plant Cell Environ 32: 758–779.

Peleg, Z., Saranga, Y., Suprunova, T., Ronin, Y., Röder, M.S., Kilian, A., Korol, A.B., and Fahima, T. (2008). High-density genetic map of durum wheat × wild emmer wheat based on SSR and DArT markers. Theor Appl Genet 117: 103–115.

Pellicer, J., Fay, M.F., and Leitch, I.J. (2010). The largest eukaryotic genome of them all? Bot J Linnean Soc 164: 10–15.

Peterson, B.K., Weber, J.N., Kay, E.H., Fisher, H.S., and Hoekstra, H.E. (2012). Double Digest RADseq: an inexpensive method for *de novo* SNP discovery and genotyping in model and non-model species. PLoS ONE **7**: e37135.

Platt, A.R., Woodhall, R.W., and George, A.L., Jr (2007). Improved DNA sequencing quality and efficiency using an optimized fast cycle sequencing protocol. Biotechniques 43: 58–62.

Puglisi, C., Yao, T.L., Milne, R., Möller, M., and Middleton, D.J. (2016). Generic recircumscription in the Loxocarpinae (Gesneriaceae), as inferred by phylogenetic and morphological data. Taxon 65: 277–292.

Puglisi, C.,Wei, Y.G., Nishii, K., and Möller, M. (2011) *Oreocharis* × *heterandra* (Gesneriaceae): a natural hybrid from the Shengtangshan Mountains, Guangxi, China. Phytotaxa 38: 1–18.

Pyke, K.A. and Leech, R.M. (1994). A genetic analysis of chloroplast division and expansion in *Arabidopsis thaliana*. Plant Physio 104: 201–207.

Quinlan, A.R. and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26: 841–842.

R Development Core Team (2008). R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria. Available at: http://www.R-project.org/.

Raghavan, C. and Collard, B.C.Y. (2012). Effect of small mapping population sizes on reliability of quantitative trait locus (QTL) mapping. Afr J Biotechnol 11: 10661–10674.

Ram, M.S.N. and Mohandas, S. (2003). Transformation of African violet (*Saintpaulia ionantha*) with glucanase-chitinase genes using *Agrobacterium* tumefaciens. Acta Hortic 624: 471–478.

Rascoe, J., Berg, M., Melcher, U., Mitchell, F.L., Bruton, B.D., Pair, S.D., and Fletcher, J. (2003). Identification, phylogenetic analysis, and biological characterization of *Serratia marcescens* strains causing cucurbit yellow vine disease. Phytopathology 93: 1233–1239.

Ratter, J.A. (1963). Some chromosome numbers in the Gesneriaceae. Notes R Bot Gard Edinb 24: 221–229.

Rauh, R.A. (2001). The role of suppression of the morphogenesis and phylogeny of *Streptocarpus prolixus* (Gesneriaceae). PhD thesis. City University of New York. New York. USA.

Rauh, R.A. and Basile, D.V. (2003). Phenovariation induced in *Streptocarpus prolixus* (Gesneriaceae) by β-glucosyl Yariv reagent. Can J Bot 81: 338–344.

Reinten, E.Y., Coetzee, J.H., and van Wyk, B.E. (2011). The potential of South African indigenous plants for the international cut flower trade. S Afr J Bot 77: 934–946.

Ren, T., Zheng, W., Han, K., Zeng, S., Zhao, J., and Liu, Z.-L. (2016). Characterization of the complete chloroplast genome sequence of *Lysionotus pauciflorus* (Gesneriaceae). Conserv Genet Resour 9: 185–187.

Ren, X., Li, R., Wei, X., Bi, Y., Ho, V.W.S., Ding, Q., Xu, Z., Zhang, Z., Hsieh, C.-L., Young, A., Zeng, J., Liu, X., and Zhao, Z. (2018). Genomic basis of recombination suppression in the hybrid between *Caenorhabditis briggsae* and *C. nigoni*. Nucleic Acids Res 46: 1295–1307.

Reuter, J.A., Spacek, D.V., and Snyder, M.P. (2015). High-throughput sequencing technologies. Mol Cell 58: 586–597.

Rhee, S.Y., Beavis, W., Berardini, T.Z., Chen, G., Dixon, D., Doyle, A., Garcia-Hernandez, M., Huala, E., Lander, G., Montoya, M., Miller, N., Mueller, L.A., Mundodi, S., Reiser, L., Tacklind, J., Weems, D.C., Wu, Y., Xu, I., Yoo, D., Yoon, J., and Zhang, P. (2003). The *Arabidopsis* Information Resource (TAIR): a model organism database providing a centralized, curated gateway to *Arabidopsis* biology, research materials and community. Nucleic Acids Res 31: 224–228.

Rieseberg, L.H., Archer, M.A., and Wayne, R.K. (1999). Transgressive segregation, adaptation and speciation. Heredity 83: 363–372.

Ristaino, J.B., Johnson, A., Plant, M.B.-M., 2007 (2007). Identification of the tobacco blue mold pathogen, *Peronospora tabacina*, by polymerase chain reaction. Plant Dis 91: 685–691.

Roberts, W.R. and Roalson, E.H. (2017). Comparative transcriptome analyses of flower development in four species of *Achimenes* (Gesneriaceae). BMC Genomics 18: 240–26.

Rosenblum, I.M. and Basile, D.V. (1984). Hormonal regulation of morphogenesis in *Streptocarpus* and its relevance to evolutionary history of the Gesneriaceae. Am J Bot 71: 52–64.

Royston, J.P. (1982). An extension of Shapiro and Wilk's W test for normality to large samples. J R Stat Soc Ser C Appl Stat 31: 115–124.

Russell, J.J., Theriot, J.A., Sood, P., Marshall, W.F., Landweber, L.F., Fritz-Laylin, L., Polka, J.K., Oliferenko, S., Gerbich, T., Gladfelter, A., Umen, J., Bezanilla, M., Lancaster, M.A., He, S., Gibson, M.C., Goldstein, B., Tanaka, E.M., Hu, C.-K., and Brunet, A. (2017). Non-model model organisms. BMC Biol 15: 55–86.

Sakaguchi, S., Sugino, T., Tsumura, Y., Ito, M., Crisp, M.D., Bowman, D.M.J.S., Nagano, A.J., Honjo, M.N., Yasugi, M., Kudoh, H., Matsuki, Y., Suyama, Y. and Isagi, Y (2015). High-throughput linkage mapping of Australian white cypress pine (*Callitris glaucophylla*) and map transferability to related species. Tree Genet Genomes 11: 403–413.

Sakamoto, W., Miyagishima, S.-Y., and Jarvis, P. (2008). Chloroplast biogenesis: control of plastid development, protein Import, division and inheritance. *Arabidopsis* Book 6: e0110.

Sánchez, C., Villacreses, J., Blanc, N., Espinoza, L., Martinez, C., Pastor, G., Manque, P., Undurraga, S.F., and Polanco, V. (2016). High quality RNA extraction from Maqui berry for its application in next-generation sequencing. Springerplus 5: 1243.

Sánchez-Monge, A., Flores, L., Salazar, L., Hockland, S., and Bert, W. (2015). An updated list of the plants associated with plant-parasitic *Aphelenchoides* (Nematoda: Aphelenchoididae) and its implications for plant-parasitism within this genus. Zootaxa 4013: 207–224.

Sanger, F., Nicklen, S., and Coulson, A.R. (1977). DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci USA 74: 5463–5467.

Sato, S., Nakamura, Y., Kaneko, T., Asamizu, E., and Tabata, S. (1999). Complete structure of the chloroplast genome of *Arabidopsis thaliana*. DNA Res 6: 283–290.

Sboner, A., Mu, X.J., Greenbaum, D., Auerbach, R.K., and Gerstein, M.B. (2011). The real cost of sequencing: higher than you think! Genome Biol 12: 125–135.

Schmieder, R. and Edwards, R. (2011). Quality control and preprocessing of metagenomic datasets. Bioinformatics 27: 863–864.

Schmülling, T., Werner, T., Riefler, M., Krupková, E., and Bartrina y Manns, I. (2003). Structure and function of cytokinin oxidase/dehydrogenase genes of maize, rice, *Arabidopsis* and other species. J Plant Res 116: 241–252.

Schnable, P.S., Ware, D., Fulton, R.S., Stein, J.C., Wei, F., Pasternak, S., Liang, C., Zhang, J., Fulton, L., Graves, T.A., Minx, P., Reily, A.D., Courtney, L., Kruchowski, S.S., Tomlinson, C., Strong, C., Delehaunty, K., Fronick, C., Courtney, B., Rock, S.M., Belter, E., Du, F., Kim, K., Abbott, R.M., Cotton, M., Levy, A., Marchetto, P., Ochoa, K., Jackson, S.M., Gillam, B., Chen, W., Yan, L., Higginbotham, J., Cardenas, M., Waligorski, J., Applebaum, E., Phelps, L., Falcone, J., Kanchi, K., Thane, T., Scimone, A., Thane, N., Henke, J., Wang, T., Ruppert, J., Shah, N., Rotter, K., Hodges, J., Ingenthron, E., Cordes, M., Kohlberg, S., Sgro, J., Delgado, B., Mead, K., Chinwalla, A., Leonard, S., Crouse, K., Collura, K., Kudrna, D., Currie, J., He, R., Angelova, A., Rajasekar, S., Mueller, T., Lomeli, R., Scara, G., Ko, A., Delaney, K., Wissotski, M., Lopez, G., Campos, D., Braidotti, M., Ashley, E., Golser, W., Kim, H., Lee, S., Lin, J., Dujmic, Z., Kim, W., Talag, J., Zuccolo, A., Fan, C., Sebastian, A., Kramer, M., Spiegel, L., Nascimento, L., Zutavern, T., Miller, B., Ambroise, C., Muller, S., Spooner, W., Narechania, A., Ren, L., Wei, S., Kumari, S., Faga, B., Levy, M.J., McMahan, L., Van Buren, P., Vaughn, M.W., Ying, K., Yeh, C.-T., Emrich, S.J., Jia, Y., Kalyanaraman, A., Hsia, A.-P., Barbazuk, W.B., Baucom, R.S., Brutnell, T.P., Carpita, N.C., Chaparro, C., Chia, J.-M., Deragon, J.-M., Estill, J.C., Fu, Y., Jeddeloh, J.A., Han, Y., Lee, H., Li, P., Lisch, D.R., Liu, S., Liu, Z., Nagel, D.H., McCann, M.C., SanMiguel, P., Myers, A.M., Nettleton, D., Nguyen, J., Penning, B.W., Ponnala, L., Schneider, K.L., Schwartz, D.C., Sharma, A., Soderlund, C., Springer, N.M., Sun, Q., Wang, H., Waterman, M., Westerman, R., Wolfgruber, T.K., Yang, L., Yu, Y., Zhang, L., Zhou, S., Zhu, Q., Bennetzen, J.L., Dawe, R.K., Jiang, J., Jiang, N., Presting, G.G., Wessler, S.R., Aluru, S., Martienssen, R.A., Clifton, S.W., McCombie, W.R., Wing, R.A., and Wilson, R.K.(2009). The B73 maize genome: complexity, diversity, and dynamics. Science 326: 1112–1115.

Schneeberger, K. (2014). Using next-generation sequencing to isolate mutant genes from forward genetic screens. Nat Rev Genet 15: 662–676.

Schneider, C.A., Rasband, W.S., and Eliceiri, K.W. (2012). NIH Image to ImageJ: 25 years of image analysis. Nat Meth 9: 671–675.

Scott-Moncrieff, R. (1936). A biochemical survey of some mendelian factors for flower colour. J Genet 32: 117–170.

Sehgal, D., Singh, R., and Rajpal, V.R. (2016). ‘Quantitative Trait Loci Mapping in Plants: Concepts and Approaches’. In *Molecular Breeding for Sustainable Crop Improvement, Sustainable Development and Biodiversity* (ed. by Rajpal, V. R., Rao, S. R., Raina, S. N.). Springer Science & Business Media. Berlin. Germany.

Semagn, K., Bjørnstad, Å., and Ndjiondjop, M.N. (2006). Principles, requirements and prospects of genetic mapping in plants. Afr J Biotechnol 5: 2569–2587.

Serrano-Serrano, M.L., Marcionetti, A., Perret, M., and Salamin, N. (2017). Transcriptomic resources for an endemic neotropical plant lineage (Gesneriaceae). Appl Plant Sci 5: 1600135–1600142.

Shafer, A.B.A., Peart, C.R., Tusso, S., Maayan, I., Brelsford, A., Wheat, C.W., and Wolf, J.B.W. (2016). Bioinformatic processing of RAD-seq data dramatically impacts downstream population genetic inference. Methods Ecol Evol 8: 907–917.

Shaw, J., Lickey, E.B., Schilling, E.E., and Small, R.L. (2007). Comparison of whole chloroplast genome sequences to choose noncoding regions for phylogenetic studies in angiosperms: the tortoise and the hare III. Am J Bot 94: 275–288.

Shen, M.-J.R., Boutell, J.M., Stephens, K.M., Ronaghi, M., Gunderson, K., Venkatesan, B.M., Bowen, M.S., Vijayan, K., Illumina, Inc. (2014). *United States Patent No. WO2013188582A1.* Retrieved from https://patents.google.com/patent/WO2013188582A1.

Sheridan, H., Nestor, C., O'Driscoll, L., and Hook, I. (2011). Isolation, structure elucidation, and cytotoxic evaluation of furanonaphthoquinones from in vitro plantlets and cultures of *Streptocarpus dunnii*. J Nat Prod 74: 82–85.

Shinozaki, K., Ohme, M., Tanaka, M., Wakasugi, T., Hayashida, N., Matsubayashi, T., Zaita, N., Chunwongse, J., Obokata, J., Yamaguchi-Shinozaki, K., Ohto, C., Torazawa, K., Meng, B.Y., Sugita, M., Deno, H., Kamogashira, T., Yamada, K., Kusuda, J., Takaiwa, F., Kato, A., Tohdoh, N., Shimada, H., and Sugiura, M. (1986). The complete nucleotide sequence of the tobacco chloroplast genome: its gene organization and expression. EMBO J 5: 2043–2049.

Siedow, J.N. and Umbach, A.L. (1995). Plant mitochondrial electron transfer and molecular biology. Plant Cell 7: 821–831.

Siegfried, K.R., Eshed, Y., Baum, S.F., Otsuga, D., Drews, G.N., and Bowman, J.L. (1999). Members of the *YABBY* gene family specify abaxial cell fate in *Arabidopsis*. Development 126: 4117–4128.

Sierro, N., Battey, J.N.D., Ouadi, S., Bakaher, N., Bovet, L., Willig, A., Goepfert, S., Peitsch, M.C., and Ivanov, N.V. (2014). The tobacco genome sequence and its comparison with those of tomato and potato. Nat Commun 5: 3833–3842.

Silverstone, A.L., Tseng, T.S., Swain, S.M., Dill, A., Jeong, S.Y., Olszewski, N.E., and Sun, T.P. (2006). Functional analysis of SPINDLY in gibberellin signaling in *Arabidopsis*. Plant Physio 143: 987–1000.

Simão, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V., and Zdobnov, E.M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 31: 3210–3212.

Simbolo, M., Gottardi, M., Corbo, V., Fassan, M., Mafficini, A., Malpeli, G., Lawlor, R.T., and Scarpa, A. (2013). DNA qualification workflow for next generation sequencing of histopathological samples. PLoS ONE 8: e62692.

Simpson, J.T., Wong, K., Jackman, S.D., Schein, J.E., Jones, S.J.M., and Birol, I. (2009). ABySS: a parallel assembler for short read sequence data. Genome Res 19: 1117–1123.

Sims, D., Sudbery, I., Ilott, N.E., Heger, A., and Ponting, C.P. (2014). Sequencing depth and coverage: key considerations in genomic analyses. Nat Rev Genet 15: 121–132.

Singh, V.K. and Jain, M. (2014). Transcriptome profiling for discovery of genes involved in shoot apical meristem and flower development. Genom Data 2: 135–138.

Slusarenko, A.J. and Schlaich, N.L. (2003). Downy mildew of *Arabidopsis thaliana* caused by *Hyaloperonospora parasitica* (formerly *Peronospora parasitica*). Mol Plant Pathol 4: 159–170.

Small, C.M., Bassham, S., Catchen, J., Amores, A., Fuiten, A.M., Brown, R.S., Jones, A.G., and Cresko, W.A. (2016). The genome of the Gulf pipefish enables understanding of evolutionary innovations. Genome Biol 17: 258–281.

Smith, D.R. (2016). Goodbye genome paper, hello genome report: the increasing popularity of "genome announcements" and their impact on science. Brief Funct Genomics 16: 156-162.

Soyars, C.L., James, S.R., and Nimchuk, Z.L. (2016). Ready, aim, shoot: stem cell regulation of the shoot apical meristem. Curr Opin Plant Biol 29: 163–168.

Stahle, M.I., Kuehlich, J., Staron, L., Arnim, von, A.G., and Golz, J.F. (2009). YABBYs and the transcriptional corepressors LEUNIG and LEUNIG_HOMOLOG maintain leaf polarity and meristem activity in *Arabidopsis*. Plant Cell 21: 3105–3118.

Stange, M., Utz, H.F., Schrag, T.A., Melchinger, A.E., and Würschum, T. (2013). High-density genotyping: an overkill for QTL mapping? Lessons learned from a case study in maize and simulations. Theor Appl Genet 126: 2563–2574.

Stein, L.D. (2010). The case for cloud computing in genome informatics. Genome Biol 11: 207.

Stöckigt, J., Srocka, U., and Zenk, M.H. (1973). Structure and biosynthesis of a new anthraquinone from *Streptocarpus dunnii*. Phytochemistry 12: 2389–2391.

Sturtevant, A.H. (1913). The linear arrangement of six sex-linked factors in *Drosophila*, as shown by their mode of association. J Exp Zool 14: 43–59.

Stuurman, J., Jäggi, F., and Kuhlemeier, C. (2002). Shoot meristem maintenance is controlled by a *GRAS*-gene mediated signal from differentiating cells. Genes Dev 16: 2213–2218.

Sugiyama, Y., Watase, Y., Nagase, M., Makita, N., Yagura, S., Hirai, A., and Sugiura, M. (2004). The complete nucleotide sequence and multipartite organization of the tobacco mitochondrial genome: comparative analysis of mitochondrial genomes in higher plants. Mol Genet Genomics 272: 603–615.

Sun, L., Wang, J., Zhu, X., Jiang, L., Gosik, K., Sang, M., Sun, F., Cheng, T., Zhang, Q., and Wu, R. (2017). HpQTL: a geometric morphometric platform to compute the genetic architecture of heterophylly. Brief Bioinform 19: 603–612.

Tan, S.C. and Yiap, B.C. (2009). DNA, RNA, and protein extraction: the past and the present. J Biomed Biotechnol 2009: 574398–574408.

Tan, Y. (2018). Trichome morphology and development in the genus. PhD thesis. University of Edinburgh. Edinburgh, UK.

Tao, Q., Guo, D., Wei, B., Zhang, F., Pang, C., Jiang, H., Zhang, J., Wei, T., Gu, H., Qu, L.-J., and Qin, G. (2013). The TIE1 transcriptional repressor links TCP transcription factors with TOPLESS/TOPLESS-RELATED corepressors and modulates leaf development in *Arabidopsis*. Plant Cell 25: 421–437.

Tatsuzawa, F. and Hosokawa, M. (2015). Flower colors and their anthocyanins in *Saintpaulia* cultivars (Gesneriaceae). Hort J 85: 63–69.

Tillich, M., Lehwark, P., Pellizzer, T., Ulbricht-Jones, E.S., Fischer, A., Bock, R., and Greiner, S. (2017). GeSeq – versatile and accurate annotation of organelle genomes. Nucleic Acids Res 45: W6–W11.

Tononi, P., Möller, M., Bencivenga, S., and Spada, A. (2010). *GRAMINIFOLIA* homolog expression in *Streptocarpus rexii* is associated with the basal meristems in phyllomorphs, a morphological novelty in Gesneriaceae. Evol Dev 12: 61–73.

Triboush, S.O., Danilenko, N.G., and Davydenko, O.G. (1998). A method for isolation of chloroplast DNA and mitochondrial DNA from sunflower. Plant Mol Biol Rep 16: 183–183.

Ullrich, A., Shine, J., Chirgwin, J., Pictet, R., Tischer, E., Rutter, W., and Goodman, H. (1977). Rat insulin genes: construction of plasmids containing the coding sequences. Science 196: 1313–1319.

Unseld, M., Marienfeld, J.R., Brandt, P., and Brennicke, A. (1997). The mitochondrial genome of *Arabidopsis thaliana* contains 57 genes in 366,924 nucleotides. Nat Genet 15: 57–61.

Vales, M.I., Schön, C.C., Capettini, F., Chen, X.M., Corey, A.E., Mather, D.E., Mundt, C.C., Richardson, K.L., Sandoval-Islas, J.S., Utz, H.F., and Hayes, P.M. (2005). Effect of population size on the estimation of QTL: a test using resistance to barley stripe rust. Theor Appl Genet 111: 1260–1270.

Van Dijk, E.L., Jaszczyszyn, Y., Naquin, D., and Thermes, C. (2018). The third revolution in sequencing technology. Trends Genet 34: 666-681.

Van Nimwegen, K.J.M., van Soest, R.A., Veltman, J.A., Nelen, M.R., van der Wilt, G.J., Vissers, L.E.L.M., and Grutters, J.P.C. (2016). Is the $1000 genome as near as we think? a cost analysis of next-generation sequencing. Clin Chem 62: 1458–1464.

Van Ooijen, J.W. (2006). JoinMap ® 4, Software for the calculation of genetic linkage maps in experimental populations. Available at: https://www.kyazma.nl/index.php/JoinMap/

Van Ooijen, J.W. and Jansen, J. (2013). Genetic Mapping in Experimental Populations. Cambridge University Press. Cambridge. UK.

Van Tassell, C.P., Smith, T.P.L., Matukumalli, L.K., Taylor, J.F., Schnabel, R.D., Lawley, C.T., Haudenschild, C.D., Moore, S.S., Warren, W.C., and Sonstegard, T.S. (2008). SNP discovery and allele frequency estimation by deep sequencing of reduced representation libraries. Nat Meth 5: 247–252.

Verdan, M.H. and Stefanello, M.É.A. (2012). Secondary metabolites and biological properties of Gesneriaceae species. Chem Biodivers 9: 2701–2731.

Viana, J.M.S., Silva, F.F., Mundim, G.B., Azevedo, C.F., and Jan, H.U. (2016). Efficiency of low heritability QTL mapping under high SNP density. Euphytica 213: 13–24.

Visser, E.A., Wegrzyn, J.L., Steenkmap, E.T., Myburg, A.A., and Naidoo, S. (2015). Combined *de novo* and genome guided assembly and annotation of the *Pinus patula* juvenile shoot transcriptome. BMC Genomics 16: 1057.

Voorrips, R.E. (2002). MapChart: software for the graphical presentation of linkage maps and QTLs. J Hered 93: 77–78.

Vurture, G.W., Sedlazeck, F.J., Nattestad, M., Underwood, C.J., Fang, H., Gurtowski, J., and Schatz, M.C. (2017). GenomeScope: fast reference-free genome profiling from short reads. Bioinformatics 33: 2202–2204.

Wang, D., Cao, G., Fang, P., Xia, L., and Cheng, B. (2017a). Comparative transcription analysis of different *Antirrhinum* phyllotaxy nodes identifies major signal networks involved in vegetative-reproductive transition. PLoS ONE 12: e0178424.

Wang, L., Xia, Q., Zhang, Y., Zhu, X., Zhu, X., Li, D., Ni, X., Gao, Y., Xiang, H., Wei, X., Yu, J., Quan, Z., and Zhang, X. (2016). Updated sesame genome assembly and fine mapping of plant height and seed coat color QTLs using a new high-density genetic map. BMC Genomics 17: 31.

Wang, N., Thomson, M., Bodles, W.J.A., Crawford, R.M.M., Hunt, H.V., Featherstone, A.W., Pellicer, J., and Buggs, R.J.A. (2013). Genome sequence of dwarf birch (*Betula nana*) and cross-species RAD markers. Mol Ecol 22: 3098–3111.

Wang, Y., Liu, K., Bi, D., Zhou, S., and Shao, J. (2017). Characterization of the transcriptome and EST-SSR development in *Boea clarkeana*, a desiccation-tolerant plant endemic to China. PeerJ 5: e3422.

Wang, Z., Gerstein, M., and Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet 10: 57–63.

Waters, M.T. and Langdale, J.A. (2009). The making of a chloroplast. EMBO J 28: 2861–2873.

Weber, A. (2004). ‘Gesneriaceae’. in *Flowering Plants · Dicotyledons* (ed. by Kubitzki, K., Rohwer, J. G., Bittrich, V.). Springer Publishing. New York. USA. pp. 63–158.

Weber, A., Clark, J.L., and Möller, M. (2013). A new formal classification of Gesneriaceae. Selbyana 31: 68–94.

Weiss, J., Delgado-Benarroch, L., and Egea-Cortines, M. (2005). Genetic control of floral size and proportions. Int J Dev Biol 49: 513–525.

Wessinger, C.A., Hileman, L.C., and Rausher, M.D. (2014). Identification of major quantitative trait loci underlying floral pollination syndrome divergence in *Penstemon*. Phil Trans R Soc B 369: 20130349–20130349.

Wheeler, T.J. and Eddy, S.R. (2013). nhmmer: DNA homology search with profile HMMs. Bioinformatics 29: 2487–2489.

Wicke, S., Schneeweiss, G.M., dePamphilis, C.W., Müller, K.F., and Quandt, D. (2011). The evolution of the plastid chromosome in land plants: gene content, gene order, gene function. Plant Mol Biol 76: 273–297.

Wolf, J.B.W. (2013). Principles of transcriptome analysis and gene expression quantification: an RNA-seq tutorial. Mol Ecol Resour 13: 559–572.

Xiao, L., Yang, G., Zhang, L., Yang, X., Zhao, S., Ji, Z., Zhou, Q., Hu, M., Wang, Y., Chen, M., Xu, Y., Jin, H., Xiao, X., Hu, G., Bao, F., Hu, Y., Wan, P., Li, L., Deng, X., Kuang, T., Xiang, C., Zhu, J.-K., Oliver, M.J., and He, Y. (2015). The resurrection genome of *Boea hygrometrica*: A blueprint for survival of dehydration. Proc Natl Acad Sci USA 112: 5833–5837.

Xu, S. and Atchley, W.R. (1996). Mapping quantitative trait loci for complex binary diseases using line crosses. Genetics 143: 1417–1424.

Yandell, M. and Ence, D. (2012). A beginner's guide to eukaryotic genome annotation.Nat Rev Genet 13: 329–342.

Yang, C., Chen, Z., Zhuang, C., Mei, M., and Liu, Y. (2004). Genetic and physical fine-mapping of the Sc locus conferring *indica-japonica* hybrid sterility in rice (*Oryza sativa* L.). Chinese Sci Bull 49: 1718–1725.

Yang, H., Tao, Y., Zheng, Z., Zhang, Q., Zhou, G., Sweetingham, M.W., Howieson, J.G., and Li, C. (2013). Draft genome sequence, and a sequence-defined genetic linkage map of the legume crop species *Lupinus angustifolius* L. PLoS ONE 8: e64799.

Yang, Y., Xie, B., and Yan, J. (2014). Application of Next-generation Sequencing Technology in Forensic Science. Genom Proteom Bioinf 12: 190–197.

Yanofsky, M.F. (1995). Floral meristems to floral organs: genes controlling early events in *Arabidopsis* flower development. Annu Rev Plant Physiol Plant Mol Biol 46: 167–188.

Ye, Z., Li, J., and Wang, G. (2014). *Agrobacterium*-mediated genetic transformation of *AtTIP5*; 1 gene into *Saintpaulia ionantha*. Acta Botanica Boreali-Occidentalia Sinica 34: 2412–2417.

Zeng, Z.B. (1994). Precision mapping of quantitative trait loci. Genetics 136: 1457–1468.

Zhang, J., Sci, J.S.J.C., 2000 (2000). Economical and rapid method for extracting cotton genomic DNA. J Cotton Sci 4: 193–201.

Zhang, R., Calixto, C.P.G., Marquez, Y., Venhuizen, P., Tzioutziou, N.A., Guo, W., Spensley, M., Entizne, J.C., Lewandowska, D., Have, ten, S., Frei dit Frey, N., Hirt, H., James, A.B., Nimmo, H.G., Barta, A., Kalyna, M., and Brown, J.W.S. (2017). A high quality *Arabidopsis* transcriptome for accurate transcript-level analysis of alternative splicing. Nucleic Acids Res 45: 5061–5073.

Zhang, T., Fang, Y., Wang, X., Deng, X., Zhang, X., Hu, S., and Yu, J. (2012). The complete chloroplast and mitochondrial genome sequences of *Boea hygrometrica*: insights into the evolution of plant organellar genomes. PLoS ONE 7: e30531.

Zhang, Z.-L., Ogawa, M., Fleet, C.M., Zentella, R., Hu, J., Heo, J.-O., Lim, J., Kamiya, Y., Yamaguchi, S., and Sun, T.-P. (2011). SCARECROW-LIKE 3 promotes gibberellin signaling by antagonizing master growth repressor DELLA in *Arabidopsis*. Proc Natl Acad Sci USA 108: 2160–2165.

Zhao, L., Nakazawa, M., Takase, T., Manabe, K., Kobayashi, M., Seki, M., Shinozaki, K., and Matsui, M. (2004). Overexpression of *LSH1*, a member of an uncharacterised gene family, causes enhanced light regulation of seedling development. Plant J. 37: 694–706.

Zhao, P., Zhou, H.-J., Potter, D., Hu, Y.-H., Feng, X.-J., Dang, M., Feng, L., Zulfiqar, S., Liu, W.-Z., Zhao, G.-F., and Woeste, K. (2018). Population genetics, phylogenomics and hybrid speciation of *Juglans* in China determined from whole chloroplast genomes, transcriptomes, and genotyping-by-sequencing (GBS). Mol Phylogenet Evol 126: 250–265.

Zhou, Z., Jiang, Y., Wang, Z., Gou, Z., Lyu, J., Li, W., Yu, Y., Shu, L., Zhao, Y., Ma, Y., Fang, C., Shen, Y., Liu, T., Li, C., Li, Q., Wu, M., Wang, M., Wu, Y., Dong, Y., Wan, W., Wang, X., Ding, Z., Gao, Y., Xiang, H., Zhu, B., Lee, S.-H., Wang, W., and Tian, Z. (2015). Resequencing 302 wild and cultivated accessions identifies genes related to domestication and improvement in soybean. Nature Biotechnol 33: 408–414.

Appendices

Appendix 2.1

4% CTAB extraction method

Chemicals:

4% CTAB solution
(100 mM Tris HCl pH 8.0, 1.4 M NaCl, 20 mM EDTA, 4% CTAB)
β-mercaptoethanol (Sigma-Aldrich, Merck, Darmstadt, Germany)
Chloroform:isoamyl alcohol (24:1)
Isopropanol (Sigma-Aldrich)
Wash buffer (10 mM ammonium acetate, 76% ethanol)
TE buffer (10 mM Tris, 1 mM EDTA, pH 8.0, Sigma-Aldrich)

Procedures:

Day 1

1. Preheat CTAB buffer (1 ml 4% CTAB; 2 µl β-ME; 2% PVPP) at 65°C
2. Add 1 ml of preheated buffer to the sample and gently shake
3. Incubate the sample in a 65°C heat-block for at least 60 minutes. Occasionally mix by inverting
4. Remove the tube from the heat-block and keep at room temperature for ~5 minutes
5. Add 500 µl chloroform-IAA (24:1) and shake vigorously. Open the lid of the Eppendorf tube to release gas, then put in an orbital shaker to shake at minimum speed for 30 minutes
6. Gently mix by shaking, then centrifuge at 11,000 rpm for 10 minutes
7. Carefully transfer the aqueous phase (about 800 - 850 µl) to a clean Eppendorf tube
8. Repeat step 5 - 7 (transfer about 650 – 700 µl of the aqueous phase this time)
9. Add an equal amount (700 µl) of ice-cold isopropanol and rock gently
10. Store the sample in the -20°C freezer overnight

Day 2

11. Centrifuge at 8,000 rpm for 10 minutes
12. Remove supernatant and add 500 µl of wash buffer. Shake gently and check that the pellet is floating. Leave at room temperature for at least 30 minutes
13. Centrifuge at 8,000 rpm for 10 minutes
14. Remove the supernatant and dry the pellet in a SpeedVac machine for 10 -15 minutes
15. Dissolve the pellet in 50 µl TE buffer
16. Incubate the sample at 50°C for 10 minutes, then leave at room temperature for 1 hour to fully dissolve the DNA
17. Store the stock DNA at -20°C for long-term storage

Appendix 2.2

ChargeSwitch gDNA Plant Kit protocol

Note: the protocol included below refers to the ChargeSwitch Kit$^{\text{Extend time}}$ protocol

Chemicals:

From the kit –

Precipitation buffer (N5)
Lysis buffer (L18)
10% SDS
10% Detergent (D1)
Magnetic beads solution
Wash buffer (W12)
Elution buffer (E6)

Procedures:

1. Chill the precipitation buffer (N5) on ice
2. Grind the leaf tissue (4 leaf discs for one sample)
3. Add 1 ml of Lysis buffer (L18) to the sample and incubate at room temperature for 1 hour
4. Vortex the ground tissue until the sample is completely resuspended
5. Add 100 µl 10% SDS to the 1 ml plant lysate and leave for 30 minutes at room temperature
6. Add 400 µl of Precipitation buffer (N5). Mix by inversion and leave for 30 minutes on ice
7. Centrifuge at maximum speed for 5 minutes
8. Transfer the clear lysate to a clean tube
9. Thoroughly vortex the magnetic beads tube and fully resuspend the beads
10. Add 100 µl 10% Detergent (D1) to the lysate
11. Add 40 µl beads to the lysate, and mix gently by pipetting up and down five times
12. Incubate for 30 minutes
13. Place the tubes on the MagnaRack until a tight pellet is formed. Carefully remove the supernatant without removing the tubes from the rack
14. Remove the tubes from the rack
15. Add 1 ml Wash buffer (W12) and gently pipette five times to mix
16. Place the tubes on the MagnaRack until a tight pellet is formed. Remove the supernatant.
17. Repeat steps 15-16
18. Make sure no supernatant remains, and remove the tubes from the rack
19. Add 150 µl Elution buffer (E6), and pipette at least 30 times to mix, until no bead clumps are visible
20. Incubate at room temperature for 30 minutes
21. Place the tube on the rack until a tight pellet is formed. Transfer the clear supernatant to a clean Eppendorf tube
22. Store the stock DNA at -20°C for long-term storage

Appendix 2.3

DNAzol method

Chemicals:

Plant DNAzol^{TM} reagent (Invitrogen, Thermo Fisher Scientific)
100% ethanol
75% ethanol
TE buffer (10 mM Tris, 1 mM EDTA, pH 8.0, Sigma-Aldrich)

Procedures:

1. Grind the tissue in liquid nitrogen
2. Add 0.3 ml of DNAzol to the tube and mix vigorously
3. Shake for 30 minutes
4. Add 0.3 ml of chloroform, mix vigorously and shake for 5 minutes
5. Centrifuge at 10,000 rpm for 10 minutes
6. Transfer the aqueous phase to a clean tube
7. Add 225 µl of 100% EtOH, mix by inverting 6-8 times and incubate for 5 minutes
8. Centrifuge at 7,000 rpm for 4 minutes
9. Add 0.3 µl of DNAzol-EtOH wash solution, mix by vortex
   * Wash solution: 1 volume of DNAzol with 0.75 volumes of 100% EtOH
   To prepare 0.6 ml of wash solution, mix 343 µl DNAzol with 257 µl EtOH
10. Incubate the sample for 5 minutes
11. Centrifuge at 7,000 rpm for 4 minutes
12. Remove the EtOH and air dry
13. Dissolve the DNA pellet in 70 µl TE buffer
14. Store the stock DNA at -20°C for long-term storage

Appendix 2.4

DNeasy Plant Mini Kit protocol

Note: the protocol included below refers to the DNeasy Kit$^{\text{Extend time}}$ protocol

Chemicals:

From the DNeasy Kit –

- P3 buffer
- AW1 buffer
- AW2 buffer
- AE buffer

Procedures:

1. Grind the tissue in liquid nitrogen
2. Add 400 µl AP1 to the tube and vortex
3. Incubate at 65°C for 30 minutes. Mix occasionally by inverting 2-3 times
4. Add 130 µl P3 buffer. Mix and incubate on ice for 5 minutes
5. Centrifuge at 13,000 rpm (~20,000 xg) for 5 minutes
6. Transfer the lysate to a QIAshredder Mini Spin column in a 2 ml collection tube
7. Centrifuge at 13,000 rpm for 2 minutes
8. Transfer the flow-through to a clean tube
9. Add 1.5 volumes AW1 and mix by pipetting
10. Transfer 650 µl of the mixture into a DNeasy Mini Spin column placed in a 2 ml collection tube
11. Centrifuge at ≥ 8,000 rpm (≥6000 xg) for 1 minute
12. Discard flow-through
13. Repeat steps 9 - 12 with the rest of the sample
14. Place the DNeasy Mini Spin column into new 2 ml collection tube
15. Add 500 µl AW2
16. Centrifuge at ≥ 8,000 rpm (≥6000 xg) for 1 minute
17. Discard the flow-through
18. Add 500 µl AW2
19. Centrifuge at 13,000 rpm (20000 xg) for 2 minutes
20. Discard the flow-through
21. Transfer the column to a clean Eppendorf tube
22. Add 100 µl AE to the membrane and incubate for 5 minutes
23. Centrifuge at ≥ 8000 rpm (≥6000 xg) for 1 minute to collect the DNA
24. Keep the DNA at -20°C for long-term storage

Appendix 2.5

CTAB extraction + RNase A treatment + phenol-chloroform purification (used for preparation of RAD sequencing samples)

Chemicals:
4% CTAB solution (100 mM Tris HCl pH 8.0, 1.4 M NaCl, 20 mM EDTA, 4% CTAB)
β-mercaptoethanol (Sigma-Aldrich, Merck, Darmstadt, Germany)
Chloroform:isoamyl alcohol (24:1)
Isopropanol (Sigma-Aldrich)
Wash buffer (10 mM ammonium acetate, 76% ethanol)
TE buffer (10 mM Tris, 1 mM EDTA, pH 8.0, Sigma-Aldrich)
RNase A (1/5 from the original stock, 12091021, Invitrogen, Thermo Fisher Scientific)
Phenol:chloroform:isoamyl alcohol 25:24:1 (pH 8.0)
Chloroform:isoamyl alcohol 24:1
3 M sodium acetate (Sigma-Aldrich, Merck)
100% ethanol

Note: perform four CTAB reactions (totally 16 leaf discs) for each individual to be extracted, which will be combined before the RNase A treatment.

Procedures:
Day 1

1. Preheat CTAB buffer (1 ml 4% CTAB; 2 µl β-ME; 2% PVPP) at 65°C
2. Add 1 ml of preheated buffer to the sample and gently shake
3. Incubate the sample in a 65°C heat block for at least 60 minutes. Occasionally mix by inverting
4. Remove the tube from heat block and keep at room temperature for ~5 minutes
5. Add 500 µl chloroform:isoamyl alcohol 24:1 and shake vigorously. Open the lid of the Eppendorf tube to release gas, then put on an orbital shaker to shake at minimum speed for 30 minutes
6. Gently mix by shaking, then centrifuge at 11,000 rpm for 10 minutes
7. Carefully transfer the aqueous phase (about 800 - 850 µl) to a clean Eppendorf tube
8. Repeat step 5 - 7 (transfer about 650 – 700 µl of the aqueous phase this time)
9. Add an equal amount (700 µl) of ice-cold isopropanol and rock gently
10. Store the sample in -20°C freezer overnight

Day 2

11. Centrifuge at 8,000 rpm for 10 minutes
12. Remove supernatant and add 500 µl of wash buffer. Shake gently and check that the pellet is floating. Leave at room temperature for at least 30 minutes
13. Centrifuge at 8,000 rpm for 10 minutes
14. Remove the supernatant and dry the pellet in a SpeedVac machine for 10-15 minutes
15. Dissolve the pellet in 100 µl of TE buffer
16. Incubate the sample at 50°C for 10 minutes, then keep at room temperature for 1 hour to fully dissolve the pellet
17. Briefly spin down the sample, than carefully mix up all 4 tubes of extractions from the same individual (so the total amount is 400 µl). Make sure to transfer all liquid and possible pellet

18. Add in 2 µL RNase and mix by inversion. Keep at room temperature for 5-10 minutes (<10 minutes)
19. Add 400 µl phenol:chloroform:isoamyl alcohol (pH 8.0) and mix by shaking vigorously
20. Centrifuge at 11,000rpm for 10 minutes
21. Transfer the aqueous phase to a clean Eppendorf tube (~400 µl)
22. Add 400 µl chloroform:isoamyl alcohol 24:1 and mix by shaking vigorously
23. Centrifuge at 11,000 rpm for 10 minutes
24. Transfer the aqueous phase to a clean Eppendorf tube (~400 µl)
25. Add in 1/10 times volume of 3 M sodium acetate (about 40 µl), followed by 2-2.5 times volume ethanol (about 1,000 µl)
26. Keep the sample in -20°C freezer overnight

Day 3

27. Centrifuge at 8,000 rpm for 10 minutes
28. Remove the supernatant
29. Add 1 ml of 70% EtOH for washing. Keep for 30 minutes
30. Centrifuge at 11,000 rpm for 5 minutes
31. Remove the supernatant
32. Dry the pellet using the SpeedVac machine for 20-40 minutes. Dry the pellet completely
33. Add 20 µl of TE buffer
34. Incubate the sample at 55°C for 30 minutes and leave for a while at 4°C to fully dissolve the DNA. If the pellet did not dissolve, add an additional 10 µl of TE buffer
35. Take 1 µl of DNA and mix with 9 µl of sterilised water for evaluation (1/10x dilution)

    Check a. NanoVue (use 3 µl diluted sample)
    b. Gel electrophoresis (use 5 µl diluted sample)
    c. Qubit assay (use 2 µl diluted sample)
36. Store the stock DNA at -20°C for long-term storage

Appendix 2.6

ChargeSwitch Kit$^{\text{Extend time}}$ + RNase A treatment + phenol-chloroform purification (used for preparation of whole genome shotgun sequencing samples)

Chemicals:

From the kit –

Precipitation buffer (N5)
Lysis buffer (L18)
10% SDS
10% Detergent (D1)
Magnetic beads solution
Wash buffer (W12)
Elution buffer (E6)

RNase A (1/5 from the original stock, 12091021, Invitrogen, Thermo Fisher Scientific)
Phenol:chloroform:isoamyl alcohol 25:24:1 (pH 8.0)
Chloroform:isoamyl alcohol 24:1
3 M sodium acetate (Sigma-Aldrich, Merck)
100% ethanol
TE buffer (10 mM Tris, 1 mM EDTA, pH 8.0, Sigma-Aldrich)

Note: process at least 2 tubes of reactions at a time, which will be mixed before the RNase A treatment.

Procedures:

Day 1

1. Chill the Precipitation buffer (N5) on ice
2. Grind the leaf tissue (4 leaf discs for one sample)
3. Add 1 ml Lysis buffer (L18) to the sample and incubate at room temperature for 1 hour
4. Vortex the ground tissue until the sample is completely resuspended
5. Add 100 µl 10% SDS to the 1 ml plant lysate and leave for 30 minutes at room temperature
6. Add 400 µl Precipitation buffer (N5). Mix by inversion and leave for 30 minutes on ice
7. Centrifuge at maximum speed for 5 minutes
8. Transfer the clear lysate to a clean tube
9. Thoroughly vortex the magnetic beads solution until the beads are completely resuspended
10. Add 100 µl 10% Detergent (D1) to the lysate
11. Add 40 µl of magnetic beads to the lysate, and mix gently by pipetting up and down for five times
12. Incubate for 30 minutes
13. Place the tubes on the MagnaRack until a tight pellet is formed. Carefully remove the supernatant without removing the tubes from the magnetic rack
14. Remove the tubes from the rack
15. Add 1 ml Wash buffer (W12) and gently pipette for 5 times
16. Place tubes on the MagnaRack until a tight pellet is formed. Remove supernatant.

17. Repeat steps 15-16
18. Make sure no supernatant remains and remove the tubes from the rack
19. Add 150 µl Elution buffer (E6), and pipette for at least 30 times until no bead clumps are visible
20. Incubate at room temperature for 30 minutes
21. Place tube on rack until a tight pellet is formed. Transfer the clear supernatant to a clean Eppendorf tube
22. Combine both tubes of eluate, making a total volume of 300 µl
23. Warm up the DNA solution to 65°C for 30 minutes, with occasional inversion
24. Add 2 µl RNase A, mix by inversion and incubate for 5-10 minutes
25. Add 300 µl of phenol:chloroform:isoamyl alcohol pH 8.0 (PCI) solution to the sample
26. Shake on an orbital shaker for 30 minutes
27. Centrifuge at 12,000 g (ca.11,000 rpm) for 10 minutes
28. Transfer the supernatant to a clean Eppendorf tube
29. Add 300 µl PCI pH 8.0 to the sample
30. Shake on an orbital shaker for 30 minutes
31. Centrifuge at 12,000 g (ca.11,000 rpm) for 10 minutes
32. Transfer the supernatant to a clean Eppendorf tube
33. Add equal amount of chloroform:isoamyl alcohol 24:1 to the sample
34. Shake on an orbital shaker for 30 minutes
35. Centrifuge at 12,000 g (ca.11,000rpm) for 10 minutes
36. Transfer the supernatant to a clean Eppendorf tube
37. Add 1/10 volumes of 3 M sodium acetate
38. Add 2.5 volumes of 100% ethanol
39. Keep the sample at -20°C overnight

Day 2

40. Centrifuge at 12,000 g (ca. 11,000 rpm) for 10 minutes
41. Discard the supernatant
42. Add 1000 µl of 70% ethanol for washing. Leave for 30 minutes with occasional gently tapping
43. Centrifuge at 12,000 g (ca. 11,000 rpm) for 10 minutes
44. Discard the supernatant
45. Dry the pellet completely using the SpeedVac machine
46. Dissolve the pellet in 20 µl TE buffer. Keep at 65°C for 1 hour to fully dissolve the pellet
47. Take 1 µl of DNA and mix with 9 µl of sterilised water for quality check (1/10x dilution)

    Check a. NanoVue (use 3 µl of diluted sample)
    b. Gel electrophoresis (use 5µl of diluted sample)
    c. Qubit assay (use 2 µl of diluted sample)
48. Store the stock DNA at -20°C for long-term storage

Appendix 2.7

TRIzol reagent RNA extraction + acidic phenol purification + PureLink RNA Mini Kit clean up

Chemicals:

TRIzol reagent (Invitrogen, Thermo Fisher Scientific)
Chloroform (BDH, VWR International)
Isopropanol
DEPC water
Phenol:chloroform 5:1 pH 4.3 – 4.7 (Sigma-Aldrich, Merck)
75% ethanol
β-mercaptoethanol (Sigma-Aldrich, Merck, Darmstadt, Germany)
From the PureLink RNA Mini Kit –

- Lysis buffer
- Wash buffer I
- Wash buffer II
- RNase-free water

Note: Extract multiple tubes for each tissue type. The tubes are combined before the acidic phenol purification.

Procedures:

Day 1

1. Place 1 ml of TRIzol in a 2 ml Eppendorf tube
2. Grind the tissue to fine powder using mortar and pestle with liquid nitrogen
3. Transfer the powder into the tube containing the TRIzol. The total amount of the mixture should be around 1.2 – 1.8 ml (avoid too little or too much tissue)
4. Gently shake on orbital shaker and incubate for 55 minutes to 1 hour
5. Centrifuge at 11,000 rpm and 4°C for 10 minutes
6. Transfer the supernatant to a clean 2 ml tube
7. Add 0.2 ml chloroform to the sample and mix by shaking vigorously for 15 seconds. Leave the tube for 2-3 minutes
8. Centrifuge at 11,000 rpm and 4°C for 15 minutes
9. Transfer the aqueous phase to a clean 1.5 ml Eppendorf tube
10. Add 0.5 ml of ice-cold isopropanol
11. Keep the sample at -20°C for more than 1 hour
12. Centrifuge at 11,000 rpm and 4°C for 10 minutes
13. Add in DEPC water to dissolve the pellet. The amount added is according to the note below:

    For 4 extraction tubes add 75 µl to each tube. Combine all 4 tubes to obtain a total of 300 µl.

    For 3 extraction tubes add 100 µl to each tube. Combine all 3 tubes to obtain a total of 300 µl.

    For 2 extraction tubes add 150 µl to each tube. Combine both tubes to obtain a total of 300 µl.
14. Add 300 µl phenol:chloroform 5:1 to the sample and mix by shaking vigorously
15. Centrifuge at 11,000 rpm for 10 minutes

16. Transfer the supernatant to a clean 1.5 ml Eppendorf tube
17. Repeat steps 14 – 16
18. Add 300 µl of ice-cold isopropanol
19. Keep the sample at -20°C overnight

Day 2

20. Centrifuge at 11,000 rpm for 10 minutes
21. Discard the supernatant
22. Add 500 µl of 75% ethanol
23. Centrifuge at 9,000 rpm for 5 minutes
24. Discard the supernatant and remove any remaining liquid with a pipette
25. Dilute the pellet in 50 µl DEPC water

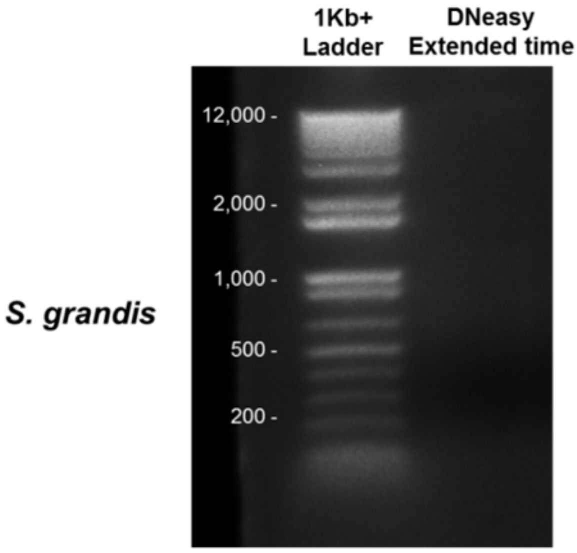
Day 2 - PureLink

# Prepare the lysis buffer freshly before the experiment: for each reaction add 4 µl of β-mercaptoethanol with 0.4 ml of Lysis buffer

26. Add 0.4 ml of freshly prepared Lysis buffer to the RNA sample
27. Vortex and incubate for 3 minutes
28. Add 0.2 ml ethanol (0.5 volumes)
29. Transfer up to 700 µl of the sample to the spin cartridge attached to the collection tube
30. Centrifuge at 11,000 rpm for 1 minute (>15 sec)
31. Discard the flow-through
32. Add 600 µl Wash buffer I
33. Centrifuge at 11,000 rpm for 1 minute (>15 sec)
34. Replace the collection tube with a clean one
35. Add 400µl Wash buffer II
36. Centrifuge at 11,000 rpm for 1 minute (>15sec)
37. Discard the flow-through
38. Repeat steps 35 - 37
39. Centrifuge at 11,000 rpm for 2 minutes
40. Add 15 - 30 µl RNase free water and incubate for 5 - 10 minutes
41. Centrifuge at 11,000 rpm for 2 minutes to collect the RNA
42. Keep the RNA at -80°C for long-term storage

Appendix 2.8

Gel electrophoresis result of *S. grandis* DNA extracted using DNeasy kit protocol



Appendix 2.9

Leaf disc weight measurement

Six leaf discs were freshly collected from *S. grandis* and *S. rexii.* The weights were measured and compared. The average weight per leaf disc for *S. grandis* is 0.0277 g (27.7 mg), and for *S. rexii* is 0.0433 g (43.3 mg). The difference between the average leaf disc weight is statistically different (Unpaired *t* test, d.f. = 10, *P*-value = 0.0006).

Table. Average fresh weight of leaf disc

Species	N	Average (g)	Median	Standard deviation
S. grandis	6	0.0277	0.0283	0.0042
S. rexii	6	0.0433	0.0428	0.0065

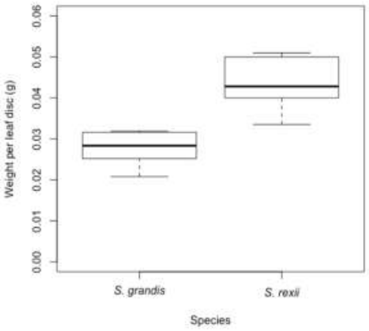


Figure. Box plot of the leaf disc weight of *S. grandis* and *S. rexii* (N = 6)

Appendix 3.1

Materials used for whole genome sequencing

Streptocarpus rexii

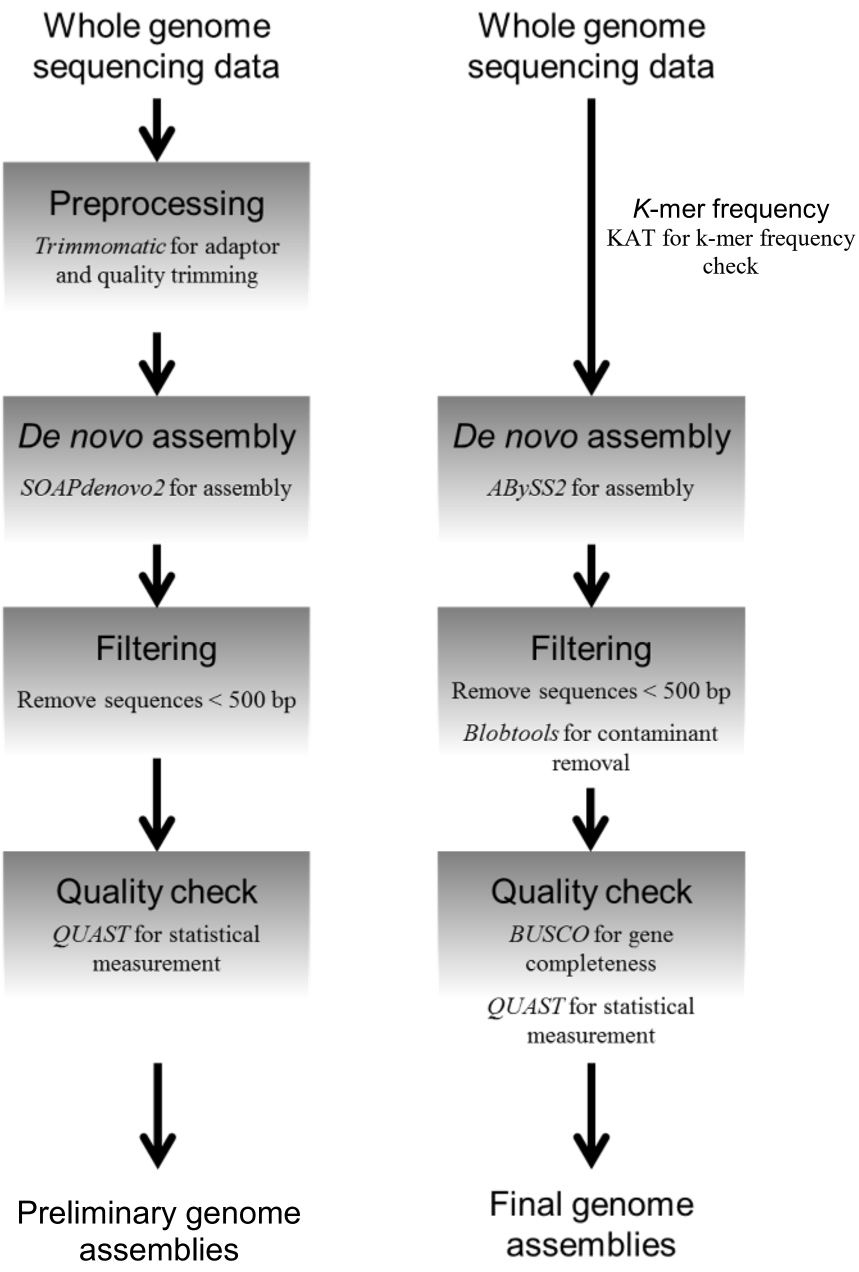
Accession	Qualifier	Used for genome seq	Lineage and collection information
20150819	A	Yes	Descendent of 19990270 by selfing
19990270	E	-	Descendent of 19870333 by selfing
19870333	N/A	-	Collector: Jong, K. (Collector no. K JNG1226) Collection date: 29th October 1986 Collection location: Grahamstown, 'Faraway' Estate, South Eastern Cape Province, South Africa

Streptocarpus grandis

Accession	Qualifier	Used for genome seq	Lineage and collection information
20150821	A	Yes	Descendent of 20130764 by selfing
20130764	A	-	Descendent of 20120713 by selfing
20120713	A	-	Descendent of 19771210 by selfing
19771210	A	-	Collector: Hilliard, O. and Burtt, B. L. (Collector no. HBT5923) Collection date: March 1977 Collection location: Ngome forest, Zululand, Natal Province, South Africa

Appendix 3.2

Flowchart of the genome assembly, filtering and analysis procedures


Whole genome sequencing data
Preprocessing
Trimmomatic for adaptor and quality trimming
De novo assembly
SOAPdenovo2 for assembly
Filtering
Remove sequences < 500 bp
Quality check
QUAST for statistical measurement
Preliminary genome assemblies
Whole genome sequencing data
K-mer frequency
KAT for k-mer frequency check
De novo assembly
ABySS2 for assembly
Filtering
Remove sequences < 500 bp
Blobtools for contaminant removal
Quality check
BUSCO for gene completeness
QUAST for statistical measurement
Final genome assemblies

Appendix 3.3

List of contaminant species identified in the *S. rexii* genome assembly

Phylum Actinobacteria

Acidipropionibacterium acidipropionici, *Actinoalloteichus hoggarensis*, *Actinoplanes missouriensis*, *Actinoplanes* sp., *Actinosynnema mirum*, *Actinosynnema pretiosum*, *Aeromicrobium erythreum*, *Agromyces aureus*, *Agromyces flavus*, *Amycolatopsis mediterranei*, *Amycolatopsis methanolica*, *Amycolatopsis orientalis*, *Auraticoccus monumenti*, *Beutenbergia cavernae*, *Brachybacterium* sp., *Cellulomonas gilvus*, *Clavibacter capsici*, *Clavibacter insidiosus*, *Clavibacter michiganensis*, *Clavibacter sepedonicus*, *Cnuibacter physcomitrellae*, *Corynebacterium doosanense*, *Corynebacterium frankenforstense*, *Cryobacterium arcticum*, *Cryobacterium* sp., *Cupriavidus necator*, *Curtobacterium pusillum*, *Curtobacterium* sp., *Frankia* sp., *Friedmanniella luteola*, *Friedmanniella sagamiharensis*, *Frondihabitans* sp., *Geodermatophilus obscurus*, *Gordonia bronchialis*, *Gordonia* sp., *Intrasporangium calvum*, *Jatrophihabitans* sp., *Jiangella alkaliphila*, *Jiangella* sp., *Kocuria palustris*, *Kribbella flavida*, *Leifsonia* sp., *Leifsonia xyli*, *Microbacterium aurum*, *Microbacterium pygmaeum*, *Microbacterium* sp., *Microcella alkaliphila*, *Micrococcus* sp., *Micromonospora echinaurantiaca*, *Micromonospora echinospora*, *Micromonospora inositola*, *Microterricola viridarii*, *Mycobacterium abscessus*, *Mycobacterium aurum*, *Mycobacterium avium*, *Mycobacterium bovis*, Mycobacterium canettii, *Mycobacterium chelonae*, *Mycobacterium chimaera*, *Mycobacterium chubuense*, *Mycobacterium colombiense*, *Mycobacterium dioxanotrophicus*, *Mycobacterium fortuitum*, *Mycobacterium gilvum*, *Mycobacterium goodii*, *Mycobacterium haemophilum*, *Mycobacterium immunogenum*, My*cobacterium indicus*, *Mycobacterium intracellulare*, *Mycobacterium kansasii*, *Mycobacterium leprae*, *Mycobacterium lepraemurium*, *Mycobacterium liflandii*, *Mycobacterium litorale*, *Mycobacterium marinum*, *Mycobacterium marseillense*, *Mycobacterium paraintracellulare*, *Mycobacterium phlei*, *Mycobacterium rhodesiae*, *Mycobacterium rutilum*, *Mycobacterium shigaense*, *Mycobacterium simiae*, *Mycobacterium sinense*, *Mycobacterium smegmatis*, *Mycobacterium* sp., *Mycobacterium stephanolepidis*, *Mycobacterium terrae*, *Mycobacterium thermoresistibile*, *Mycobacterium tuberculosis*, *Mycobacterium ulcerans*, *Mycobacterium vaccae*, *Mycobacterium vanbaalenii*, *Mycobacterium yongonense*, *Nakamurella multipartita*, *Nocardia asteroides*, *Nocardia brasiliensis*, *Nocardia cyriacigeorgica*, *Nocardia farcinica*, *Nocardia nova*, *Nocardia seriolae*, *Nocardia soli*, *Nocardiopsis dassonvillei*, *Nonomuraea gerenzanensis*, *Nonomuraea* sp., *Plantibacter flavus*, *Propionibacterium freudenreichii*, *Pseudomonas* sp., *Pseudonocardia dioxanivorans*, *Pseudonocardia* sp., *Rathayibacter tritici*, *Rhodococcus hoagii*, *Rhodococcus opacus*, *Rhodococcus rhodochrous*, *Rhodococcus ruber*, *Rhodococcus* sp., *Saccharothrix espanaensis*, *Sinomonas atrocyanea*, *Streptomyces albulus*, *Streptomyces albus*, *Streptomyces cattleya*, *Streptomyces chartreusis*, *Streptomyces coelicolor*, *Streptomyces griseochromogenes*, *Streptomyces hygroscopicus*, *Streptomyces parvulus*, *Streptomyces rapamycinicus*, *Streptomyces* sp., *Streptomyces venezuelae*, *Streptosporangium roseum*, *Tessaracoccus flavus*, *Thermobispora bispora*, uncultured *Mycobacterium*

Phylum Apicomplexa

Toxoplasma gondii

Phylum Arthropoda

Culicoides sonorensis, *Dendroctonus ponderosae*, *Nasonia vitripennis*, *Odontocepheus oblongus*, *Platynothrus peltifer*, *Zeugodacus cucurbitae*

Phylum Ascomycota

Aspergillus fumigatus, *Aspergillus nidulans*, *Aspergillus oryzae*, *Dandida albicans*, *Dandida dubliniensis*, *Dandida parapsilosis*, *Dapnobotryella renispora*, *Dhaetothyriales sp.*, *Dladophialophora bantiana*, *Exophiala mesophila*, *Kluyveromyces lactis*, *Naumovozyma castellii*, *Neurospora crassa*, *Parapenidiella pseudotasmaniensis*, *Phialophora verrucosa*, *Pseudocercospora mori*, *Saccharomycopsis fibuligera*, *Suhomyces tanzawaensis*, *Talaromyces marneffei*, *Talaromyces stipitatus*, *Tropicoporus linteus*, *Zasmidium cellare*, *Zymoseptoria tritici*

Phylum Bacteroidetes

Uncultured *Porphyromonas*

Phylum Basidiomycota

Cryptococcus gattii, *Cryptococcus neoformans*, *Exobasidium pachysporum*, *Fibroporia vaillantii*, *Geminibasidium donsium*, *Kalmanozyma brasiliensis*, *Kwoniella bestiolae*, *Kwoniella mangrovensis*, *Kwoniella pini*, *Malassezia sympodialis*, *Melanopsichium pennsylvanicum*, *Moesziomyces bullatus*, *Pseudozyma* sp., *Saitozyma ninhbinhensis*, *Sporisorium scitamineum*, *Tilletia laevis*, *Tremella fuciformis, Ustilago bromivora, Ustilago esculenta, Ustilago maydis*, *Wallemia mellicola*

Phylum Chlorophyta

Edaphochlorella mirabilis, *Stichococcus bacillaris*

Phylum Chordata

Cyprinus carpio, Oryzias *latipes*

Phylum Mucoromycota

Lichtheimia ramose, *Mucoromycota* sp., *Rhizopus microsporus*

Phylum Nematoda

Aphelenchoides fragariae, *Heterakis gallinarum*

Phylum Proteobacteria

Achromobacter spanius, Achromobacter xylosoxidans, Acidovorax citrulli, Alicycliphilus denitrificans, Aminobacter aminovorans, Amycolatopsis mediterranei, Archangium gephyra, Aureimonas sp.*, Azorhizobium caulinodans, Azospirillum brasilense, Azospirillum lipoferum, Azospirillum thiophilum, Azotobacter chroococcum, Beijerinckia indica, Blastochloris viridis, Bordetella bronchialis, Bordetella bronchiseptica, Bordetella hinzii, Bordetella pertussis, Bosea* sp.*, Bosea vaviloviae, Bradyrhizobium canariense, Bradyrhizobium diazoefficiens, Bradyrhizobium erythrophlei, Bradyrhizobium japonicum, Bradyrhizobium oligotrophicum, Bradyrhizobium* sp.*, Brucella vulpis, Burkholderia ambifaria, Burkholderia cenocepacia, Burkholderia cepacia, Burkholderia gladioli, Burkholderia glumae, Burkholderia lata, Burkholderia mallei, Burkholderia multivorans, Burkholderia oklahomensis, Burkholderia pyrrocinia, Burkholderia* sp.*, Burkholderia stabilis, Burkholderia territorii, Burkholderia thailandensis, Burkholderia ubonensis, Burkholderia vietnamiensis, Castellaniella defragrans, Caulobacter henricii, Caulobacter mirabilis, Caulobacter* sp.*, Caulobacter vibrioides, Chelatococcus daeguensis, Chelatococcus* sp.*, Chromobacterium vaccinii, Chromobacterium violaceum, Cupriavidus oxalaticus, Desulfovibrio vulgaris, Devosia* sp.*, Dokdonella koreensis, Dyella japonica, Dyella jiangningensis, Dyella* sp.*, Dyella thiooxydans, Ensifer adhaerens, Escherichia coli*, *Frateuria aurantia, Gluconacetobacter diazotrophicus, Granulibacter bethesdensis, Luteibacter rhizovicinus, Luteitalea pratensis, Lysobacter antibioticus, Lysobacter capsici, Lysobacter enzymogenes, Lysobacter gummosus, Magnetospirillum* sp.,

Mannheimia haemolytica, Martelella mediterranea, Massilia sp.*, Mesorhizobium amorphae, Mesorhizobium australicum, Mesorhizobium ciceri, Mesorhizobium loti, Methylibium petroleiphilum, Methylobacterium aquaticum, Methylobacterium extorquens, Methylobacterium nodulans, Methylobacterium oryzae, Methylobacterium phyllosphaerae, Methylobacterium populi, Methylobacterium radiotolerans, Methylobacterium* sp.*, Methylobacterium zatmanii, Methylocella silvestris, Methylocystis bryophila, Methylocystis* sp.*, Methylophilus* sp.*, Methylosinus trichosporium, Micromonospora narathiwatensis, Microvirga ossetica*, *Nitrobacter winogradskyi*, *Oligotropha carboxidovorans, Pandoraea pnomenusa, Pandoraea thiooxydans, Pantholops hodgsonii, Paraburkholderia aromaticivorans, Paraburkholderia caribensis, Paraburkholderia fungorum, Paraburkholderia hospita, Paraburkholderia phymatum, Paraburkholderia phytofirmans, Paraburkholderia sprentiae, Paraburkholderia xenovorans, Polaromonas* sp.*, Polymorphum gilvum, Pseudomonas aeruginosa, Pseudomonas citronellolis, Pseudomonas putida, Pseudomonas stutzeri, Pseudomonas veronii, Pseudoxanthomonas spadix, Pseudoxanthomonas suwonensis, Ralstonia mannitolilytica, Ralstonia solanacearum, Rheinheimera* sp.*, Rhizobium gallicum, Rhizobium leguminosarum, Rhizobium phaseoli, Rhizobium* sp.*, Rhodanobacter denitrificans, Rhodobacter* sp.*, Rhodomicrobium vannielii, Rhodoplanes* sp.*, Rhodopseudomonas palustris, Rhodospirillum rubrum, Roseomonas gilardii, Rubrivivax gelatinosus, Ruegeria pomeroyi, Shinella* sp.*, Sinorhizobium americanum, Sinorhizobium fredii, Sinorhizobium meliloti, Sphingomonas melonis, Sphingomonas panacis, Sphingomonas wittichii, Sphingopyxis macrogoltabida, Starkeya novella, Stenotrophomonas acidaminiphila, Stenotrophomonas maltophilia, Stenotrophomonas rhizophila, Stenotrophomonas* sp.*, Steroidobacter denitrificans*, *Verminephrobacter eiseniae, Xanthomonas albilineans, Xanthomonas campestris, Xanthomonas citri, Xanthomonas fuscans, Xanthomonas oryzae, Xanthomonas sacchari, Xanthomonas translucens, Pseudomonas mesoacidophila*, uncultured *Pseudomonas*, uncultured *Shewanella,* uncultured bacterium
Undefined eukaryota
Peronospora tabacina, *Pythium ultimum*, uncultured *alveolate*, uncultured eukaryote
Undefined bacteria
Uncultured bacterium
Viruses undefined
Dahlia mosaic, Escherichia virus, Mycobacterium phage

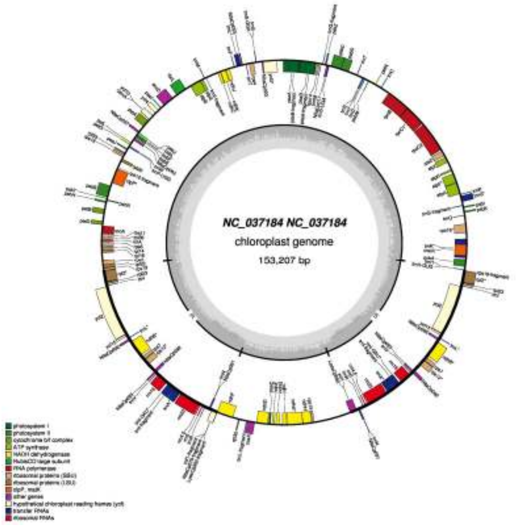
Appendix 3.4

List of contaminant species identified in the *S. grandis* genome assembly

Phylum Chordata
Cyanistes caeruleus
Phylum Proteobacteria
Alcaligenes faecalis, *Aquaspirillum* sp., *Beggiatoa leptomitoformis*, *Bordetella holmesii*, *Bordetella* sp., *Bradyrhizobium erythrophlei*, *Candidatus Methylopumilus*, *Chelatococcus daeguensis*, *Collimonas arenae*, *Collimonas pratensis*, *Dechloromonas aromatic*, *Escherichia coli*, *Gallionella capsiferriformans*, *Herbaspirillum seropedicae*, *Herminiimonas arsenitoxidans*, *Janthinobacterium agaricidamnosum*, *Laribacter hongkongensis*, *Methylobacillus flagellates*, *Methylobacterium nodulans*, *Methylomonas clara*, *Methylomonas* sp., *Methylophaga nitratireducenticrescens*, *Methylophilus methylotrophus*, *Methylophilus* sp., *Methylotenera mobilis*, *Methylotenera versatilis*, *Methylovorus* sp., *Moraxella osloensis*, *Neisseria meningitides*, *Nitrobacter winogradskyi*, *Nitrosomonas* sp., *Oligotropha carboxidovorans*, *Oxalobacter formigenes*, *Pandoraea sputorum*, *Paracoccus yeei*, *Pectobacterium polaris*, *Pseudoalteromonas piscicida*, *Pseudomonas aeruginosa*, *Pseudomonas cichorii*, *Pseudomonas fluorescens*, *Ricinus communis*, *Salmonella enterica*, *Serratia marcescens*, *Shewanella baltica*, *Vibrio fluvialis*, *Xenorhabdus poinarii*, *Zhongshania aliphaticivorans*
Undefined Eukaryota
Environmental Viridiplantae, uncultured eukaryote
Undefined Eukaryota
Environmental Viridiplantae, uncultured eukaryote
Undefined viruses
Escherichia virus, *Streptomyces* phage

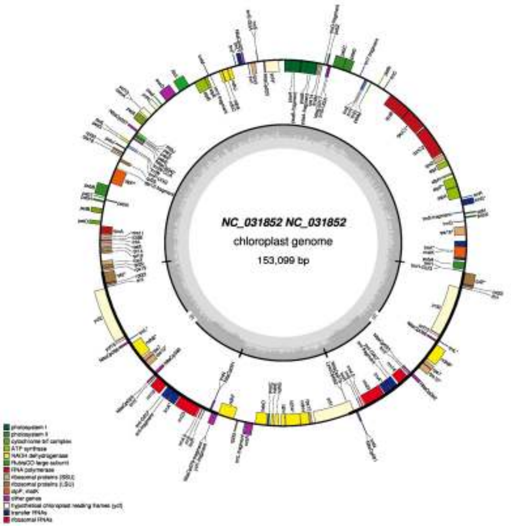
Appendix 3.5

Annotation of *Streptocarpus teitensis* and *Haberlea rhodopensis* chloroplasts
Annotation of both chloroplasts was carried out using GeSeq as described in section 3.2.3



Streptocarpus teitensis

rRNA	Location on the genome (bp, start location .. stop location)
rrn16	100,136..101,626
rrn23	104,068..106,880
rrn4.5	106,979..107,081
rrn5	107,306..107,426
rrn5	129,884..130,004 (complementary strand)
rrn4.5	130,229..130,331 (complementary strand)
rrn23	130,430..133,242 (complementary strand)
rrn16	135,684..137,174 (complementary strand)



Haberlea rhodopensis

rRNA	Location on the genome (bp, start location .. stop location)
rrn16	100,525..102,015
rrn23	104,444..107,253
rrn4.5	107,352..107,454
rrn5	107,679..107,799
rrn5	129,755..129,875 (complementary strand)
rrn4.5	130,100..130,202 (complementary strand)
rrn23	130,301..133,110 (complementary strand)
rrn16	135,539..137,029 (complementary strand)

Appendix 3.6

Annotation of *Dorcoceras hygrometricum* and *Erythrantes guttata* mitochondria

Annotation of both mitochondrias was carried out using GeSeq as described in section 3.2.3

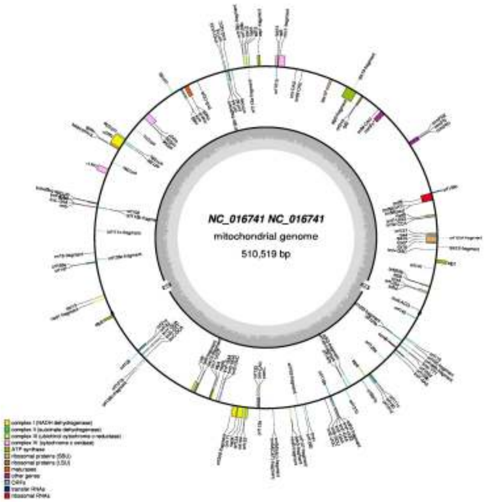


Figure. *D. hygrometricum* mitochondrion (NC_016741), 152 protein coding genes annotated

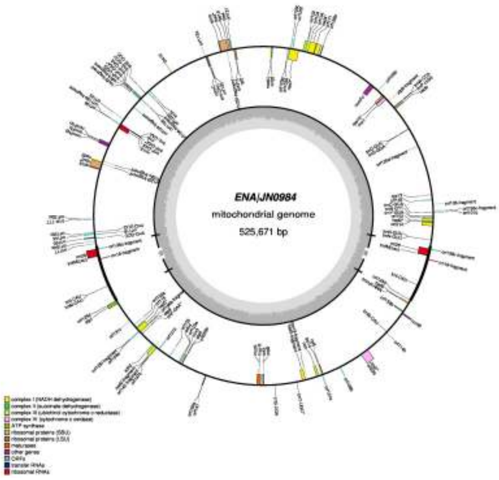


Figure. *E. guttata* mitochondrion (JN098455), 149 protein coding genes annotated

Appendix 4.1

Statistical summary of orthogroup analysis result in the *S. rexii* and *S. grandis* transcriptome assemblies

	S. rexii ***de novo* assembly**	***S. rexii*** **ref-guided assembly**	***S. grandis*** ***de novo* assembly**	***S. grandis*** **ref-guided assembly**
No. contigs	60500	53322	51267	46429
No. contigs assigned to orthogroups	51375	47580	45908	42555
No. unassigned contigs	9125	5742	5359	3874
Contigs assigned to orthogroups (%)	84.9	89.2	89.5	91.7
Unassigned contigs (%)	15.1	10.8	10.5	8.3
No. species-specific orthogroups	7	3	1	5
No. contigs in species-specific orthogroups	30	19	2	19
Contigs in species-specific orthogroups (%)	0	0	0	0

Appendix 5.1

(a) List of materials used for RAD-Seq, and the amount of data obtained before and after preprocessing

				Raw data		**After preprocessing**	
DNA ID	**Taxon**	**Accession No.**	**Qualifier**	**Total Bases**	**Read Count**	**Total Bases**	**Read Count**
YYD17	*S grandis*	20150821	C	312,506,427	6,127,577	119,192,300	2,383,846
YYD33	*S grandis*	20151810	S	220,846,932	4,330,332	83,450,250	1,669,005
YYD16	*S rexii*	20150819	A	118,086,624	2,315,424	32,638,150	652,763
YYD19	*S rexii*	19990270	I	99,823,116	1,957,316	34,431,200	688,624
YYD1001	(*S.grandis* × *rexii*) × *grandis*	20150825	A	3,912,822	76,722	1,033,600	20,672
YYD1002	(*S.grandis* × *rexii*) × *grandis*	20150825	B	1,600,329	31,379	269,900	5,398
YYD1003	(*S.grandis* × *rexii*) × *grandis*	20150825	C	4,157,367	81,517	1,314,250	26,285
YYD1004	(*S.grandis* × *rexii*) × *grandis*	20150825	D	19,539,681	383,131	9,128,850	182,577
YYD1005	(*S.grandis* × *rexii*) × *grandis*	20150825	E	138,372,741	2,713,191	62,472,750	1,249,455
YYD1006	(*S.grandis* × *rexii*) × *grandis*	20150825	F	83,970,888	1,646,488	26,162,750	523,255
YYD1007	(*S.grandis* × *rexii*) × *grandis*	20150825	G	85,332,282	1,673,182	44,084,000	881,680
YYD1008	(*S.grandis* × *rexii*) × *grandis*	20150825	H	19,406,061	380,511	7,641,900	152,838
YYD1010	(*S.grandis* × *rexii*) × *grandis*	20150825	J	76,875,411	1,507,361	35,307,250	706,145
YYD1011	(*S.grandis* × *rexii*) × *grandis*	20150825	K	55,234,122	1,083,022	20,714,650	414,293
YYD1012	(*S.grandis* × *rexii*) × *grandis*	20150825	L	37,202,205	729,455	8,910,000	178,200
YYD1013	(*S.grandis* × *rexii*) × *grandis*	20150825	M	318,338,940	6,241,940	120,943,450	2,418,869
YYD1014	(*S.grandis* × *rexii*) × *grandis*	20150825	N	5,912,073	115,923	2,396,700	47,934
YYD1015	(*S.grandis* × *rexii*) × *grandis*	20150825	O	61,176,846	1,199,546	22,919,700	458,394
YYD1016	(*S.grandis* × *rexii*) × *grandis*	20150825	P	33,795,048	662,648	14,715,550	294,311

DNA ID	Taxon	Accession No.	Qualifier	Total Bases (raw)	Read Count (raw)	Total Bases (Preprocessed)	Read Count (Preprocessed)
YYD1017	(*S.grandis* × *rexii*) × *grandis*	20150825	Q	240,459,135	4,714,885	108,180,850	2,163,617
YYD1018	(*S.grandis* × *rexii*) × *grandis*	20150825	R	90,328,038	1,771,138	33,559,750	671,195
YYD1019	(*S.grandis* × *rexii*) × *grandis*	20150825	S	9,721,518	190,618	4,903,950	98,079
YYD1020	(*S.grandis* × *rexii*) × *grandis*	20150825	T	142,060,500	2,785,500	38,033,650	760,673
YYD1021	(*S.grandis* × *rexii*) × *grandis*	20150825	U	168,920,415	3,312,165	49,218,400	984,368
YYD1022	(*S.grandis* × *rexii*) × *grandis*	20150825	V	149,462,181	2,930,631	67,513,600	1,350,272
YYD1023	(*S.grandis* × *rexii*) × *grandis*	20150825	W	40,151,076	787,276	19,287,350	385,747
YYD1024	(*S.grandis* × *rexii*) × *grandis*	20150825	X	191,774,739	3,760,289	49,841,350	996,827
YYD1025	(*S.grandis* × *rexii*) × *grandis*	20150825	Y	19,918,764	390,564	10,803,650	216,073
YYD1026	(*S.grandis* × *rexii*) × *grandis*	20150825	Z	10,462,089	205,139	5,742,350	114,847
YYD1027	(*S.grandis* × *rexii*) × *grandis*	20150825	AA	137,677,560	2,699,560	57,162,100	1,143,242
YYD1028	(*S.grandis* × *rexii*) × *grandis*	20150825	AB	13,729,302	269,202	5,391,550	107,831
YYD1029	(*S.grandis* × *rexii*) × *grandis*	20150825	AC	20,805,858	407,958	8,960,200	179,204
YYD1030	(*S.grandis* × *rexii*) × *grandis*	20150825	AD	30,731,019	602,569	14,759,950	295,199
YYD1031	(*S.grandis* × *rexii*) × *grandis*	20150825	AE	127,202,568	2,494,168	51,506,600	1,030,132
YYD1032	(*S.grandis* × *rexii*) × *grandis*	20150825	AF	75,740,508	1,485,108	19,862,800	397,256
YYD1033	(*S.grandis* × *rexii*) × *grandis*	20150825	AG	8,989,413	176,263	3,890,300	77,806
YYD1034	(*S.grandis* × *rexii*) × *grandis*	20150825	AH	35,656,803	699,153	18,756,300	375,126
YYD1035	(*S.grandis* × *rexii*) × *grandis*	20150825	AI	18,826,140	369,140	9,287,600	185,752
YYD1036	(*S.grandis* × *rexii*) × *grandis*	20150825	AJ	114,254,841	2,240,291	16,474,650	329,493

DNA ID	Taxon	Accession No.	Qualifier	Total Bases (raw)	Read Count (raw)	Total Bases (Preprocessed)	Read Count (Preprocessed)
YYD1037	(*S.grandis* × *rexii*) × *grandis*	20150825	AK	28,448,820	557,820	10,757,200	215,144
YYD1038	(*S.grandis* × *rexii*) × *grandis*	20150825	AL	135,336,456	2,653,656	67,186,750	1,343,735
YYD1040	(*S.grandis* × *rexii*) × *grandis*	20150825	AN	39,906,021	782,471	17,137,650	342,753
YYD1041	(*S.grandis* × *rexii*) × *grandis*	20150825	AO	53,147,814	1,042,114	23,253,850	465,077
YYD1042	(*S.grandis* × *rexii*) × *grandis*	20150825	AP	165,504,027	3,245,177	67,327,250	1,346,545
YYD1043	(*S.grandis* × *rexii*) × *grandis*	20150825	AQ	109,334,412	2,143,812	56,704,600	1,134,092
YYD1044	(*S.grandis* × *rexii*) × *grandis*	20150825	AR	9,475,086	185,786	5,207,150	104,143
YYD1045	(*S.grandis* × *rexii*) × *grandis*	20150825	AS	40,705,854	798,154	17,668,350	353,367
YYD1046	(*S.grandis* × *rexii*) × *grandis*	20150825	AT	40,964,322	803,222	23,162,150	463,243
YYD1047	(*S.grandis* × *rexii*) × *grandis*	20150825	AU	23,431,491	459,441	12,327,550	246,551
YYD1048	(*S.grandis* × *rexii*) × *grandis*	20150825	AV	42,251,409	828,459	19,790,300	395,806
YYD1049	(*S.grandis* × *rexii*) × *grandis*	20150825	AW	55,059,957	1,079,607	20,850,300	417,006
YYD1050	(*S.grandis* × *rexii*) × *grandis*	20150825	AX	270,472,227	5,303,377	90,661,000	1,813,220
YYD1051	(*S.grandis* × *rexii*) × *grandis*	20150825	AY	9,536,949	186,999	4,026,200	80,524
YYD1052	(*S.grandis* × *rexii*) × *grandis*	20150825	AZ	73,666,950	1,444,450	25,675,150	513,503
YYD1053	(*S.grandis* × *rexii*) × *grandis*	20150825	BA	127,931,103	2,508,453	54,381,750	1,087,635
YYD1056	(*S.grandis* × *rexii*) × *grandis*	20150825	BD	135,333,294	2,653,594	38,058,000	761,160
YYD1058	(*S.grandis* × *rexii*) × *grandis*	20150825	BF	211,712,118	4,151,218	75,272,600	1,505,452
YYD1059	(*S.grandis* × *rexii*) × *grandis*	20150825	BG	81,889,680	1,605,680	25,917,900	518,358
YYD1060	(*S.grandis* × *rexii*) × *grandis*	20150825	BH	249,284,430	4,887,930	115,172,950	2,303,459

DNA ID	Taxon	Accession No.	Qualifier	Total Bases (raw)	Read Count (raw)	Total Bases (Preprocessed)	Read Count (Preprocessed)
YYD1061	(*S.grandis* × *rexii*) × *grandis*	20150825	BI	167,866,551	3,291,501	65,181,050	1,303,621
YYD1062	(*S.grandis* × *rexii*) × *grandis*	20150825	BJ	52,228,437	1,024,087	30,935,950	618,719
YYD1063	(*S.grandis* × *rexii*) × *grandis*	20150825	BK	54,243,600	1,063,600	23,959,600	479,192
YYD1064	(*S.grandis* × *rexii*) × *grandis*	20150825	BL	66,148,326	1,297,026	28,886,550	577,731
YYD1065	(*S.grandis* × *rexii*) × *grandis*	20150825	BM	15,757,266	308,966	7,881,250	157,625
YYD1066	(*S.grandis* × *rexii*) × *grandis*	20150825	BN	61,391,658	1,203,758	18,461,550	369,231
YYD1067	(*S.grandis* × *rexii*) × *grandis*	20150825	BO	100,624,377	1,973,027	33,326,950	666,539
YYD1069	(*S.grandis* × *rexii*) × *grandis*	20150825	BQ	21,978,297	430,947	9,996,050	199,921
YYD1070	(*S.grandis* × *rexii*) × *grandis*	20150825	BR	181,928,526	3,567,226	62,792,500	1,255,850
YYD1071	(*S.grandis* × *rexii*) × *grandis*	20150825	BS	84,470,076	1,656,276	29,470,600	589,412
YYD1072	(*S.grandis* × *rexii*) × *grandis*	20150825	BT	27,896,847	546,997	10,134,700	202,694
YYD1073	(*S.grandis* × *rexii*) × *grandis*	20150825	BU	84,691,161	1,660,611	28,400,600	568,012
YYD1074	(*S.grandis* × *rexii*) × *grandis*	20150825	BV	213,076,725	4,177,975	80,051,950	1,601,039
YYD1075	(*S.grandis* × *rexii*) × *grandis*	20150825	BW	37,865,205	742,455	20,669,350	413,387
YYD1076	(*S.grandis* × *rexii*) × *grandis*	20150825	BX	28,876,404	566,204	10,656,550	213,131
YYD1077	(*S.grandis* × *rexii*) × *grandis*	20150825	BY	42,124,572	825,972	17,329,250	346,585
YYD1078	(*S.grandis* × *rexii*) × *grandis*	20150825	BZ	185,661,930	3,640,430	81,810,700	1,636,214
YYD1079	(*S.grandis* × *rexii*) × *grandis*	20150825	CA	486,147,963	9,532,313	181,971,550	3,639,431
YYD1080	(*S.grandis* × *rexii*) × *grandis*	20150825	CB	111,675,924	2,189,724	40,156,800	803,136
YYD1081	(*S.grandis* × *rexii*) × *grandis*	20150825	CC	73,183,980	1,434,980	26,684,700	533,694

DNA ID	Taxon	Accession No.	Qualifier	Total Bases (raw)	Read Count (raw)	Total Bases (Preprocessed)	Read Count (Preprocessed)
YYD1082	(*S.grandis* × *rexii*) × *grandis*	20150825	CD	116,755,524	2,289,324	36,518,850	730,377
YYD1083	(*S.grandis* × *rexii*) × *grandis*	20150825	CE	221,016,660	4,333,660	88,413,000	1,768,260
YYD1084	(*S.grandis* × *rexii*) × *grandis*	20150825	CF	32,581,758	638,858	13,027,650	260,553
YYD1085	(*S.grandis* × *rexii*) × *grandis*	20150825	CG	158,930,943	3,116,293	65,525,600	1,310,512
YYD1086	(*S.grandis* × *rexii*) × *grandis*	20150825	CH	34,696,422	680,322	13,126,200	262,524
YYD1087	(*S.grandis* × *rexii*) × *grandis*	20150825	CI	199,554,891	3,912,841	86,746,850	1,734,937
YYD1088	(*S.grandis* × *rexii*) × *grandis*	20150825	CJ	80,734,683	1,583,033	32,349,400	646,988
YYD1089	(*S.grandis* × *rexii*) × *grandis*	20150825	CK	101,587,512	1,991,912	32,540,500	650,810
YYD1090	(*S.grandis* × *rexii*) × *grandis*	20150825	CL	292,302,828	5,731,428	123,615,800	2,472,316
YYD1091	(*S.grandis* × *rexii*) × *grandis*	20150825	CM	184,437,930	3,616,430	83,473,250	1,669,465
YYD1092	(*S.grandis* × *rexii*) × *grandis*	20150825	CN	12,223,272	239,672	5,017,500	100,350
YYD1093	(*S.grandis* × *rexii*) × *grandis*	20150825	CO	3,538,839	69,389	172,850	3,457
YYD1094	(*S.grandis* × *rexii*) × *grandis*	20150825	CP	45,060,336	883,536	17,543,050	350,861
YYD1095	(*S.grandis* × *rexii*) × *grandis*	20150825	CQ	56,440,731	1,106,681	27,067,950	541,359
YYD1096	(*S.grandis* × *rexii*) × *grandis*	20150825	CR	29,826,126	584,826	15,323,450	306,469
YYD1098	(*S.grandis* × *rexii*) × *grandis*	20150825	CT	94,231,221	1,847,671	40,868,900	817,378
YYD1099	(*S.grandis* × *rexii*) × *grandis*	20150825	CU	40,737,780	798,780	19,087,800	381,756
YYD1100	(*S.grandis* × *rexii*) × *grandis*	20150825	CV	25,544,268	500,868	11,886,500	237,730
YYD1101	(*S.grandis* × *rexii*) × *grandis*	20150825	CW	14,418,159	282,709	7,240,250	144,805
YYD1102	(*S.grandis* × *rexii*) × *grandis*	20150825	CX	274,179,417	5,376,067	76,873,150	1,537,463

DNA ID	Taxon	Accession No.	Qualifier	Total Bases (raw)	Read Count (raw)	Total Bases (Preprocessed)	Read Count (Preprocessed)
YYD1103	(*S.grandis* × *rexii*) × *grandis*	20150825	CY	110,732,526	2,171,226	35,739,850	714,797
YYD1104	(*S.grandis* × *rexii*) × *grandis*	20150825	CZ	5,607,399	109,949	833,900	16,678
YYD1105	(*S.grandis* × *rexii*) × *grandis*	20150825	DA	139,092,708	2,727,308	48,056,800	961,136
YYD1106	(*S.grandis* × *rexii*) × *grandis*	20150825	DB	111,340,752	2,183,152	42,815,700	856,314
YYD1107	(*S.grandis* × *rexii*) × *grandis*	20150825	DC	9,277,665	181,915	4,854,150	97,083
YYD1108	(*S.grandis* × *rexii*) × *grandis*	20150825	DD	38,192,880	748,880	19,190,900	383,818
YYD1109	(*S.grandis* × *rexii*) × *grandis*	20150825	DE	81,108,615	1,590,365	28,108,900	562,178
YYD1110	(*S.grandis* × *rexii*) × *grandis*	20150825	DF	194,473,149	3,813,199	59,054,650	1,181,093
YYD1111	(*S.grandis* × *rexii*) × *grandis*	20150825	DG	111,168,780	2,179,780	46,401,350	928,027
YYD1112	(*S.grandis* × *rexii*) × *grandis*	20150825	DH	292,349,748	5,732,348	81,701,600	1,634,032
YYD1113	(*S.grandis* × *rexii*) × *grandis*	20150825	DI	47,569,332	932,732	18,786,850	375,737
YYD1114	(*S.grandis* × *rexii*) × *grandis*	20150825	DJ	74,048,940	1,451,940	34,488,700	689,774
YYD1115	(*S.grandis* × *rexii*) × *grandis*	20150825	DK	1,871,496	36,696	842,600	16,852
YYD1116	(*S.grandis* × *rexii*) × *grandis*	20150825	DL	546,006	10,706	209,500	4,190
YYD1117	(*S.grandis* × *rexii*) × *grandis*	20150825	DM	51,162,486	1,003,186	16,978,600	339,572
YYD1118	(*S.grandis* × *rexii*) × *grandis*	20150825	DN	3,167,151	62,101	1,383,800	27,676
YYD1119	(*S.grandis* × *rexii*) × *grandis*	20150825	DO	7,143,468	140,068	3,035,400	60,708
YYD1120	(*S.grandis* × *rexii*) × *grandis*	20150825	DP	2,494,716	48,916	1,129,200	22,584
YYD1121	(*S.grandis* × *rexii*) × *grandis*	20150825	DQ	5,671,200	111,200	2,550,400	51,008
YYD1122	(*S.grandis* × *rexii*) × *grandis*	20150825	DR	5,976,435	117,185	2,959,850	59,197

DNA ID	Taxon	Accession No.	Qualifier	Total Bases (raw)	Read Count (raw)	Total Bases (Preprocessed)	Read Count (Preprocessed)
YYD1123	(*S.grandis* × *rexii*) × *grandis*	20150825	DS	41,305,665	809,915	16,450,550	329,011
YYD1124	(*S.grandis* × *rexii*) × *grandis*	20150825	DT	51,364,446	1,007,146	23,406,650	468,133
YYD1125	(*S.grandis* × *rexii*) × *grandis*	20150825	DU	14,736,705	288,955	6,235,450	124,709
YYD1127	(*S.grandis* × *rexii*) × *grandis*	20150825	DW	7,942,995	155,745	3,294,100	65,882
YYD1128	(*S.grandis* × *rexii*) × *grandis*	20150825	DX	27,799,029	545,079	11,845,800	236,916
YYD1129	(*S.grandis* × *rexii*) × *grandis*	20150825	DY	11,828,940	231,940	4,093,600	81,872
YYD1130	(*S.grandis* × *rexii*) × *grandis*	20150825	DZ	14,185,242	278,142	5,636,650	112,733
YYD1131	(*S.grandis* × *rexii*) × *grandis*	20150825	EA	8,442,285	165,535	3,839,500	76,790
YYD1132	(*S.grandis* × *rexii*) × *grandis*	20150825	EB	192,718,494	3,778,794	37,343,000	746,860
YYD1133	(*S.grandis* × *rexii*) × *grandis*	20150825	EC	17,845,104	349,904	8,718,950	174,379
YYD1134	(*S.grandis* × *rexii*) × *grandis*	20150825	ED	10,517,373	206,223	4,583,100	91,662
YYD1135	(*S.grandis* × *rexii*) × *grandis*	20150825	EE	91,672,602	1,797,502	33,457,400	669,148
YYD1136	(*S.grandis* × *rexii*) × *grandis*	20150825	EF	23,931,087	469,237	7,356,000	147,120
YYD1137	(*S.grandis* × *rexii*) × *grandis*	20150825	EG	36,713,370	719,870	13,406,950	268,139
YYD1138	(*S.grandis* × *rexii*) × *grandis*	20150825	EH	186,869,712	3,664,112	67,481,200	1,349,624
YYD1139	(*S.grandis* × *rexii*) × *grandis*	20150825	EI	5,312,772	104,172	2,594,650	51,893
YYD1140	(*S.grandis* × *rexii*) × *grandis*	20150825	EJ	7,342,470	143,970	2,633,450	52,669
YYD1141	(*S.grandis* × *rexii*) × *grandis*	20150825	EK	10,973,160	215,160	5,006,100	100,122
YYD1142	(*S.grandis* × *rexii*) × *grandis*	20150825	EL	1,628,022	31,922	818,050	16,361
YYD1143	(*S.grandis* × *rexii*) × *grandis*	20150825	EM	2,773,176	54,376	612,800	12,256

DNA ID	Taxon	Accession No.	Qualifier	Total Bases (raw)	Read Count (raw)	Total Bases (Preprocessed)	Read Count (Preprocessed)
YYD1144	(*S.grandis* × *rexii*) × *grandis*	20150825	EN	30,578,937	599,587	14,421,750	288,435
YYD1145	(*S.grandis* × *rexii*) × *grandis*	20150825	EO	483,344,595	9,477,345	169,305,950	3,386,119
YYD1146	(*S.grandis* × *rexii*) × *grandis*	20150825	EP	6,497,196	127,396	3,062,250	61,245
YYD1147	(*S.grandis* × *rexii*) × *grandis*	20150825	EQ	10,763,040	211,040	4,973,800	99,476
YYD1148	(*S.grandis* × *rexii*) × *grandis*	20150825	ER	7,228,536	141,736	2,408,550	48,171
YYD1149	(*S.grandis* × *rexii*) × *grandis*	20150825	ES	66,071,826	1,295,526	35,476,800	709,536
YYD1150	(*S.grandis* × *rexii*) × *grandis*	20150825	ET	205,599,258	4,031,358	80,476,700	1,609,534
YYD1151	(*S.grandis* × *rexii*) × *grandis*	20150825	EU	184,422,477	3,616,127	48,961,600	979,232
YYD1152	(*S.grandis* × *rexii*) × *grandis*	20150825	EV	15,942,447	312,597	6,248,900	124,978
YYD1153	(*S.grandis* × *rexii*) × *grandis*	20150825	EW	142,576,824	2,795,624	37,805,800	756,116
YYD1154	(*S.grandis* × *rexii*) × *grandis*	20150825	EX	18,075,930	354,430	6,651,750	133,035
YYD1155	(*S.grandis* × *rexii*) × *grandis*	20150825	EY	205,922,190	4,037,690	64,165,850	1,283,317
YYD1156	(*S.grandis* × *rexii*) × *grandis*	20150825	EZ	27,866,706	546,406	9,888,700	197,774
YYD1157	(*S.grandis* × *rexii*) × *grandis*	20150825	FA	12,184,716	238,916	5,604,100	112,082
YYD1158	(*S.grandis* × *rexii*) × *grandis*	20150825	FB	10,628,706	208,406	4,284,250	85,685
YYD1159	(*S.grandis* × *rexii*) × *grandis*	20150825	FC	13,950,591	273,541	4,302,450	86,049
YYD1160	(*S.grandis* × *rexii*) × *grandis*	20150825	FD	23,614,581	463,031	8,833,600	176,672
YYD1161	(*S.grandis* × *rexii*) × *grandis*	20150825	FE	17,254,320	338,320	6,006,700	120,134
YYD1162	(*S.grandis* × *rexii*) × *grandis*	20150825	FF	32,017,494	627,794	11,714,850	234,297
YYD1163	(*S.grandis* × *rexii*) × *grandis*	20150825	FG	94,249,581	1,848,031	33,364,200	667,284

DNA ID	Taxon	Accession No.	Qualifier	Total Bases (raw)	Read Count (raw)	Total Bases (Preprocessed)	Read Count (Preprocessed)
YYD1164	(*S.grandis* × *rexii*) × *grandis*	20150825	FH	310,167,312	6,081,712	126,140,300	2,522,806
YYD1165	(*S.grandis* × *rexii*) × *grandis*	20150825	FI	27,757,311	544,261	12,353,500	247,070
YYD1166	(*S.grandis* × *rexii*) × *grandis*	20150825	FJ	8,012,049	157,099	4,042,250	80,845
YYD1167	(*S.grandis* × *rexii*) × *grandis*	20150825	FK	21,590,085	423,335	8,998,750	179,975
YYD1168	(*S.grandis* × *rexii*) × *grandis*	20150825	FL	30,350,304	595,104	11,295,350	225,907
YYD1169	(*S.grandis* × *rexii*) × *grandis*	20150825	FM	14,135,517	277,167	3,396,800	67,936
YYD1170	(*S.grandis* × *rexii*) × *grandis*	20150825	FN	109,648,725	2,149,975	29,023,550	580,471
YYD1171	(*S.grandis* × *rexii*) × *grandis*	20150825	FO	16,519,665	323,915	7,137,150	142,743
YYD1172	(*S.grandis* × *rexii*) × *grandis*	20150825	FP	5,228,367	102,517	2,955,650	59,113
YYD1173	(*S.grandis* × *rexii*) × *grandis*	20150825	FQ	6,785,652	133,052	3,061,950	61,239
YYD1174	(*S.grandis* × *rexii*) × *grandis*	20150825	FR	36,968,574	724,874	15,000,950	300,019
YYD1175	(*S.grandis* × *rexii*) × *grandis*	20150825	FS	14,612,010	286,510	6,374,100	127,482
YYD1176	(*S.grandis* × *rexii*) × *grandis*	20150825	FT	9,994,725	195,975	3,794,000	75,880
YYD1177	(*S.grandis* × *rexii*) × *grandis*	20150825	FU	119,514,930	2,343,430	29,273,050	585,461
YYD1178	(*S.grandis* × *rexii*) × *grandis*	20150825	FV	112,144,359	2,198,909	54,624,850	1,092,497
YYD1180	(*S.grandis* × *rexii*) × *grandis*	20150825	FX	15,026,844	294,644	6,694,250	133,885
YYD1181	(*S.grandis* × *rexii*) × *grandis*	20150825	FY	119,809,302	2,349,202	36,633,350	732,667
YYD1183	(*S.grandis* × *rexii*) × *grandis*	20150825	GA	21,536,076	422,276	9,650,000	193,000
YYD1184	(*S.grandis* × *rexii*) × *grandis*	20150825	GB	31,931,049	626,099	11,911,200	238,224
YYD1185	(*S.grandis* × *rexii*) × *grandis*	20150825	GC	223,307,274	4,378,574	63,450,200	1,269,004

DNA ID	Taxon	Accession No.	Qualifier	Total Bases (raw)	Read Count (raw)	Total Bases (Preprocessed)	Read Count (Preprocessed)
YYD1188	(*S.grandis* × *rexii*) × *grandis*	20150825	GF	23,446,944	459,744	10,995,200	219,904
YYD1189	(*S.grandis* × *rexii*) × *grandis*	20150825	GG	242,056,455	4,746,205	58,370,400	1,167,408
YYD1190	(*S.grandis* × *rexii*) × *grandis*	20150825	GH	142,536,330	2,794,830	60,443,200	1,208,864
YYD1191	(*S.grandis* × *rexii*) × *grandis*	20150825	GI	47,192,187	925,337	12,266,400	245,328
YYD1192	(*S.grandis* × *rexii*) × *grandis*	20150825	GJ	35,802,459	702,009	16,243,750	324,875
YYD1193	(*S.grandis* × *rexii*) × *grandis*	20150825	GK	92,502,729	1,813,779	38,733,750	774,675
YYD1194	(*S.grandis* × *rexii*) × *grandis*	20150825	GL	5,611,479	110,029	1,698,850	33,977
YYD1195	(*S.grandis* × *rexii*) × *grandis*	20150825	GM	143,617,071	2,816,021	42,027,650	840,553
YYD1196	(*S.grandis* × *rexii*) × *grandis*	20150825	GN	24,900,189	488,239	9,832,600	196,652
YYD1197	(*S.grandis* × *rexii*) × *grandis*	20150825	GO	118,473,102	2,323,002	37,770,650	755,413
YYD1198	(*S.grandis* × *rexii*) × *grandis*	20150825	GP	47,556,990	932,490	19,524,000	390,480
YYD1199	(*S.grandis* × *rexii*) × *grandis*	20150825	GQ	431,722,191	8,465,141	171,791,400	3,435,828
YYD1200	(*S.grandis* × *rexii*) × *grandis*	20150825	GR	45,186,204	886,004	14,554,150	291,083
YYD1201	(*S.grandis* × *rexii*) × *grandis*	20150825	GS	126,339,291	2,477,241	42,255,400	845,108
YYD1203	(*S.grandis* × *rexii*) × *grandis*	20150825	GU	11,557,059	226,609	4,639,700	92,794
YYD1204	(*S.grandis* × *rexii*) × *grandis*	20150825	GV	153,200,430	3,003,930	47,538,000	950,760
YYD1205	(*S.grandis* × *rexii*) × *grandis*	20150825	GW	25,366,890	497,390	10,616,500	212,330
YYD1206	(*S.grandis* × *rexii*) × *grandis*	20150825	GX	67,899,666	1,331,366	21,341,750	426,835
YYD1207	(*S.grandis* × *rexii*) × *grandis*	20150825	GY	41,944,287	822,437	14,508,450	290,169
YYD1208	(*S.grandis* × *rexii*) × *grandis*	20150825	GZ	37,442,568	734,168	13,849,250	276,985

DNA ID	Taxon	Accession No.	Qualifier	Total Bases (raw)	Read Count (raw)	Total Bases (Preprocessed)	Read Count (Preprocessed)
YYD1209	(*S.grandis* × *rexii*) × *grandis*	20150825	HA	31,766,472	622,872	10,109,300	202,186
YYD1210	(*S.grandis* × *rexii*) × *grandis*	20150825	HB	152,959,710	2,999,210	60,450,900	1,209,018
YYD1211	(*S.grandis* × *rexii*) × *grandis*	20150825	HC	68,802,162	1,349,062	20,893,300	417,866
YYD1212	(*S.grandis* × *rexii*) × *grandis*	20150825	HD	42,265,638	828,738	18,431,400	368,628
YYD1213	(*S.grandis* × *rexii*) × *grandis*	20150825	HE	22,158,735	434,485	9,267,050	185,341
YYD1214	(*S.grandis* × *rexii*) × *grandis*	20150825	HF	100,445,520	1,969,520	34,413,600	688,272
YYD1215	(*S.grandis* × *rexii*) × *grandis*	20150825	HG	402,140,253	7,885,103	153,480,850	3,069,617
YYD1216	(*S.grandis* × *rexii*) × *grandis*	20150825	HH	28,111,353	551,203	10,446,400	208,928
YYD1217	(*S.grandis* × *rexii*) × *grandis*	20150825	HI	75,722,505	1,484,755	27,336,000	546,720
YYD1218	(*S.grandis* × *rexii*) × *grandis*	20150825	HJ	65,964,675	1,293,425	18,461,350	369,227
YYD1219	(*S.grandis* × *rexii*) × *grandis*	20150825	HK	117,961,878	2,312,978	50,395,950	1,007,919
YYD1220	(*S.grandis* × *rexii*) × *grandis*	20150825	HL	7,851,399	153,949	2,636,900	52,738
YYD1221	(*S.grandis* × *rexii*) × *grandis*	20150825	HM	57,692,526	1,131,226	19,172,050	383,441
YYD1224	(*S.grandis* × *rexii*) × *grandis*	20150825	HP	29,078,823	570,173	8,580,600	171,612
YYD1226	(*S.grandis* × *rexii*) × *grandis*	20150825	HR	59,718,195	1,170,945	27,803,000	556,060
YYD1227	(*S.grandis* × *rexii*) × *grandis*	20150825	HS	120,454,962	2,361,862	37,522,050	750,441
YYD1228	(*S.grandis* × *rexii*) × *grandis*	20150825	HT	207,001,911	4,058,861	73,190,000	1,463,800
YYD1229	(*S.grandis* × *rexii*) × *grandis*	20150825	HU	161,189,274	3,160,574	82,588,000	1,651,760
YYD1230	(*S.grandis* × *rexii*) × *grandis*	20150825	HV	748,942,191	14,685,141	363,067,650	7,261,353
YYD1231	(*S.grandis* × *rexii*) × *grandis*	20150825	HW	78,133,377	1,532,027	32,001,900	640,038

DNA ID	**Taxon**	**Accession No.**	**Qualifier**	**Total Bases (raw)**	**Read Count (raw)**	**Total Bases (Preprocessed)**	**Read Count (Preprocessed)**
YYD1232	(*S.grandis* × *rexii*) × *grandis*	20150825	HX	70,569,006	1,383,706	27,848,150	556,963
YYD1233	(*S.grandis* × *rexii*) × *grandis*	20150825	HY	97,131,999	1,904,549	37,817,300	756,346
YYD1234	(*S.grandis* × *rexii*) × *grandis*	20150825	HZ	84,620,679	1,659,229	31,035,200	620,704
YYD1235	(*S.grandis* × *rexii*) × *grandis*	20150825	IA	16,090,602	315,502	6,225,200	124,504
YYD1236	(*S.grandis* × *rexii*) × *grandis*	20150825	IB	51,583,899	1,011,449	19,056,750	381,135
YYD1237	(*S.grandis* × *rexii*) × *grandis*	20150825	IC	12,727,254	249,554	5,034,850	100,697
YYD1238	(*S.grandis* × *rexii*) × *grandis*	20150825	ID	24,125,142	473,042	8,488,200	169,764
YYD1239	(*S.grandis* × *rexii*) × *grandis*	20150825	IE	24,016,104	470,904	7,228,950	144,579
YYD1240	(*S.grandis* × *rexii*) × *grandis*	20150825	IF	37,106,835	727,585	17,560,650	351,213
YYD1241	(*S.grandis* × *rexii*) × *grandis*	20150825	IG	170,383,605	3,340,855	47,431,450	948,629
YYD1242	(*S.grandis* × *rexii*) × *grandis*	20150825	IH	43,762,641	858,091	20,118,000	402,360
YYD1244	(*S.grandis* × *rexii*) × *grandis*	20150825	IJ	48,816,027	957,177	21,037,600	420,752
YYD1245	(*S.grandis* × *rexii*) × *grandis*	20150825	IK	44,690,433	876,283	16,859,650	337,193
YYD1246	(*S.grandis* × *rexii*) × *grandis*	20150825	IL	91,771,644	1,799,444	25,458,450	509,169
YYD1250	(*S.grandis* × *rexii*) × *grandis*	20150825	IP	22,193,160	435,160	10,188,700	203,774
YYD1251	(*S.grandis* × *rexii*) × *grandis*	20150825	IQ	20,825,238	408,338	8,963,700	179,274
YYD1252	(*S.grandis* × *rexii*) × *grandis*	20150825	IR	30,886,263	605,613	9,612,900	192,258
YYD1253	(*S.grandis* × *rexii*) × *grandis*	20150825	IS	83,329,053	1,633,903	38,189,100	763,782

(b) List of materials used for Stacks analyses and genetic map calculation. The DNA ID listed here are corresponded to the material details summarised in Appendix 5.1. v: used for analyses. -: not used for analyses.

	Used for MapA calculation					**Used for MapB calculation**			
DNA ID	**Parameter optimisation (50 indi.)**	***De novo* approach (150 indi.)**	***De novo* approach (200 indi.)**	**Ref-based approach with BWA (200 indi.)**	**Ref-based approach with Stampy (200 indi.)**	**Parameter optimisation (50 indi.)**	***De novo* approach (200 indi.)**	**Ref-based approach with BWA (200 indi.)**	**Ref-based approach with Stampy (200 indi.)**
YYD17	v	v	v	v	v	v	v	v	v
YYD33	-	-	-	-	-	-	-	-	-
YYD16	v	v	v	v	v	v	v	v	v
YYD19	v	v	v	v	v	v	v	v	v
YYD1001	-	-	-	-	-	-	-	-	-
YYD1002	-	-	-	-	-	-	-	-	-
YYD1003	-	-	-	-	-	-	-	-	-
YYD1004	-	-	v	v	v	-	v	v	v
YYD1005	v	v	v	v	v	v	v	v	v
YYD1006	-	v	v	v	v	-	v	v	v
YYD1007	v	v	v	v	v	v	v	v	v
YYD1008	-	-	v	v	v	-	v	v	v
YYD1010	-	v	v	v	v	-	v	v	v
YYD1011	-	v	v	v	v	-	v	v	v
YYD1012	-	-	v	v	v	-	v	v	v
YYD1013	v	v	v	v	v	v	v	v	v
YYD1014	-	-	-	-	-	-	-	-	-
YYD1015	-	v	v	v	v	-	v	v	v

DNA ID	**MapA Parameter optimisation (50 indi.)**	**MapA *De novo* approach (150 indi.)**	**MapA *De novo* approach (200 indi.)**	**MapA Ref-based with BWA (200 indi.)**	**MapA Ref-based with Stampy (200 indi.)**	**MapB Parameter optimisation (50 indi.)**	**MapB *De novo* approach (200 indi.)**	**MapB Ref-based with BWA (200 indi.)**	**MapB Ref-based with Stampy (200 indi.)**
YYD1016	-	v	v	v	v	-	v	v	v
YYD1017	v	v	v	v	v	v	v	v	v
YYD1018	-	v	v	v	v	-	v	v	v
YYD1019	-	-	v	v	v	-	v	v	v
YYD1020	-	v	v	v	v	-	v	v	v
YYD1021	v	v	v	v	v	v	v	v	v
YYD1022	v	v	v	v	v	v	v	v	v
YYD1023	-	v	v	v	v	-	v	v	v
YYD1024	v	v	v	v	v	v	v	v	v
YYD1025	-	-	v	v	v	-	v	v	v
YYD1026	-	-	v	v	v	-	v	v	v
YYD1027	v	v	v	v	v	v	v	v	v
YYD1028	-	-	v	v	v	-	v	v	v
YYD1029	-	-	v	v	v	-	v	v	v
YYD1030	-	v	v	v	v	-	v	v	v
YYD1031	v	v	v	v	v	v	v	v	v
YYD1032	-	v	v	v	v	-	v	v	v
YYD1033	-	-	-	-	-	-	-	-	-
YYD1034	-	v	v	v	v	-	v	v	v

DNA ID	MapA Parameter optimisation (50 indi.)	MapA *De novo* approach (150 indi.)	MapA *De novo* approach (200 indi.)	MapA Ref-based with BWA (200 indi.)	MapA Ref-based with Stampy (200 indi.)	MapB Parameter optimisation (50 indi.)	MapB *De novo* approach (200 indi.)	MapB Ref-based with BWA (200 indi.)	MapB Ref-based with Stampy (200 indi.)
YYD1035	-	-	v	v	v	-	v	v	v
YYD1036	-	v	v	v	v	-	v	v	v
YYD1037	-	-	v	v	v	-	v	v	v
YYD1038	v	v	v	v	v	v	v	v	v
YYD1040	-	v	v	v	v	-	v	v	v
YYD1041	-	v	v	v	v	-	v	v	v
YYD1042	v	v	v	v	v	v	v	v	v
YYD1043	v	v	v	v	v	v	v	v	v
YYD1044	-	-	v	v	v	-	v	v	v
YYD1045	-	v	v	v	v	-	v	v	v
YYD1046	-	v	v	v	v	-	v	v	v
YYD1047	-	v	v	v	v	-	v	v	v
YYD1048	-	v	v	v	v	-	v	v	v
YYD1049	-	v	v	v	v	-	v	v	v
YYD1050	v	v	v	v	v	v	v	v	v
YYD1051	-	-	-	-	-	-	-	-	-
YYD1052	-	v	v	v	v	-	v	v	v
YYD1053	v	v	v	v	v	v	v	v	v
YYD1056	-	v	v	v	v	-	v	v	v

DNA ID	MapA Parameter optimisation (50 indi.)	MapA *De novo* approach (150 indi.)	MapA *De novo* approach (200 indi.)	MapA Ref-based with BWA (200 indi.)	MapA Ref-based with Stampy (200 indi.)	MapB Parameter optimisation (50 indi.)	MapB *De novo* approach (200 indi.)	MapB Ref-based with BWA (200 indi.)	MapB Ref-based with Stampy (200 indi.)
YYD1058	v	v	v	v	v	v	v	v	v
YYD1059	-	v	v	v	v	-	v	v	v
YYD1060	v	v	v	v	v	v	v	v	v
YYD1061	v	v	v	v	v	v	v	v	v
YYD1062	-	v	v	v	v	-	v	v	v
YYD1063	-	v	v	v	v	-	v	v	v
YYD1064	-	v	v	v	v	-	v	v	v
YYD1065	-	-	v	v	v	-	v	v	v
YYD1066	-	v	v	v	v	-	v	v	v
YYD1067	-	v	v	v	v	-	v	v	v
YYD1069	-	-	v	v	v	-	v	v	v
YYD1070	v	v	v	v	v	v	v	v	v
YYD1071	-	v	v	v	v	-	v	v	v
YYD1072	-	-	v	v	v	-	v	v	v
YYD1073	-	v	v	v	v	-	v	v	v
YYD1074	v	v	v	v	v	v	v	v	v
YYD1075	-	v	v	v	v	-	v	v	v
YYD1076	-	-	v	v	v	-	v	v	v
YYD1077	-	v	v	v	v	-	v	v	v

DNA ID	MapA Parameter optimisation (50 indi.)	MapA *De novo* approach (150 indi.)	MapA *De novo* approach (200 indi.)	MapA Ref-based with BWA (200 indi.)	MapA Ref-based with Stampy (200 indi.)	MapB Parameter optimisation (50 indi.)	MapB *De novo* approach (200 indi.)	MapB Ref-based with BWA (200 indi.)	MapB Ref-based with Stampy (200 indi.)
YYD1078	v	v	v	v	v	v	v	v	v
YYD1079	v	v	v	v	v	v	v	v	v
YYD1080	-	v	v	v	v	-	v	v	v
YYD1081	-	v	v	v	v	-	v	v	v
YYD1082	-	v	v	v	v	-	v	v	v
YYD1083	v	v	v	v	v	v	v	v	v
YYD1084	-	v	v	v	v	-	v	v	v
YYD1085	v	v	v	v	v	v	v	v	v
YYD1086	-	v	v	v	v	-	v	v	v
YYD1087	v	v	v	v	v	v	v	v	v
YYD1088	-	v	v	v	v	-	v	v	v
YYD1089	-	v	v	v	v	-	v	v	v
YYD1090	v	v	v	v	v	v	v	v	v
YYD1091	v	v	v	v	v	v	v	v	v
YYD1092	-	-	v	v	v	-	v	v	v
YYD1093	-	-	-	-	-	-	-	-	-
YYD1094	-	v	v	v	v	-	v	v	v
YYD1095	-	v	v	v	v	-	v	v	v
YYD1096	-	v	v	v	v	-	v	v	v

DNA ID	MapA Parameter optimisation (50 indi.)	MapA *De novo* approach (150 indi.)	MapA *De novo* approach (200 indi.)	MapA Ref-based with BWA (200 indi.)	MapA Ref-based with Stampy (200 indi.)	MapB Parameter optimisation (50 indi.)	MapB *De novo* approach (200 indi.)	MapB Ref-based with BWA (200 indi.)	MapB Ref-based with Stampy (200 indi.)
YYD1098	-	v	v	v	v	-	v	v	v
YYD1099	-	v	v	v	v	-	v	v	v
YYD1100	-	v	v	v	v	-	v	v	v
YYD1101	-	-	v	v	v	-	v	v	v
YYD1102	v	v	v	v	v	v	v	v	v
YYD1103	-	v	v	v	v	-	v	v	v
YYD1104	-	-	-	-	-	-	-	-	-
YYD1105	v	v	v	v	v	v	v	v	v
YYD1106	-	v	v	v	v	-	v	v	v
YYD1107	-	-	v	v	v	-	v	v	v
YYD1108	-	v	v	v	v	-	v	v	v
YYD1109	-	v	v	v	v	-	v	v	v
YYD1110	v	v	v	v	v	v	v	v	v
YYD1111	v	v	v	v	v	v	v	v	v
YYD1112	v	v	v	v	v	v	v	v	v
YYD1113	-	v	v	v	v	-	v	v	v
YYD1114	-	v	v	v	v	-	v	v	v
YYD1115	-	-	-	-	-	-	-	-	-
YYD1116	-	-	-	-	-	-	-	-	-

DNA ID	MapA Parameter optimisation (50 indi.)	MapA *De novo* approach (150 indi.)	MapA *De novo* approach (200 indi.)	MapA Ref-based with BWA (200 indi.)	MapA Ref-based with Stampy (200 indi.)	MapB Parameter optimisation (50 indi.)	MapB *De novo* approach (200 indi.)	MapB Ref-based with BWA (200 indi.)	MapB Ref-based with Stampy (200 indi.)
YYD1117	-	v	v	v	v	-	v	v	v
YYD1118	-	-	-	-	-	-	-	-	-
YYD1119	-	-	-	-	-	-	-	-	-
YYD1120	-	-	-	-	-	-	-	-	-
YYD1121	-	-	-	-	-	-	-	-	-
YYD1122	-	-	-	-	-	-	-	-	-
YYD1123	-	v	v	v	v	-	v	v	v
YYD1124	-	v	v	v	v	-	v	v	v
YYD1125	-	-	v	v	v	-	v	v	v
YYD1127	-	-	-	-	-	-	-	-	-
YYD1128	-	v	v	v	v	-	v	v	v
YYD1129	-	-	-	-	-	-	-	-	-
YYD1130	-	-	v	v	v	-	v	v	v
YYD1131	-	-	-	-	-	-	-	-	-
YYD1132	-	v	v	v	v	-	v	v	v
YYD1133	-	-	v	v	v	-	v	v	v
YYD1134	-	-	v	v	v	-	v	v	v
YYD1135	-	v	v	v	v	-	v	v	v
YYD1136	-	-	v	v	v	-	v	v	v

DNA ID	MapA Parameter optimisation (50 indi.)	MapA *De novo* approach (150 indi.)	MapA *De novo* approach (200 indi.)	MapA Ref-based with BWA (200 indi.)	MapA Ref-based with Stampy (200 indi.)	MapB Parameter optimisation (50 indi.)	MapB *De novo* approach (200 indi.)	MapB Ref-based with BWA (200 indi.)	MapB Ref-based with Stampy (200 indi.)
YYD1137	-	v	v	v	v	-	v	v	v
YYD1138	v	v	v	v	v	v	v	v	v
YYD1139	-	-	-	-	-	-	-	-	-
YYD1140	-	-	-	-	-	-	-	-	-
YYD1141	-	-	v	v	v	-	v	v	v
YYD1142	-	-	-	-	-	-	-	-	-
YYD1143	-	-	-	-	-	-	-	-	-
YYD1144	-	v	v	v	v	-	v	v	v
YYD1145	v	v	v	v	v	v	v	v	v
YYD1146	-	-	-	-	-	-	-	-	-
YYD1147	-	-	v	v	v	-	v	v	v
YYD1148	-	-	-	-	-	-	-	-	-
YYD1149	-	v	v	v	v	-	v	v	v
YYD1150	v	v	v	v	v	v	v	v	v
YYD1151	v	v	v	v	v	v	v	v	v
YYD1152	-	-	v	v	v	-	v	v	v
YYD1153	-	v	v	v	v	-	v	v	v
YYD1154	-	-	v	v	v	-	v	v	v
YYD1155	v	v	v	v	v	v	v	v	v

DNA ID	MapA Parameter optimisation (50 indi.)	MapA *De novo* approach (150 indi.)	MapA *De novo* approach (200 indi.)	MapA Ref-based with BWA (200 indi.)	MapA Ref-based with Stampy (200 indi.)	MapB Parameter optimisation (50 indi.)	MapB *De novo* approach (200 indi.)	MapB Ref-based with BWA (200 indi.)	MapB Ref-based with Stampy (200 indi.)
YYD1156	-	-	v	v	v	-	v	v	v
YYD1157	-	-	v	v	v	-	v	v	v
YYD1158	-	-	-	-	-	-	-	-	-
YYD1159	-	-	-	-	-	-	-	-	-
YYD1160	-	-	v	v	v	-	v	v	v
YYD1161	-	-	v	v	v	-	v	v	v
YYD1162	-	v	v	v	v	-	v	v	v
YYD1163	-	v	v	v	v	-	v	v	v
YYD1164	v	v	v	v	v	v	v	v	v
YYD1165	-	v	v	v	v	-	v	v	v
YYD1166	-	-	-	-	-	-	-	-	-
YYD1167	-	-	v	v	v	-	v	v	v
YYD1168	-	v	v	v	v	-	v	v	v
YYD1169	-	-	-	-	-	-	-	-	-
YYD1170	-	v	v	v	v	-	v	v	v
YYD1171	-	-	v	v	v	-	v	v	v
YYD1172	-	-	-	-	-	-	-	-	-
YYD1173	-	-	-	-	-	-	-	-	-
YYD1174	-	v	v	v	v	-	v	v	v

DNA ID	MapA Parameter optimisation (50 indi.)	MapA *De novo* approach (150 indi.)	MapA *De novo* approach (200 indi.)	MapA Ref-based with BWA (200 indi.)	MapA Ref-based with Stampy (200 indi.)	MapB Parameter optimisation (50 indi.)	MapB *De novo* approach (200 indi.)	MapB Ref-based with BWA (200 indi.)	MapB Ref-based with Stampy (200 indi.)
YYD1175	-	-	v	v	v	-	v	v	v
YYD1176	-	-	-	-	-	-	-	-	-
YYD1177	-	v	v	v	v	-	v	v	v
YYD1178	v	v	v	v	v	v	v	v	v
YYD1180	-	-	v	v	v	-	v	v	v
YYD1181	-	v	v	v	v	-	v	v	v
YYD1183	-	-	v	v	v	-	v	v	v
YYD1184	-	v	v	v	v	-	v	v	v
YYD1185	v	v	v	v	v	v	v	v	v
YYD1188	-	v	v	v	v	-	v	v	v
YYD1189	v	v	v	v	v	v	v	v	v
YYD1190	v	v	v	v	v	v	v	v	v
YYD1191	-	v	v	v	v	-	v	v	v
YYD1192	-	v	v	v	v	-	v	v	v
YYD1193	-	v	v	v	v	-	v	v	v
YYD1194	-	-	-	-	-	-	-	-	-
YYD1195	-	v	v	v	v	-	v	v	v
YYD1196	-	-	v	v	v	-	v	v	v
YYD1197	-	v	v	v	v	-	v	v	v

DNA ID	**MapA Parameter optimisation (50 indi.)**	**MapA *De novo* approach (150 indi.)**	**MapA *De novo* approach (200 indi.)**	**MapA Ref-based with BWA (200 indi.)**	**MapA Ref-based with Stampy (200 indi.)**	**MapB Parameter optimisation (50 indi.)**	**MapB *De novo* approach (200 indi.)**	**MapB Ref-based with BWA (200 indi.)**	**MapB Ref-based with Stampy (200 indi.)**
YYD1198	-	v	v	v	v	-	v	v	v
YYD1199	v	v	v	v	v	v	v	v	v
YYD1200	-	v	v	v	v	-	v	v	v
YYD1201	-	v	v	v	v	-	v	v	v
YYD1203	-	-	v	v	v	-	v	v	v
YYD1204	v	v	v	v	v	v	v	v	v
YYD1205	-	-	v	v	v	-	v	v	v
YYD1206	-	v	v	v	v	-	v	v	v
YYD1207	-	v	v	v	v	-	v	v	v
YYD1208	-	v	v	v	v	-	v	v	v
YYD1209	-	-	v	v	v	-	v	v	v
YYD1210	v	v	v	v	v	v	v	v	v
YYD1211	-	v	v	v	v	-	v	v	v
YYD1212	-	v	v	v	v	-	v	v	v
YYD1213	-	-	v	v	v	-	v	v	v
YYD1214	-	v	v	v	v	-	v	v	v
YYD1215	v	v	v	v	v	v	v	v	v
YYD1216	-	-	v	v	v	-	v	v	v
YYD1217	-	v	v	v	v	-	v	v	v

DNA ID	MapA Parameter optimisation (50 indi.)	MapA *De novo* approach (150 indi.)	MapA *De novo* approach (200 indi.)	MapA Ref-based with BWA (200 indi.)	MapA Ref-based with Stampy (200 indi.)	MapB Parameter optimisation (50 indi.)	MapB *De novo* approach (200 indi.)	MapB Ref-based with BWA (200 indi.)	MapB Ref-based with Stampy (200 indi.)
YYD1218	-	v	v	v	v	-	v	v	v
YYD1219	v	v	v	v	v	v	v	v	v
YYD1220	-	-	-	-	-	-	-	-	-
YYD1221	-	v	v	v	v	-	v	v	v
YYD1224	-	-	v	v	v	-	v	v	v
YYD1226	-	v	v	v	v	-	v	v	v
YYD1227	-	v	v	v	v	-	v	v	v
YYD1228	v	v	v	v	v	v	v	v	v
YYD1229	v	v	v	v	v	v	v	v	v
YYD1230	v	v	v	v	v	v	v	v	v
YYD1231	-	v	v	v	v	-	v	v	v
YYD1232	-	v	v	v	v	-	v	v	v
YYD1233	-	v	v	v	v	-	v	v	v
YYD1234	-	v	v	v	v	-	v	v	v
YYD1235	-	-	v	v	v	-	v	v	v
YYD1236	-	v	v	v	v	-	v	v	v
YYD1237	-	-	v	v	v	-	v	v	v
YYD1238	-	-	v	v	v	-	v	v	v
YYD1239	-	-	v	v	v	-	v	v	v

DNA ID	MapA Parameter optimisation (50 indi.)	MapA *De novo* approach (150 indi.)	MapA *De novo* approach (200 indi.)	MapA Ref-based with BWA (200 indi.)	MapA Ref-based with Stampy (200 indi.)	MapB Parameter optimisation (50 indi.)	MapB *De novo* approach (200 indi.)	MapB Ref-based with BWA (200 indi.)	MapB Ref-based with Stampy (200 indi.)
YYD1240	-	v	v	v	v	-	v	v	v
YYD1241	v	v	v	v	v	v	v	v	v
YYD1242	-	v	v	v	v	-	v	v	v
YYD1244	-	v	v	v	v	-	v	v	v
YYD1245	-	v	v	v	v	-	v	v	v
YYD1246	-	v	v	v	v	-	v	v	v
YYD1250	-	-	v	v	v	-	v	v	v
YYD1251	-	-	v	v	v	-	v	v	v
YYD1252	-	-	v	v	v	-	v	v	v
YYD1253	-	v	v	v	v	-	v	v	v

Appendix 5.2

180 BWA-Stampy marker pairs that show identical segregation patterns in the calculation of combined-approach MapA

BWA Marker	Stampy Marker	Scaffold	Position	Strand
BW3767	ST5796	C14183102	326	-
BW3941	ST6113	C14624983	1879	+
BW4339	ST6747	scaffold10117	17371	-
BW4391	ST6839	scaffold10190	23436	-
BW4668	ST7344	scaffold10702	4089	+
BW4699	ST7401	scaffold10783	93269	-
BW4758	ST7523	scaffold10913	7401	+
BW4836	ST7681	scaffold11081	27767	-
BW5010	ST7986	scaffold11393	43412	+
BW5038	ST8049	scaffold11462	7171	+
BW5107	ST8190	scaffold11598	25969	-
BW5474	ST8892	scaffold12313	13102	+
BW5568	ST9062	scaffold1250	43526	+
BW5804	ST9457	scaffold12952	60093	-
BW5820	ST9477	scaffold12961	18454	+
BW5828	ST9495	scaffold12983	15506	-
BW5841	ST9512	scaffold12996	17314	-
BW6046	ST9891	scaffold1342	5483	+
BW6061	ST9913	scaffold13433	88946	+
BW6195	ST10143	scaffold1365	39627	-
BW6297	ST10330	scaffold13904	32869	-
BW6347	ST10435	scaffold14042	39962	+
BW6370	ST10476	scaffold14085	58879	+
BW6482	ST10663	scaffold14272	8369	+
BW6489	ST10676	scaffold14286	42131	+
BW6641	ST10943	scaffold14541	26750	-
BW6652	ST10961	scaffold14550	78473	-
BW6957	ST11548	scaffold15269	38671	-
BW6971	ST11572	scaffold15277	129542	+
BW6990	ST11595	scaffold15305	25554	-
BW6993	ST11598	scaffold15316	13317	+
BW7131	ST11859	scaffold15612	14728	+
BW7134	ST11862	scaffold15612	29442	+
BW7272	ST12100	scaffold15829	5375	+
BW7384	ST12285	scaffold16071	44967	+
BW7404	ST12323	scaffold16119	1402	+
BW7418	ST12347	scaffold16151	40170	-
BW7420	ST12349	scaffold16151	48404	+
BW7448	ST12406	scaffold16210	55227	-

BWA Marker	Stampy Marker	Scaffold	Position	Strand
BW7488	ST12485	scaffold16309	21425	-
BW7502	ST12512	scaffold16314	4805	-
BW7676	ST12868	scaffold16777	98893	+
BW7679	ST12872	scaffold16787	21837	+
BW7842	ST13201	scaffold17298	50104	+
BW7893	ST13297	scaffold17411	6876	-
BW7925	ST13356	scaffold17492	10316	-
BW8013	ST13516	scaffold17701	20205	-
BW8058	ST13606	scaffold17843	32328	-
BW8286	ST14001	scaffold18287	8783	-
BW8394	ST14215	scaffold18611	19423	+
BW8477	ST14371	scaffold18808	27271	-
BW8517	ST14440	scaffold18918	56713	+
BW8632	ST14655	scaffold1922	10652	-
BW8681	ST14758	scaffold19363	15553	+
BW8864	ST15105	scaffold19804	4115	-
BW9061	ST15462	scaffold20280	33531	+
BW9148	ST15605	scaffold20494	70756	+
BW9232	ST15747	scaffold20707	9923	+
BW9295	ST15854	scaffold20823	40154	-
BW9383	ST15993	scaffold2098	33140	-
BW9419	ST16057	scaffold21114	12335	-
BW9428	ST16071	scaffold21133	22160	-
BW9449	ST16105	scaffold21183	41365	+
BW9535	ST16245	scaffold21365	1117	+
BW9717	ST16590	scaffold21885	28980	+
BW9870	ST16858	scaffold22293	49537	+
BW9893	ST16901	scaffold22380	30140	-
BW9901	ST16915	scaffold22405	6224	-
BW10092	ST17255	scaffold22922	14237	-
BW10250	ST17535	scaffold2331	53274	+
BW10351	ST17711	scaffold23630	45200	-
BW10352	ST17712	scaffold23630	70792	+
BW10524	ST18034	scaffold24222	9012	-
BW10601	ST18183	scaffold24488	9298	+
BW10623	ST18217	scaffold24581	8741	+
BW10630	ST18225	scaffold24601	31900	-
BW10749	ST18423	scaffold24914	30692	-
BW10764	ST18445	scaffold24965	11829	-
BW11020	ST18891	scaffold25607	44619	+
BW11070	ST18980	scaffold25715	52298	-

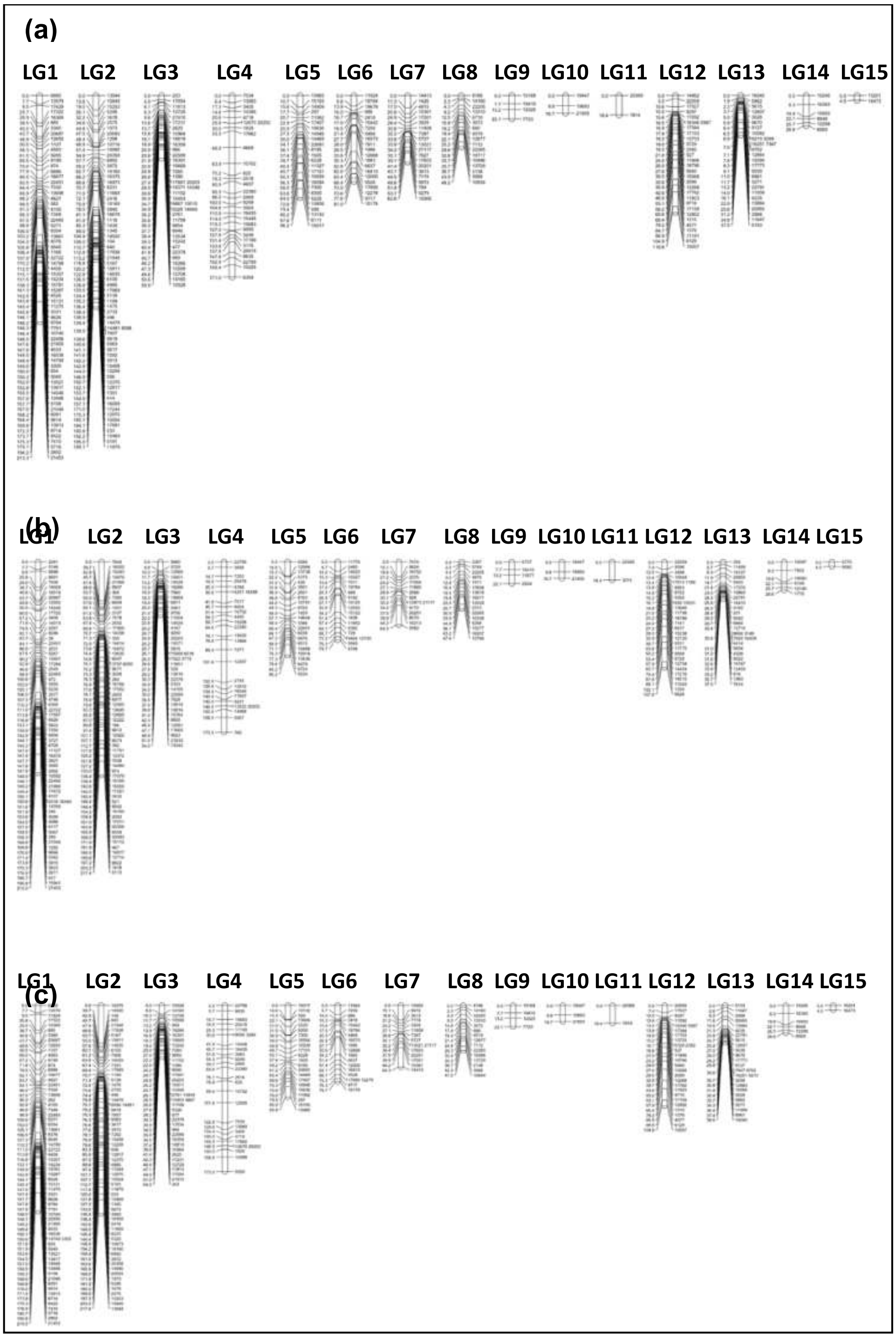
BWA Marker	Stampy Marker	Scaffold	Position	Strand
BW11088	ST19006	scaffold25779	12291	-
BW11130	ST19071	scaffold25848	49278	-
BW11216	ST19224	scaffold26126	4859	+
BW11420	ST19598	scaffold26726	68350	-
BW11582	ST19877	scaffold27202	30650	+
BW11591	ST19895	scaffold27259	19178	-
BW11594	ST19900	scaffold2726	12602	+
BW11595	ST19901	scaffold2726	12883	-
BW11684	ST20056	scaffold2751	58773	-
BW11697	ST20079	scaffold27596	48497	-
BW11717	ST20111	scaffold27621	53347	+
BW11826	ST20332	scaffold27999	4796	+
BW11831	ST20346	scaffold28029	7158	+
BW11895	ST20452	scaffold28187	28994	+
BW12036	ST20709	scaffold28705	119999	-
BW12159	ST20958	scaffold29082	105292	+
BW12176	ST20984	scaffold29097	64019	-
BW12214	ST21048	scaffold2920	113681	+
BW12266	ST21157	scaffold29374	35664	-
BW12283	ST21202	scaffold2950	26710	+
BW12302	ST21229	scaffold2959	39139	+
BW12303	ST21230	scaffold2959	45440	+
BW12361	ST21349	scaffold29891	16647	+
BW12362	ST21350	scaffold29891	30118	+
BW12399	ST21435	scaffold30120	12104	+
BW12584	ST21790	scaffold30851	29472	-
BW12648	ST21925	scaffold31095	86337	+
BW12663	ST21952	scaffold31158	58652	+
BW12696	ST22023	scaffold31294	5516	+
BW12699	ST22027	scaffold31295	54800	-
BW12983	ST22545	scaffold32543	15254	-
BW13099	ST22763	scaffold33166	14592	-
BW13130	ST22814	scaffold33385	9709	+
BW13134	ST22818	scaffold33395	21050	+
BW13163	ST22871	scaffold33618	12934	-
BW13230	ST23010	scaffold3404	25166	+
BW13264	ST23085	scaffold34277	16067	-
BW13313	ST23160	scaffold34477	35533	-
BW13346	ST23220	scaffold34647	4634	-
BW13402	ST23328	scaffold35016	29130	+
BW13574	ST23658	scaffold3606	77508	-

BWA Marker	Stampy Marker	Scaffold	Position	Strand
BW13599	ST23696	scaffold36183	25933	+
BW13693	ST23851	scaffold36602	10449	-
BW13749	ST23960	scaffold36995	57	-
BW13829	ST24107	scaffold3742	41855	+
BW13904	ST24264	scaffold3798	15826	+
BW13908	ST24268	scaffold3798	57252	-
BW14273	ST24933	scaffold4036	94508	-
BW14350	ST25076	scaffold40816	36054	+
BW14432	ST25200	scaffold4113	83743	+
BW14521	ST25363	scaffold41884	13874	+
BW14696	ST25695	scaffold43209	6656	+
BW14702	ST25706	scaffold4321	52636	+
BW14712	ST25723	scaffold43263	7987	+
BW14901	ST26079	scaffold44762	18009	+
BW14910	ST26095	scaffold44835	43066	+
BW14913	ST26099	scaffold44839	23353	+
BW15066	ST26391	scaffold46013	420	+
BW15107	ST26456	scaffold46316	3995	+
BW15286	ST26805	scaffold48317	8103	-
BW15354	ST26936	scaffold4898	28693	-
BW15427	ST27075	scaffold49619	7068	+
BW15444	ST27112	scaffold4980	33434	+
BW15453	ST27123	scaffold49854	27112	-
BW15709	ST27585	scaffold5245	129912	-
BW15710	ST27588	scaffold5245	175677	+
BW16056	ST28142	scaffold5580	30104	-
BW16213	ST28437	scaffold5810	47216	-
BW16403	ST28818	scaffold6153	3814	-
BW16404	ST28820	scaffold6153	4480	-
BW16422	ST28869	scaffold6220	21373	-
BW16603	ST29187	scaffold6421	14104	+
BW16718	ST29401	scaffold66222	6091	-
BW16934	ST29804	scaffold6932	34971	+
BW16944	ST29825	scaffold69534	39521	-
BW16956	ST29850	scaffold69601	2675	-
BW17071	ST30055	scaffold714	8615	+
BW17076	ST30065	scaffold7167	117507	-
BW17203	ST30322	scaffold7386	13802	+
BW17205	ST30326	scaffold7386	48478	-
BW17206	ST30328	scaffold7389	156178	+
BW17232	ST30368	scaffold741	32123	-

BWA Marker	Stampy Marker	Scaffold	Position	Strand
BW17250	ST30404	scaffold7435	75673	+
BW17355	ST30593	scaffold7589	166494	-
BW17458	ST30791	scaffold7734	40721	+
BW17572	ST31003	scaffold7904	9448	-
BW17642	ST31137	scaffold802	36538	-
BW17718	ST31262	scaffold811	15736	-
BW17964	ST31676	scaffold8486	60014	+
BW18217	ST32118	scaffold8764	39770	-
BW18246	ST32184	scaffold8797	119176	+
BW18295	ST32273	scaffold8874	51834	+
BW18357	ST32380	scaffold8944	21134	-
BW18551	ST32763	scaffold92180	6098	-
BW18666	ST32971	scaffold9373	24348	-
BW18783	ST33161	scaffold9482	34049	+
BW19000	ST33527	scaffold97118	4255	+
BW19004	ST33533	scaffold97142	8683	+
BW19008	ST33539	scaffold97180	2036	-
BW19041	ST33604	scaffold97354	9107	-

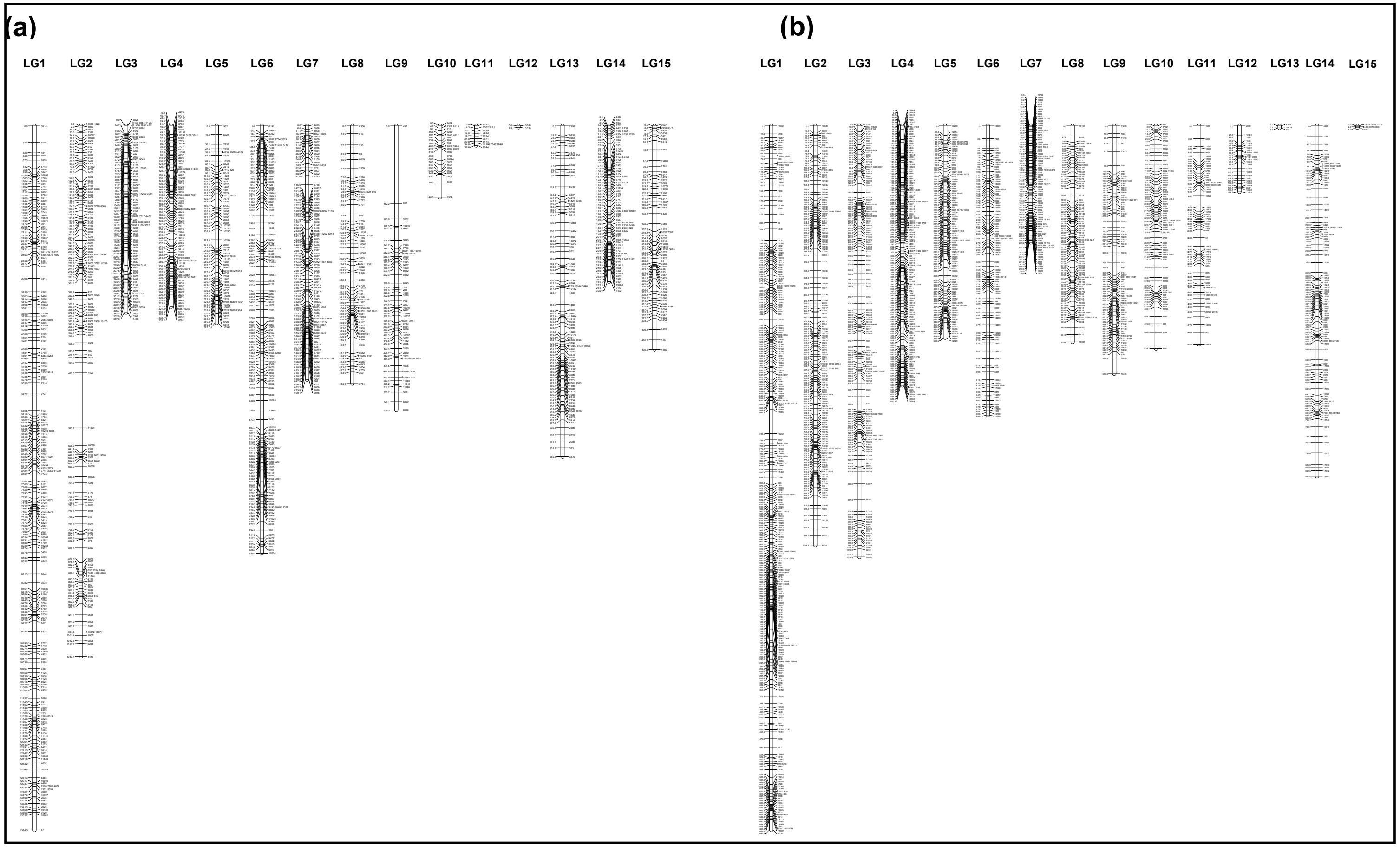
Appendix 5.3

De novo approach MapB calculated based on 50 individuals with three different parameter-settings. **(a)** Parameter: m=3 M=1 N=1 n=1. **(b)** Parameter: m=3 M=1 N=1 n=8. **(c)** Parameter: m=3 M=1 N=1 n=16. The marker positions are shown on the left (cM), and the marker names are shown on the right of each linkage.


(a)
LG1 LG2 LG3 LG4 LG5 LG6 LG7 LG8 LG9 LG10 LG11 LG12 LG13 LG14 LG15
(b)
LG1 LG2 LG3 LG4 LG5 LG6 LG7 LG8 LG9 LG10 LG11 LG12 LG13 LG14 LG15
(c)
LG1 LG2 LG3 LG4 LG5 LG6 LG7 LG8 LG9 LG10 LG11 LG12 LG13 LG14 LG15

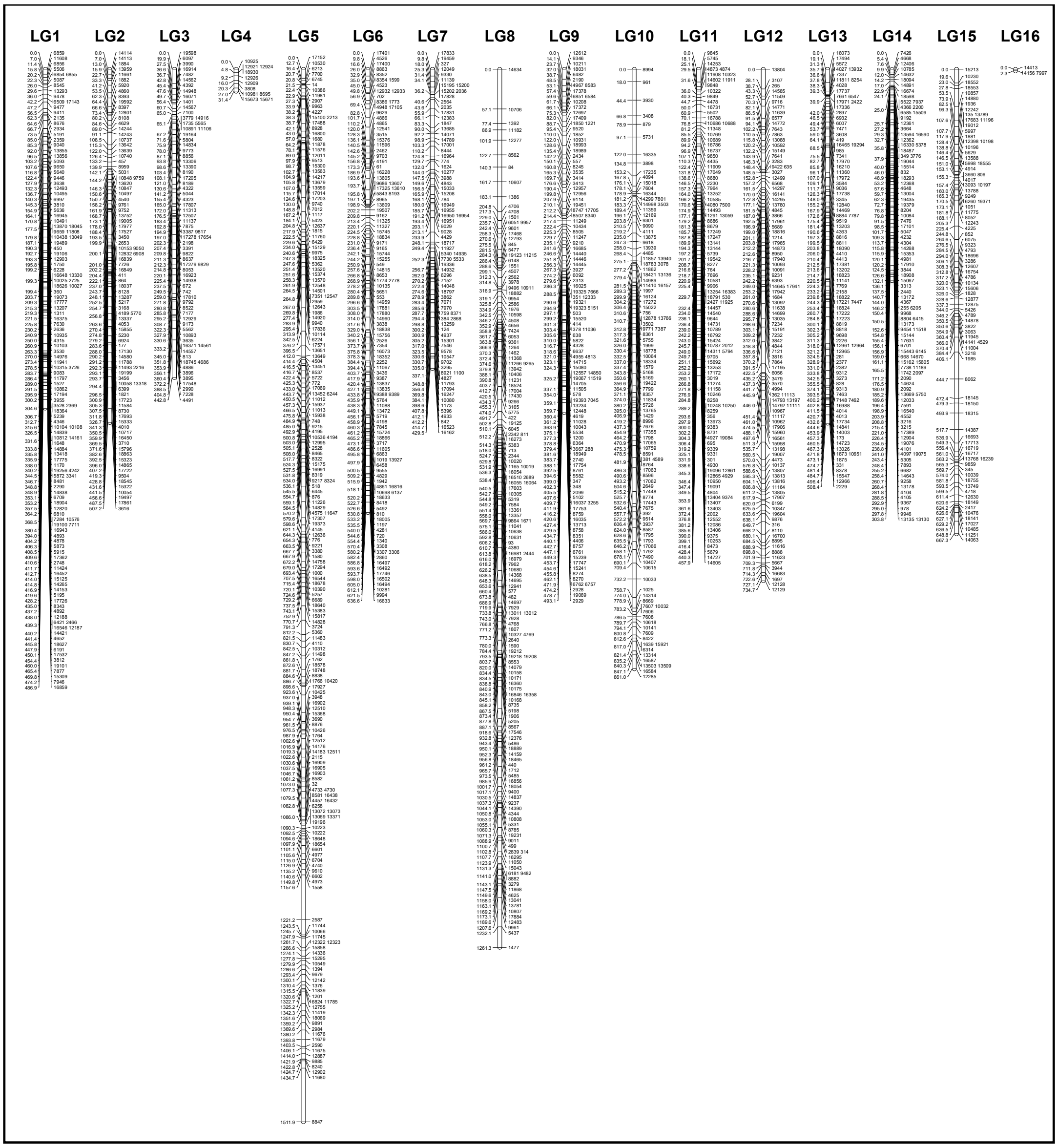
Appendix 5.4

Optimisation of the reference-based approach using the BWA aligner for MapB calculation based on 50 BC individuals. **(a)** Default BWA parameters. **(b)** Parameter: n=12 k=3. The marker positions are shown on the left (cM), and the marker names are shown on the right of each linkage.



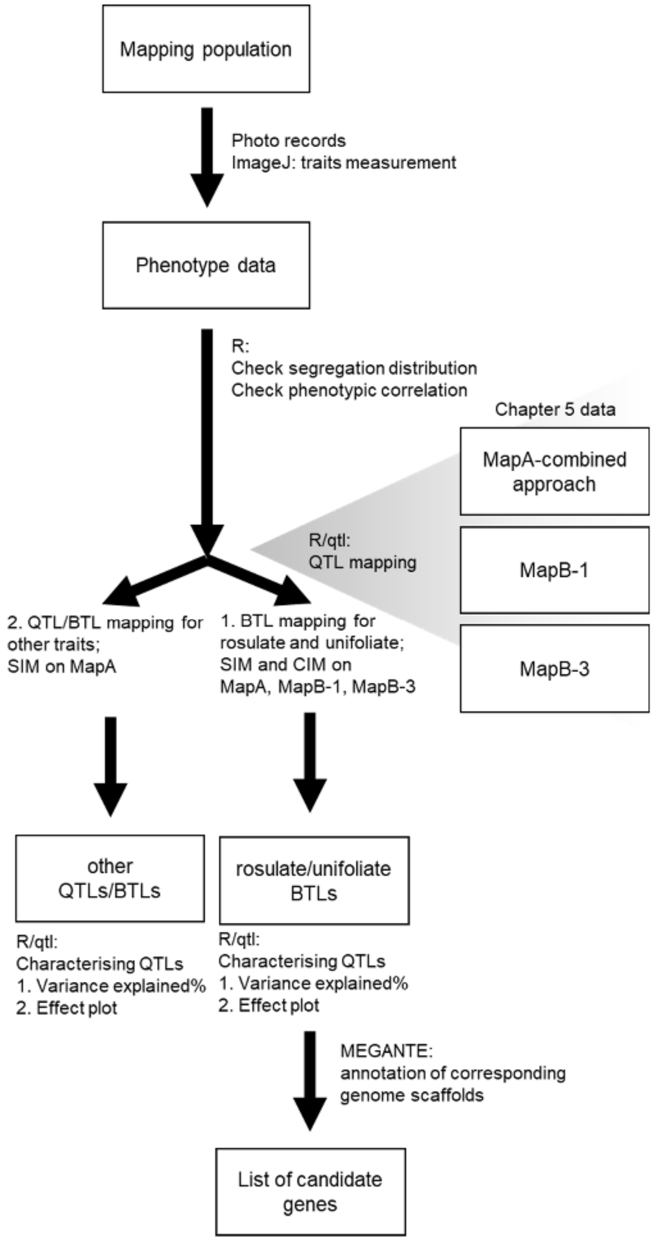
Appendix 5.5

Reference-based approach using the Stampy aligner for MapB calculation based on 50 BC individuals. The marker positions are shown on the left (cM), and the marker names are shown on the right of each linkage.



Appendix 6.1

Flowchart of QTL mapping analysis. QTL – quantitative trait loci. BTL – binary trait loci. SIM – standard interval mapping. CIM – composite interval mapping.



Appendix 6.2

Table. Statistical comparisons of the floral traits between *S. grandis*^{BC} (*N* = 12) and *S. grandis*^{F1} (*N* = 10) lineages. Showing the average values of measurements ± standard deviation.

Trait	***S. grandis***$^{\mathbf{BC}}$	***S. grandis***$^{\mathbf{F1}}$	***P*-value**
Corolla length (cm)	3.81 ± 0.31	4.49 ± 0.28	< 0.01
Undilated tube length (cm)	1.59 ± 0.11	1.53 ± 0.07	0.82
Dilated tube length (cm)	1.58 ± 0.16	2.22 ± 0.10	< 0.01
Undilated tube height (cm)	0.53 ± 0.09	0.50 ± 0.19	0.32
Dilated tube height (cm)	0.69 ± 0.06	0.72 ± 0.43	0.35
Undilated tube width (cm)	0.55 ± 0.10	0.54 ± 0.09	0.79
Dilated tube width (cm)	0.71 ± 0.15	0.76 ± 0.07	0.31
Corolla face height (cm)	1.88 ± 0.45	2.07 ± 0.20	0.42
Tube opening height (Outer) (cm)	1.03 ± 0.25	1.13 ± 0.21	0.42
Tube opening height (Inner) (cm)	0.86 ± 0.19	1.00 ± 0.00	0.18
Corolla face width (cm)	2.06 ± 0.47	2.56 ± 0.39	< 0.01
Tube opening width (Outer) (cm)	1.25 ± 0.24	1.31 ± 0.17	0.72
Tube opening width (Inner) (cm)	0.84 ± 0.15	0.92 ± 0.23	0.54
Pistil length (cm)	2.36 ± 0.16	2.65 ± 0.21	< 0.01
Ovary length (cm)	1.52 ± 0.24	1.92 ± 0.09	< 0.01
Style length (cm)	0.84 ± 0.17	0.73 ± 0.10	0.25
Calyx length (cm)	0.43 ± 0.12	0.45 ± 0.11	0.38
Stamen length (cm)	2.13 ± 0.16	2.10 ± 0.12	0.69
Filament length (attached) (cm)	1.62 ± 0.13	1.53 ± 0.08	0.14
Filament length (detached) (cm)	0.51 ± 0.06	0.56 ± 0.14	0.11
Ventral tube length (cm)	3.16 ± 0.24	3.75 ± 0.18	< 0.01
Ventral lobe length (cm)	0.66 ± 0.12	0.73 ± 0.21	0.16
Dorsal tube length (cm)	2.60 ± 0.19	2.88 ± 0.07	<0.01
Dorsal lobe length (cm)	0.61 ± 0.08	0.64 ± 0.29	0.43
Time to flowering (DAS)	377.5 ± 69.90	265 ± 0.32	< 0.01

Note. Wilcoxon rank sum test was used to compare the difference. DAS: days after sowing.

Appendix 6.3

Table. Statistical comparisons of the floral traits between *S. rexii* (*N* = 17) and *S. grandis* (data for *S. grandis* are those combined from the *S. grandis*^{F1} and *S. grandis*^{BC} lineages; *N* = 22). Showing the average values of measurements ± standard deviation.

Trait	*S. rexii*	*S. grandis*	*P*-value
Corolla length (cm)	6.79 ± 0.60	4.13 ± 0.46	< 0.01
Undilated tube length (cm)	2.65 ± 0.42	1.58 ± 0.10	< 0.01
Dilated tube length (cm)	2.74 ± 0.46	1.86 ± 0.38	< 0.01
Undilated tube height (cm)	0.49 ± 0.08	0.52 ± 0.09	0.18
Dilated tube height (cm)	1.01 ± 0.11	0.71 ± 0.06	< 0.01
Undilated tube width (cm)	0.44 ± 0.05	0.55 ± 0.09	< 0.01
Dilated tube width (cm)	1.03 ± 0.10	0.74 ± 0.13	< 0.01
Corolla face height (cm)	4.29 ± 0.86	1.97 ± 0.44	< 0.01
Tube opening height (Outer) (cm)	1.94 ± 0.49	1.08 ± 0.24	< 0.01
Tube opening height (Inner) (cm)	1.51 ± 0.34	0.93 ± 0.21	< 0.01
Corolla face width (cm)	5.06 ± 0.91	2.29 ± 0.52	< 0.01
Tube opening width (Outer) (cm)	2.73 ± 0.67	1.29 ± 0.22	< 0.01
Tube opening width (Inner) (cm)	1.88 ± 0.41	0.88 ± 0.16	< 0.01
Pistil length (cm)	3.91 ± 0.10	2.49 ± 0.23	< 0.01
Ovary length (cm)	2.60 ± 0.12	1.71 ± 0.31	< 0.01
Style length (cm)	1.31 ± 0.14	0.79 ± 0.15	< 0.01
Calyx length (cm)	0.56 ± 0.07	0.45 ± 0.11	< 0.01
Stamen length (cm)	3.44 ± 0.10	2.13 ± 0.12	< 0.01
Filament length (attached) (cm)	2.55 ± 0.08	1.59 ± 0.11	< 0.01
Filament length (detached) (cm)	0.88 ± 0.07	0.53 ± 0.08	< 0.01
Ventral tube length (cm)	5.39 ± 0.51	3.44 ± 0.40	< 0.01
Ventral lobe length (cm)	1.39 ± 0.17	0.69 ± 0.10	< 0.01
Dorsal tube length (cm)	4.62 ± 0.33	2.74 ± 0.24	< 0.01
Dorsal lobe length (cm)	1.13 ± 0.22	0.62 ± 0.08	< 0.01
Time to flowering (DAS)	237 ± 0.00	329 ± 77.08	< 0.01

Note. Wilcoxon rank sum test was used to compare the difference. DAS: days after sown.

Appendix 6.4

Table. Three way comparisons between *S. rexii* (*N* = 17), *S. grandis*^{F1} (*N* = 10) and the F1 (*N* = 17) using Dunn's post hoc test. *P*-values > 0.05 (indicate no statistical significant differences) are highlighted in grey

Trait 1. Corolla length

	(*S. grandis* × *S. rexii*) F1	*S. grandis*
S. grandis	0.0003	-
S. rexii	0.0281	< 0.0001

Trait 2. Undilated tube length

	(*S. grandis* × *S. rexii*) F1	*S. grandis*
S. grandis	0.0006	-
S. rexii	0.0073	< 0.0001

Trait 3. Dilated tube length

	(*S. grandis* × *S. rexii*) F1	*S. grandis*
S. grandis	0.0012	-
S. rexii	0.2293	0.0082

Trait 4. Undilated tube height

	(*S. grandis* × *S. rexii*) F1	*S. grandis*
S. grandis	0.2366	-
S. rexii	0.1047	0.3582

Trait 5. Dilated tube height

	(*S. grandis* × *S. rexii*) F1	*S. grandis*
S. grandis	0.0003	-
S. rexii	0.0462	< 0.01

Trait 6. Undilted tube width

	(*S. grandis* × *S. rexii*) F1	*S. grandis*
S. grandis	0.0847	-
S. rexii	0.0145	0.0006

Trait 7. Dilated tube width

	(*S. grandis* × *S. rexii*) F1	*S. grandis*
S. grandis	0.0121	-
S. rexii	0.0020	< 0.0001

Trait 8. Corolla face height

	(*S. grandis* × *S. rexii*) F1	*S. grandis*
S. grandis	0.0001	-
S. rexii	0.0909	< 0.0001

Trait 9. Tube opening height (outer)

	(*S. grandis* × *S. rexii*) F1	*S. grandis*
S. grandis	0.0003	-
S. rexii	0.1310	< 0.0001

Trait 10. Tube opening height (inner)

	(*S. grandis* × *S. rexii*) F1	*S. grandis*
S. grandis	0.0005	-
S. rexii	0.2673	0.0001

Trait 11. Corolla face width

	(*S. grandis* × *S. rexii*) F1	*S. grandis*
S. grandis	0.0003	-
S. rexii	0.0264	< 0.0001

Trait 12. Tube opening width (outer)

	(*S. grandis* × *S. rexii*) F1	*S. grandis*
S. grandis	0.0019	-
S. rexii	0.0015	< 0.0001

Trait 13. Tube opening width (inner)

	(*S. grandis* × *S. rexii*) F1	*S. grandis*
S. grandis	0.0029	-
S. rexii	0.0015	< 0.0001

Trait 14. Pistil length

	(*S. grandis* × *S. rexii*) F1	*S. grandis*
S. grandis	0.0007	-
S. rexii	0.0054	< 0.0001

Trait 15. Ovary length

	(*S. grandis* × *S. rexii*) F1	*S. grandis*
S. grandis	0.0005	-
S. rexii	0.0094	< 0.0001

Trait 16. Style length

	(*S. grandis* × *S. rexii*) F1	*S. grandis*
S. grandis	< 0.0001	-
S. rexii	0.3346	< 0.0001

Trait 17. Calyx length

	(*S. grandis* × *S. rexii*) F1	*S. grandis*
S. grandis	0.0894	-
S. rexii	0.0153	0.0007

Trait 18. Stamen length

	(*S. grandis* × *S. rexii*) F1	*S. grandis*
S. grandis	0.0042	-
S. rexii	0.0001	< 0.0001

Trait 19. Filament length (attached)

	(*S. grandis* × *S. rexii*) F1	*S. grandis*
S. grandis	0.0030	-
S. rexii	0.0002	< 0.0001

Trait 20. Filament length (free)

	(*S. grandis* × *S. rexii*) F1	*S. grandis*
S. grandis	0.0020	-
S. rexii	0.0038	< 0.0001

Trait 21. Ventral tube length

	(*S. grandis* × *S. rexii*) F1	*S. grandis*
S. grandis	0.0001	-
S. rexii	0.0766	< 0.0001

Trait 22. Ventral lobe length

	(*S. grandis* × *S. rexii*) F1	*S. grandis*
S. grandis	0.0002	-
S. rexii	0.0560	< 0.0001

Trait 23. Dorsal tube length

	(*S. grandis* × *S. rexii*) F1	*S. grandis*
S. grandis	0.0021	-
S. rexii	0.0004	< 0.0001

Trait 24. Dorsal lobe length

	(*S. grandis* × *S. rexii*) F1	*S. grandis*
S. grandis	0.0001	-
S. rexii	0.0766	< 0.0001

Appendix 6.5

Table. Traits scoring results of the 200 *Streptocarpus* backcross individuals. The listed trait numbers corresponds to that in the Table 6.4. Qual.: qualifiers of the BC plants. Unit for trait 1 to 24: cm. Unit for trait 25: days after cotyledons unfold. Trait 29.1: rosulate / unifoliate scoring Method 1. Trait 29.2: rosulate / unifoliate scoring Method 2. Trait 29.3: rosulate / unifoliate scoring Method 3. Trait 29.4: rosulate / unifoliate scoring Method 4. For binary traits 26 to 28, 30 and 31, '1' represent present and '0' represent absent. For traits 29.1 to 29.4, '1' represents rosulate and '0' represents unifoliate. -: unknown. AVG: average value (cm). STD: Standard deviation (cm).

	Trait numbers																																	
Qual.	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**	**25**	**26**	**27**	**28**	**29.1**	**29.2**	**29.3**	**29.4**	**30**	**31**
D	5.691	1.622	4.069	0.603	1.030	0.560	0.972	3.375	1.279	1.051	3.462	1.676	1.102	3.087	1.992	1.095	0.704	2.828	2.250	0.578	4.724	0.967	3.618	0.881	280	1	1	1	1	1	1	1	0	0
E	5.446	1.742	3.704	0.632	0.932	0.606	0.875	3.300	1.563	1.226	3.611	1.703	1.106	3.170	2.110	1.061	0.609	3.251	2.500	0.751	4.321	1.126	3.869	0.847	231	1	1	1	0	0	0	0	1	0
F	4.809	1.492	3.317	0.578	0.926	0.704	1.038	2.769	1.358	1.080	3.154	1.637	1.248	2.837	1.817	1.020	0.643	2.627	2.185	0.442	3.798	1.011	3.288	0.912	196	1	1	1	0	1	-	1	1	0
G	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1	1	1	1	-	0
H	4.966	1.304	3.662	0.596	0.901	0.616	0.930	2.730	1.294	1.040	3.100	1.574	0.997	2.940	1.872	1.069	0.544	2.480	1.872	0.608	3.947	1.019	3.100	0.773	259	1	1	0	0	0	-	0	1	0
J	4.483	1.509	2.974	0.525	0.880	0.543	0.992	2.599	1.099	0.810	2.807	1.445	0.997	3.221	2.088	1.133	0.544	2.926	2.262	0.664	3.508	0.975	3.216	0.712	238	1	0	1	0	1	1	1	1	0
K	4.556	1.546	3.010	0.581	0.867	0.551	0.952	1.991	1.032	0.827	2.441	1.255	0.884	2.923	1.823	1.100	0.712	2.463	1.889	0.574	3.662	0.894	2.992	0.664	455	1	1	1	0	0	0	1	0	0
L	4.278	1.478	2.801	0.427	0.681	0.421	0.670	2.187	1.089	0.914	2.226	1.202	0.899	2.936	1.936	1.000	0.508	2.780	2.069	0.711	3.501	0.778	2.878	0.697	210	1	1	0	0	0	0	0	0	0
M	5.011	1.557	3.454	0.520	0.845	0.553	0.807	2.362	1.088	0.853	2.435	1.355	0.843	3.070	1.960	1.110	0.621	2.820	2.182	0.638	4.074	0.937	3.235	0.935	210	0	0	1	0	1	1	1	1	0
O	5.048	1.456	3.592	0.653	0.890	0.581	1.071	2.695	1.465	1.244	3.279	1.808	1.168	2.867	1.830	1.037	0.626	2.412	1.752	0.660	4.071	0.977	3.023	0.682	287	1	1	0	0	0	-	0	1	0
P	3.763	1.110	2.653	0.444	0.787	0.448	0.744	2.038	0.993	0.828	2.477	1.282	0.883	2.993	1.744	1.249	0.405	2.169	1.598	0.571	2.956	0.807	2.405	0.495	245	0	0	1	0	0	0	0	0	0
Q	5.268	1.485	3.783	0.557	0.864	0.521	1.081	2.614	1.433	1.086	3.387	1.669	1.182	3.250	2.129	1.121	0.668	2.900	2.243	0.657	4.248	1.020	3.265	0.869	210	1	0	1	0	1	1	1	1	0
R	4.763	1.420	3.343	0.479	0.850	0.556	0.986	2.417	1.275	1.055	2.867	1.621	1.180	3.103	2.056	1.047	0.499	2.859	2.101	0.758	3.757	1.006	3.128	0.822	196	1	1	1	0	0	0	0	0	0
T	5.102	1.956	3.146	0.670	0.931	0.598	0.808	2.791	1.396	1.098	3.284	1.643	1.032	3.065	1.914	1.151	0.676	2.680	2.014	0.666	3.980	1.122	3.032	0.898	224	1	1	1	0	1	-	1	1	0
U	4.348	1.329	3.020	0.472	0.663	0.634	1.017	2.633	1.410	1.468	2.411	1.435	1.126	2.955	1.869	1.086	0.708	2.492	1.766	0.726	3.446	0.902	2.805	0.886	217	1	1	1	0	0	0	0	0	0
V	5.042	1.622	3.420	0.477	0.837	0.603	0.490	2.319	1.255	0.965	3.126	1.396	0.930	3.275	2.203	1.072	0.735	2.884	2.287	0.597	4.066	0.976	3.207	0.794	238	1	1	0	0	0	0	0	0	0
W	4.614	1.433	3.181	0.446	0.718	0.509	0.776	2.242	1.057	0.781	2.525	1.347	0.989	3.027	1.898	1.129	0.530	2.575	2.102	0.473	3.763	0.850	2.816	0.931	455	0	0	1	1	1	1	1	1	0
X	4.981	1.236	3.745	0.443	0.849	0.685	0.863	2.714	1.372	1.190	3.095	1.572	1.083	2.878	1.864	1.014	0.730	2.842	2.248	0.594	3.995	0.986	3.192	0.803	210	1	0	0	0	0	-	0	1	0
Y	3.565	1.255	2.310	0.488	0.776	0.577	0.780	1.606	-	-	2.750	-	-	2.711	1.869	0.842	0.490	-	-	-	2.946	0.619	2.391	0.703	371	-	0	-	0	0	1	1	1	0
Z	4.641	1.329	3.311	0.467	0.815	0.501	0.880	2.705	1.326	1.108	3.243	1.636	1.179	3.213	2.007	1.206	0.465	2.752	1.961	0.791	3.622	1.018	3.013	1.028	364	1	1	0	1	1	1	1	1	0
AA	5.785	1.949	3.837	0.662	1.170	0.886	1.319	4.660	2.320	1.979	5.205	2.968	2.104	3.868	2.781	1.087	0.867	3.294	2.600	0.694	3.969	1.817	4.069	1.143	210	1	1	0	1	1	1	1	0	0
AB	3.859	1.570	2.289	0.448	0.713	0.493	0.789	2.532	1.378	1.220	2.999	1.642	1.240	3.121	1.994	1.127	0.379	2.546	1.949	0.597	2.811	1.048	2.535	0.718	252	1	1	0	0	0	0	0	0	0
AC	5.306	1.768	3.538	0.697	0.964	0.641	0.975	3.842	1.682	1.311	3.833	1.778	1.155	3.318	2.131	1.187	0.523	2.967	2.298	0.669	3.947	1.359	3.398	0.803	252	1	0	0	1	1	1	1	1	0
AD	5.633	1.717	3.916	0.590	0.912	0.615	0.960	2.858	1.328	0.893	3.305	1.458	0.939	3.173	1.920	1.253	0.604	3.085	2.168	0.917	4.480	1.153	3.396	1.019	259	1	1	0	1	1	1	1	0	0
AE	4.631	1.371	3.260	0.471	0.851	0.510	0.894	2.720	1.216	0.955	2.951	1.538	1.148	3.499	2.318	1.181	0.536	2.802	2.146	0.656	3.693	0.938	2.971	0.808	224	1	0	0	0	0	0	0	0	0
AF	5.052	1.573	3.479	0.599	0.929	0.682	1.098	3.316	1.455	1.112	3.728	1.830	1.205	3.438	2.111	1.327	0.498	3.043	2.141	0.902	3.842	1.210	3.226	1.157	210	1	0	1	1	1	1	1	0	0
AG	5.119	1.482	3.638	0.591	0.870	0.616	0.933	2.522	1.097	0.882	3.017	1.393	1.003	2.979	1.960	1.019	0.722	2.757	2.010	0.747	4.105	1.014	3.111	0.966	364	1	0	1	0	1	1	1	1	0
AH	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1	1	1	1	0	0
AI	4.094	1.326	2.768	0.378	0.653	0.371	0.705	2.093	1.063	0.932	2.359	1.224	0.957	2.941	1.686	1.255	0.489	2.436	1.946	0.490	3.438	0.656	2.905	0.526	455	1	0	0	1	1	1	1	0	0
AJ	5.832	1.685	4.147	0.649	0.965	0.585	1.018	2.686	1.466	1.270	3.354	1.820	1.261	3.457	2.146	1.311	0.693	2.990	2.064	0.926	4.589	1.243	3.358	1.086	196	1	1	0	1	1	1	1	1	0
AK	5.159	1.882	3.277	0.607	0.898	0.624	1.004	2.594	1.138	0.892	3.189	1.538	1.056	3.307	2.200	1.107	0.760	2.652	1.969	0.683	4.109	1.050	3.083	0.846	238	0	0	1	1	1	1	1	0	0
AL	4.818	1.434	3.384	0.558	0.913	0.471	0.794	1.532	0.828	0.720	1.949	1.169	0.794	2.931	1.816	1.115	0.469	2.326	1.686	0.639	4.174	0.644	3.213	0.632	399	1	0	1	1	1	1	1	1	0
AN	4.276	1.359	2.917	0.445	0.739	0.419	0.780	2.328	1.119	0.915	2.639	1.341	0.910	2.975	1.871	1.104	0.376	2.558	1.776	0.782	3.443	0.833	2.868	0.738	364	1	1	1	1	1	1	1	0	0
AO	3.977	1.308	2.669	0.402	0.612	0.474	0.788	2.116	1.060	0.893	2.545	1.354	0.925	3.315	2.123	1.192	0.516	2.873	2.143	0.730	3.036	0.942	2.761	0.792	238	1	1	0	1	1	1	1	0	0
AP	4.303	1.178	3.125	0.525	0.794	0.654	1.006	2.470	1.356	1.096	2.961	1.533	1.095	2.815	1.758	1.057	0.542	2.305	1.776	0.529	3.410	0.893	2.601	0.818	224	1	0	1	0	0	-	1	-	-
AQ	4.434	1.273	3.161	0.555	0.842	0.639	1.015	2.447	1.116	0.977	2.947	1.609	1.609	2.799	1.854	0.945	0.527	2.485	1.817	0.668	3.532	0.902	2.930	0.804	252	1	0	1	1	1	1	1	0	0
AS	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1	1	1	1	0	0
AT	4.151	1.198	2.953	0.377	0.672	0.455	0.807	2.358	1.117	0.933	2.771	1.379	0.988	2.848	1.854	0.994	0.502	2.509	1.791	0.718	3.213	0.938	2.644	0.676	455	1	1	0	1	1	1	1	1	0
AU	4.458	1.375	3.083	0.404	0.847	0.508	0.989	2.247	1.125	0.912	2.710	1.506	1.030	3.178	1.853	1.325	0.488	2.528	2.181	0.347	3.510	0.948	2.949	0.866	238	1	1	1	1	1	1	1	0	1

	Trait numbers																																	
Qual.	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**	**25**	**26**	**27**	**28**	**29.1**	**29.2**	**29.3**	**29.4**	**30**	**31**
AV	3.976	1.432	2.544	0.506	0.669	0.558	0.910	2.458	1.053	0.835	2.649	1.336	0.865	2.981	2.123	0.858	0.375	2.813	2.075	0.738	3.158	0.818	2.513	0.872	315	0	0	1	1	1	1	1	0	0
AW	5.042	1.467	3.575	0.592	0.814	0.587	1.159	2.012	1.235	1.014	2.899	1.511	1.129	2.995	2.066	0.929	0.644	2.632	2.035	0.597	4.047	0.995	3.131	0.805	238	1	1	0	1	1	1	1	1	0
AX	4.547	1.314	3.233	0.522	0.738	0.588	0.980	2.923	1.443	1.131	3.341	1.705	1.150	3.394	2.192	1.202	0.468	2.923	2.085	0.838	3.433	1.115	2.861	0.941	189	1	1	0	1	0	1	0	0	0
AZ	5.494	1.755	3.739	0.598	0.972	0.470	1.055	3.655	1.625	1.344	3.851	1.927	1.300	3.334	2.215	1.119	0.747	3.017	2.166	0.851	4.190	1.304	3.434	1.388	245	1	1	0	1	1	1	1	1	0
BA	4.852	1.539	3.313	0.660	0.820	0.758	1.136	2.955	1.394	1.180	3.279	1.836	1.197	3.207	2.166	1.042	0.425	2.673	1.906	0.767	3.832	1.020	2.990	0.834	231	1	0	0	1	1	1	1	0	0
BD	4.487	1.378	3.109	0.485	0.689	0.587	0.991	2.555	1.278	1.024	2.870	1.608	1.174	3.171	2.062	1.109	0.470	2.875	2.203	0.672	3.624	0.863	3.008	0.769	259	1	1	0	1	0	1	0	0	0
BF	5.016	1.555	3.461	0.589	0.931	0.627	0.969	2.585	1.430	1.123	3.036	1.492	0.956	3.193	1.930	1.263	0.520	3.060	2.270	0.790	4.074	0.942	3.125	0.832	238	1	1	1	0	0	0	0	0	0
BG	4.578	1.328	3.250	0.496	0.826	0.573	0.904	2.873	1.618	1.387	2.796	1.646	1.138	2.943	1.751	1.192	0.551	2.694	2.033	0.661	3.671	0.907	3.036	0.832	210	1	1	0	1	1	1	1	1	1
BH	4.687	1.275	3.413	0.528	0.848	0.605	0.928	2.590	1.356	1.121	2.895	1.526	1.020	2.859	1.862	0.997	0.440	2.642	1.858	0.784	3.700	0.988	3.007	0.803	252	1	1	0	1	1	1	1	0	0
BI	5.258	1.645	3.613	0.613	0.920	0.595	1.043	2.987	1.443	1.182	3.252	1.585	1.022	3.277	2.309	0.968	0.429	2.629	1.974	0.655	4.070	1.188	3.210	1.045	266	1	1	0	1	1	1	1	-	0
BJ	4.130	1.291	2.840	0.402	0.671	0.449	0.781	2.291	0.971	0.833	2.504	1.132	0.770	-	-	-	0.450	-	-	-	3.156	0.975	2.769	0.947	385	1	0	1	1	1	1	1	1	0
BK	4.871	1.491	3.380	0.507	0.741	0.434	0.838	1.959	1.055	0.861	2.590	1.438	0.966	2.964	2.032	0.932	0.407	2.741	1.991	0.750	4.091	0.780	3.394	0.845	406	1	1	0	1	1	1	1	1	0
BL	3.699	1.066	2.633	0.428	0.680	0.534	0.823	1.924	0.945	0.809	2.239	1.149	0.769	2.806	1.991	0.815	0.519	2.133	1.631	0.502	3.044	0.656	2.347	0.619	364	1	1	0	1	1	1	1	1	0
BM	5.205	1.486	3.719	0.557	0.803	0.557	0.967	2.263	0.972	0.812	2.532	1.187	0.823	2.940	1.948	0.992	0.522	2.697	1.874	0.823	4.329	0.877	3.512	0.772	301	0	0	1	1	1	1	1	-	0
BN	5.007	1.487	3.520	0.522	0.793	0.607	1.031	3.219	1.581	1.301	3.677	1.664	1.063	3.014	2.029	0.985	0.589	2.649	1.890	0.759	3.935	1.072	3.173	0.878	252	1	1	0	0	0	0	0	0	0
BO	4.477	1.298	3.179	0.508	0.791	0.555	0.861	2.682	1.475	1.249	3.106	1.804	1.219	3.204	1.921	1.283	0.623	2.844	2.085	0.759	3.652	0.825	2.984	0.705	203	1	1	0	0	0	0	0	0	0
BQ	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1	1	1	1	0	0
BR	4.944	1.293	3.651	0.535	0.896	0.676	1.087	3.277	1.337	1.088	3.363	1.577	0.988	2.950	1.917	1.033	0.421	2.716	1.971	0.745	3.776	1.168	2.971	0.826	364	1	1	0	0	1	-	1	1	0
BS	4.877	1.464	3.413	0.522	0.826	0.501	0.968	2.738	1.207	0.997	2.949	1.538	1.032	3.231	2.228	1.003	0.611	2.587	2.077	0.510	3.882	0.995	3.205	0.955	238	1	0	1	0	0	-	0	-	0
BT	4.702	1.349	3.353	0.440	0.721	0.481	0.827	2.522	1.060	0.959	2.768	1.333	0.827	2.674	1.720	0.954	0.369	2.488	1.750	0.738	3.788	0.914	2.934	0.708	364	1	1	0	1	1	1	1	1	0
BU	5.690	1.868	3.822	0.683	0.951	0.649	1.113	3.964	1.919	1.545	4.358	2.260	1.577	3.516	2.517	0.999	0.603	3.097	2.141	0.956	4.231	1.459	3.487	1.23	259	1	1	0	1	1	1	1	1	0
BV	5.034	1.782	3.252	0.620	0.891	0.499	0.873	2.608	1.076	0.913	2.630	1.478	0.924	2.994	1.945	1.049	0.531	2.567	1.993	0.574	4.108	0.926	3.059	0.839	287	1	0	1	0	1	-	1	1	0
BW	4.139	1.163	2.976	0.390	0.674	0.536	0.846	2.640	1.149	0.993	2.770	1.334	0.919	2.987	1.800	1.187	0.621	2.373	1.709	0.664	3.168	0.971	2.570	0.68	315	1	1	0	0	1	1	1	1	0
BX	4.547	1.212	3.336	0.480	0.757	0.523	0.874	2.624	1.365	1.128	2.856	1.453	1.028	2.887	1.770	1.117	0.414	2.479	1.770	0.709	3.572	0.975	2.669	0.85	259	1	1	0	0	1	-	1	-	0
BY	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1	1	1	1	0	0
BZ	4.330	1.420	2.910	0.452	0.671	0.611	0.988	2.079	0.982	0.909	2.551	1.203	0.965	2.879	1.863	1.016	0.409	2.699	1.895	0.804	3.402	0.928	2.970	0.837	280	1	1	0	0	1	-	1	1	0
CA	3.859	1.173	2.686	0.512	0.653	0.496	0.758	2.091	0.947	0.750	2.253	1.254	0.868	2.668	1.724	0.944	0.426	2.456	1.695	0.761	3.211	0.648	2.540	0.724	266	1	1	0	0	0	0	0	0	0
CB	5.307	1.467	3.840	0.593	0.824	0.547	0.875	2.065	0.991	0.794	2.579	1.215	0.832	3.190	1.883	1.307	0.622	2.854	2.213	0.641	4.402	0.905	3.543	0.808	210	1	0	0	0	1	-	1	1	0
CC	5.180	1.481	3.699	0.517	0.762	0.629	0.998	2.651	1.279	1.047	2.804	1.458	1.010	3.475	2.130	1.345	0.595	2.803	2.054	0.749	4.012	1.168	3.166	1.06	210	1	1	0	0	1	-	1	1	0
CD	5.331	1.508	3.823	0.550	0.842	0.556	0.963	2.470	1.244	1.035	3.284	1.450	0.893	3.228	1.883	1.345	0.885	3.045	2.007	1.038	4.061	1.270	3.154	0.955	224	1	1	0	1	1	1	1	0	0
CE	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1	1	1	1	0	0
CF	4.702	1.295	3.406	0.582	0.933	0.669	1.062	3.086	1.380	1.175	3.472	1.706	1.130	2.998	2.065	0.933	0.501	2.956	2.224	0.732	3.594	1.108	3.250	1.248	364	1	1	0	1	1	1	1	0	0
CG	4.010	1.368	2.642	0.412	0.821	0.531	0.914	3.065	1.545	1.215	3.105	1.574	1.033	2.999	2.050	0.949	0.485	2.816	2.082	0.734	2.865	1.145	3.032	0.915	259	1	1	0	0	1	-	1	-	0
CH	5.706	1.614	4.092	0.589	0.817	0.689	1.075	2.787	1.318	0.975	3.129	1.565	0.966	3.571	2.435	1.136	0.631	3.112	2.144	0.968	4.529	1.177	3.456	1.024	210	1	1	1	0	0	-	1	1	0
CI	4.483	1.466	3.017	0.438	0.708	0.506	0.849	2.697	1.210	1.780	3.147	1.398	1.066	3.487	2.120	1.367	0.560	3.153	2.307	0.846	3.428	1.055	3.051	0.806	203	1	0	1	0	0	0	0	0	0
CJ	4.529	1.155	3.374	0.523	0.817	0.593	0.977	2.288	1.163	0.993	2.632	1.418	1.027	3.197	1.928	1.269	0.557	2.586	1.867	0.719	3.523	1.006	2.775	0.902	231	1	1	0	0	1	1	1	1	0
CK	5.225	1.435	3.790	0.718	0.933	0.654	1.053	2.897	1.479	1.231	3.629	1.804	1.333	2.784	1.932	0.852	0.641	2.506	1.739	0.767	4.269	0.956	3.230	0.933	280	1	1	0	0	1	1	1	1	0
CL	4.661	1.506	3.155	0.557	0.802	0.518	0.838	2.549	1.107	0.792	2.820	1.417	0.968	3.014	1.936	1.078	0.518	2.669	1.959	0.711	3.628	1.033	3.174	0.818	455	1	1	0	1	1	1	1	-	0
CM	3.922	1.304	2.618	0.498	0.693	0.551	0.834	2.823	1.209	1.072	3.036	1.549	1.151	-	-	-	0.371	-	-	-	2.832	1.090	2.587	0.785	420	0	0	1	1	1	1	1	1	0
CN	4.580	1.441	3.139	0.407	0.659	0.450	0.822	2.206	1.024	0.820	2.650	1.411	0.952	3.130	2.020	1.111	0.389	2.627	1.951	0.676	3.682	0.897	2.825	0.789	385	1	0	0	1	1	1	1	-	0
CP	4.273	1.268	3.004	0.539	0.762	0.528	0.856	2.495	1.088	0.926	2.779	1.274	0.867	2.883	2.153	0.730	0.404	2.581	1.946	0.635	3.430	0.843	2.884	1.021	315	1	1	0	1	1	1	1	-	0
CQ	4.158	1.349	2.809	0.452	0.740	0.432	0.743	2.057	0.887	0.683	2.372	1.225	0.845	2.684	1.728	0.955	0.481	2.371	1.705	0.666	3.333	0.825	2.676	0.751	455	0	0	0	1	1	1	1	1	0
CR	4.668	1.394	3.274	0.381	0.686	0.423	0.682	1.580	0.800	0.682	2.271	1.164	0.830	2.982	1.555	1.427	0.413	2.754	2.052	0.702	3.822	0.845	3.329	0.77	455	1	0	0	1	1	1	1	0	0
CT	4.867	1.203	3.665	0.469	0.712	0.599	0.991	2.398	1.014	0.904	2.797	1.385	0.948	2.782	1.962	0.820	0.615	2.685	1.951	0.734	3.782	1.086	2.974	0.981	364	1	1	0	1	1	1	1	-	0

	Trait numbers																																	
Qual.	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**	**25**	**26**	**27**	**28**	**29.1**	**29.2**	**29.3**	**29.4**	**30**	**31**
CU	4.467	1.315	3.153	0.489	0.723	0.501	0.810	2.529	1.029	0.899	2.917	1.381	0.962	2.798	1.726	1.072	0.475	2.407	1.697	0.710	3.503	0.964	2.943	0.651	364	1	1	0	0	0	-	1	1	0
CV	4.397	1.681	2.716	0.550	0.811	0.474	0.845	2.718	1.367	1.192	3.254	1.781	1.276	3.025	1.873	1.152	0.454	2.829	2.053	0.776	3.607	0.790	3.228	0.908	266	1	0	1	1	1	1	1	1	0
CW	4.362	1.220	3.142	0.558	0.861	0.610	0.915	2.292	1.256	1.027	2.741	1.468	0.998	2.748	1.883	0.865	0.564	2.485	1.769	0.716	3.434	0.928	2.869	0.839	287	1	1	0	1	1	1	1	0	0
CX	4.686	1.492	3.194	0.437	0.697	0.594	0.947	2.809	1.285	1.106	2.928	1.548	1.160	3.299	1.967	1.332	0.464	2.892	2.140	0.752	3.691	0.995	3.000	0.95	203	1	1	0	1	1	1	1	1	0
CY	4.413	1.411	3.002	0.521	0.731	0.579	0.938	3.473	1.622	1.365	3.881	1.919	1.366	2.907	1.969	0.938	0.604	2.457	1.649	0.808	3.378	1.035	2.926	0.802	210	1	1	0	0	0	0	0	0	0
DA	5.406	1.579	3.827	0.548	0.797	0.494	0.897	2.266	1.097	0.920	2.690	1.254	0.944	2.762	1.825	0.937	0.594	2.463	1.775	0.688	4.441	0.965	3.123	1.061	224	1	1	1	0	0	-	1	1	0
DB	5.200	1.452	3.748	0.468	0.785	0.576	0.964	2.740	1.202	0.947	3.091	1.502	1.074	3.020	2.021	0.999	0.767	2.779	2.051	0.728	4.164	1.036	3.189	1.041	238	1	0	0	0	0	-	1	-	0
DD	5.802	1.915	3.887	0.603	0.987	0.643	1.091	2.775	1.252	1.035	3.080	1.396	0.931	3.140	2.241	0.899	0.923	2.904	2.120	0.784	4.681	1.121	3.554	1.052	210	1	1	0	0	1	1	1	1	0
DE	5.829	1.451	4.378	0.613	0.817	0.646	1.053	2.800	1.167	0.745	3.833	1.767	1.244	3.262	2.212	1.050	0.568	2.860	2.035	0.825	4.652	1.177	3.500	1.192	245	1	1	0	1	1	1	1	1	0
DF	5.211	1.601	3.610	0.476	0.776	0.596	1.134	3.599	1.518	1.222	3.715	1.812	1.316	3.440	2.283	1.157	0.722	3.094	2.179	0.915	3.976	1.235	3.190	1.318	224	1	1	0	1	1	1	1	-	0
DG	4.720	1.454	3.266	0.502	0.828	0.541	0.941	2.245	1.058	0.902	2.825	1.390	1.033	2.882	1.960	0.922	0.456	2.620	1.856	0.764	3.760	0.960	3.224	0.996	266	1	0	0	1	1	1	1	0	0
DH	5.034	1.521	3.513	0.507	0.812	0.585	0.976	3.476	1.535	1.251	3.554	1.762	1.180	3.448	2.226	1.222	0.653	2.855	2.153	0.702	3.993	1.041	3.288	0.86	203	0	0	1	1	0	1	0	0	0
DI	5.262	1.885	3.377	0.514	0.819	0.638	0.934	2.167	1.141	0.956	2.618	1.329	1.002	3.199	2.280	0.919	0.764	2.657	1.982	0.675	4.266	0.996	3.271	1.004	210	1	0	1	0	1	-	1	-	0
DJ	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1	1	1	1	0	0
DM	5.069	1.506	3.563	0.504	0.851	0.544	0.790	2.523	1.128	0.888	2.820	1.264	0.943	2.932	1.964	0.968	0.800	2.633	2.013	0.620	4.024	1.045	3.308	0.986	224	1	0	0	1	1	1	1	1	0
DO	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1	1	1	1	0	0
DS	4.434	1.344	3.090	0.483	0.753	0.611	0.861	2.288	1.076	0.883	2.423	1.284	0.888	2.780	1.831	0.949	0.506	2.722	2.068	0.654	3.615	0.819	2.846	0.835	364	1	1	0	1	1	1	1	1	0
DT	4.208	1.264	2.944	0.439	0.651	0.516	0.847	2.618	1.104	0.803	2.812	1.449	0.978	3.270	2.175	1.095	0.603	2.923	2.227	0.696	3.167	1.041	2.764	0.805	182	0	0	1	0	0	-	1	-	0
DU	5.184	1.495	3.689	0.611	0.769	0.585	0.943	2.567	1.386	1.038	3.213	1.707	1.221	3.261	2.275	0.986	0.813	2.864	2.081	0.783	4.124	1.060	3.004	1.038	259	1	1	0	1	1	1	1	0	0
DX	4.584	1.272	3.311	0.482	0.710	0.489	0.803	2.973	0.849	0.981	3.041	1.359	0.979	2.809	1.977	0.832	0.410	2.560	1.897	0.663	3.526	1.058	3.048	0.989	315	1	1	0	1	1	1	1	1	0
DZ	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	455	1	0	1	0	1	-	1	1	0
EA	4.199	1.183	3.015	0.428	0.785	0.439	0.798	2.691	1.238	1.003	2.892	1.434	0.920	2.653	1.678	0.975	0.335	2.217	1.481	0.736	3.233	0.965	2.508	0.967	315	1	1	0	0	0	-	1	1	0
EB	4.283	1.361	2.922	0.429	0.630	0.418	0.725	2.313	1.090	0.824	2.666	1.411	0.863	2.964	1.753	1.211	0.412	2.761	1.813	0.948	3.524	0.759	2.890	0.819	196	0	0	1	1	1	1	1	1	0
EC	5.168	1.570	3.598	0.526	0.785	0.531	0.970	2.792	1.201	0.960	3.283	1.560	1.201	3.068	2.043	1.025	0.495	2.705	2.043	0.662	4.220	0.948	3.336	0.913	266	1	0	0	1	1	1	1	1	0
EE	5.809	1.662	4.147	0.544	0.865	0.615	0.971	3.226	1.368	0.996	3.683	1.665	1.245	3.562	2.202	1.360	0.550	3.085	2.318	0.767	4.593	1.216	3.364	1.173	231	1	1	0	0	1	1	1	1	0
EF	4.864	1.570	3.294	0.543	0.882	0.665	0.977	2.776	1.162	1.011	2.926	1.586	1.134	3.298	2.111	1.187	0.528	3.030	2.411	0.619	3.978	0.886	3.497	0.85	210	1	0	0	1	1	1	1	-	1
EG	3.905	1.571	2.334	0.424	0.722	0.532	0.836	3.090	1.265	1.035	3.019	1.438	1.043	3.064	2.215	0.849	0.614	3.010	2.151	0.859	2.845	1.060	2.852	1.032	266	0	0	1	1	0	1	1	0	0
EH	4.332	1.371	2.961	0.499	0.769	0.568	0.852	2.655	1.081	0.858	3.008	1.505	1.023	3.032	2.013	1.019	0.783	2.563	1.990	0.573	3.371	0.961	2.781	1.008	210	1	0	1	0	1	-	-	1	0
EK	3.630	1.223	2.407	0.389	0.672	0.444	0.737	2.412	1.077	0.939	2.663	1.406	0.942	2.683	1.738	0.946	0.418	2.096	1.614	0.482	3.015	0.615	2.481	0.607	476	1	1	0	1	1	1	1	0	0
EN	4.948	1.602	3.346	0.588	0.828	0.557	0.925	2.964	1.132	0.933	3.244	1.613	1.119	3.471	2.357	1.114	0.674	3.283	2.499	0.784	3.820	1.128	3.371	0.984	210	1	0	1	1	1	1	1	1	0
EO	5.204	1.801	3.403	0.589	0.944	0.621	0.957	3.632	1.424	1.117	3.842	1.589	1.159	3.035	2.090	0.945	0.892	2.862	2.147	0.715	4.016	1.188	3.724	1.126	224	1	0	0	0	1	-	1	1	0
EQ	3.268	0.824	2.444	0.634	1.045	0.413	0.739	1.426	0.629	0.575	1.553	0.767	0.572	-	-	-	0.534	-	-	-	2.715	0.553	2.780	0.919	364	1	1	0	1	1	1	1	0	0
ES	4.754	1.448	3.306	0.603	0.891	0.667	0.971	2.664	1.138	0.891	3.252	1.549	1.076	2.771	1.844	0.927	0.625	2.577	1.855	0.722	3.760	0.994	2.888	0.887	287	1	0	0	0	1	-	1	1	0
ET	4.755	1.773	2.982	0.520	0.798	0.606	0.844	2.258	1.143	0.830	2.822	1.578	1.054	3.166	2.198	0.968	0.698	2.789	2.112	0.677	3.895	0.860	3.111	1.088	224	1	1	1	0	0	0	0	0	0
EU	4.644	1.447	3.197	0.592	0.747	0.617	0.932	2.721	1.164	0.875	2.964	1.481	1.038	2.616	1.840	0.776	0.406	2.426	1.816	0.610	3.751	0.893	2.976	0.997	210	1	1	0	0	0	0	0	0	0
EV	5.381	1.729	3.652	0.597	1.007	0.632	0.951	2.641	1.202	0.995	2.873	1.419	1.012	3.178	2.248	0.930	0.763	3.134	2.165	0.969	4.337	1.044	4.158	1.153	238	1	1	0	0	1	1	0	1	0
EW	4.616	1.359	3.257	0.543	0.820	0.473	0.807	2.365	1.164	0.966	2.775	1.408	0.944	2.945	1.928	1.017	0.378	2.599	1.962	0.637	3.722	0.894	2.987	0.899	455	1	1	0	1	1	1	1	0	0
EX	5.315	1.393	3.922	0.646	0.913	0.656	1.007	2.188	1.167	0.950	2.922	1.515	1.144	2.742	1.930	0.812	0.611	2.478	1.777	0.701	4.204	1.111	3.231	1.124	259	1	1	1	0	0	-	0	1	0
EY	5.725	1.675	4.050	0.580	0.913	0.653	1.006	3.064	1.362	1.032	3.481	1.581	1.164	3.254	2.319	0.935	0.896	3.071	2.412	0.659	4.612	1.113	3.475	1.346	224	0	0	1	0	0	-	1	1	0
EZ	4.051	1.362	2.689	0.453	0.690	0.577	0.897	2.687	0.996	0.809	2.863	1.453	1.090	2.951	1.995	0.956	0.538	2.519	1.824	0.695	3.255	0.796	2.964	0.762	287	0	0	1	0	0	-	1	1	0
FA	3.928	1.290	2.638	0.452	0.751	0.482	0.806	2.395	1.047	0.857	2.552	1.370	0.904	2.794	1.801	0.993	0.430	2.373	1.749	0.624	3.148	0.781	2.701	0.869	364	1	0	0	1	1	1	1	1	0
FB	4.366	1.172	3.194	0.598	0.823	0.532	0.890	2.772	1.142	0.908	3.120	1.428	0.879	2.631	1.635	0.996	0.392	2.370	1.677	0.693	3.397	0.968	2.597	0.975	364	0	0	1	1	1	1	1	1	0
FC	5.232	1.956	3.276	0.496	0.864	0.569	0.963	2.888	1.243	0.982	3.404	1.695	1.172	3.147	2.197	0.950	0.587	3.021	2.338	0.683	4.167	1.065	3.826	1.018	238	1	0	0	0	1	1	1	1	0
FD	4.093	1.094	2.999	0.540	0.707	0.506	0.888	2.155	1.090	0.869	2.543	1.532	0.907	2.989	1.908	1.081	0.459	2.530	1.772	0.758	3.190	0.903	2.535	0.665	266	1	1	0	1	1	1	1	0	1

	Trait numbers																																	
Qual.	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**	**25**	**26**	**27**	**28**	**29.1**	**29.2**	**29.3**	**29.4**	**30**	**31**
FE	3.946	1.310	2.636	0.497	0.775	0.551	0.801	1.748	0.918	0.782	2.080	1.247	0.871	2.656	1.509	1.147	0.350	-	-	-	3.126	0.820	2.504	0.742	399	1	1	1	1	1	1	1	0	0
FF	5.690	1.612	4.078	0.630	0.923	0.673	1.052	3.000	1.241	0.973	3.722	1.785	1.145	2.896	2.098	0.798	0.670	2.705	1.804	0.901	4.512	1.178	3.331	1.253	266	0	0	1	1	1	1	1	1	0
FG	4.927	1.511	3.416	0.567	0.887	0.584	0.837	2.859	1.275	0.986	3.005	1.556	1.037	3.074	2.115	0.959	0.495	2.506	1.777	0.729	4.052	0.875	3.164	1.044	238	1	1	0	1	1	1	1	1	1
FH	5.676	1.592	4.084	0.632	0.987	0.551	0.809	2.736	1.221	0.927	3.150	1.560	0.942	3.088	2.202	0.886	0.488	2.620	1.935	0.685	4.734	0.942	3.467	1.055	203	1	0	0	0	0	-	0	1	0
FI	5.198	1.594	3.604	0.522	0.832	0.659	0.908	2.564	1.179	0.947	3.030	1.542	1.043	2.921	1.989	0.932	0.613	2.758	2.034	0.724	4.066	1.132	3.292	1.035	287	1	1	0	1	1	1	1	0	0
FK	4.096	1.289	2.807	0.466	0.599	0.520	0.857	2.407	1.169	0.922	2.896	1.348	0.847	2.956	2.074	0.882	0.534	2.701	1.984	0.717	3.057	1.039	2.820	0.82	280	1	1	0	0	1	1	1	1	0
FL	4.284	1.218	3.066	0.490	0.741	0.518	0.830	2.481	1.092	0.849	2.545	1.467	1.048	2.927	2.045	0.882	0.499	2.583	1.803	0.780	3.528	0.756	2.741	0.746	245	1	1	0	0	1	-	1	-	0
FN	4.762	1.588	3.174	0.572	0.860	0.614	0.841	3.120	1.438	1.147	3.186	1.719	1.196	-	-	-	0.643	-	-	-	3.755	1.007	3.192	0.945	210	1	1	1	1	1	1	1	1	0
FO	4.300	1.280	3.021	0.497	0.745	0.507	0.800	2.153	1.005	0.838	2.481	1.320	0.865	2.675	1.732	0.944	0.471	2.478	1.850	0.628	3.539	0.761	2.894	0.802	287	0	0	1	0	1	1	1	1	0
FR	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	315	-	-	-	1	1	1	1	0	0
FS	4.473	1.342	3.131	0.519	0.767	0.609	0.759	2.484	1.009	0.747	2.862	1.411	0.823	2.878	-	-	0.558	-	-	-	3.344	1.129	2.551	0.957	315	1	1	0	1	1	1	1	0	0
FT	4.485	1.282	3.203	0.487	0.755	0.518	0.767	2.287	0.898	0.678	2.957	1.217	0.761	2.689	1.732	0.957	0.464	2.565	1.864	0.701	3.461	1.024	2.663	1	238	1	0	0	0	0	0	1	0	0
FU	5.626	1.791	3.835	0.597	0.913	0.697	1.025	3.548	1.383	0.988	3.906	1.829	1.161	3.462	2.300	1.162	0.601	3.183	2.427	0.756	4.527	1.099	3.670	1.128	280	1	0	1	0	0	-	0	0	0
FV	3.729	1.117	2.611	0.425	0.691	0.434	0.692	1.812	0.963	0.833	2.274	1.190	0.776	2.345	1.428	0.917	0.326	2.448	1.796	0.652	2.996	0.733	2.552	0.713	364	0	0	1	1	1	1	1	1	0
FX	5.416	1.609	3.807	0.560	0.858	0.617	0.995	2.905	1.318	1.115	3.444	1.856	1.219	3.071	2.013	1.058	0.483	2.780	2.146	0.634	4.456	0.960	3.451	1.144	364	1	1	0	1	1	1	1	1	0
FY	5.729	1.694	4.035	0.662	0.906	0.602	1.032	3.367	1.226	0.922	3.680	1.550	1.055	3.471	2.475	0.996	0.598	2.885	2.169	0.716	4.468	1.261	3.420	1.228	259	1	1	0	0	0	0	0	1	0
GA	4.470	1.238	3.233	0.487	0.793	0.514	0.856	2.888	1.213		3.115	1.503	1.050	3.128	1.963	1.165	0.480	2.619	1.811	0.808	3.337	1.133	2.914	0.871	364	1	0	1	1	1	1	1	1	0
GB	4.181	1.209	2.972	0.523	0.695	0.549	0.906	2.845	1.193	1.027	2.909	1.481	0.972	2.607	1.696	0.911	0.407	2.358	1.694	0.664	3.148	1.033	2.720	1.165	280	1	1	0	0	1	-	1	1	0
GC	5.550	1.689	3.861	0.560	0.915	0.648	1.032	3.523	1.452	1.130	3.767	1.704	1.163	3.469	2.231	1.238	0.960	3.041	2.199	0.842	4.365	1.185	3.518	1.242	231	1	1	0	0	1	1	1	1	0
GF	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1	1	1	1	0	0
GG	5.242	1.410	3.832	0.550	0.802	0.502	0.851	3.136	1.288	0.996	3.174	1.473	0.894	2.984	2.084	0.900	0.468	2.426	1.846	0.580	4.077	1.165	3.197	1.085	189	0	0	1	0	0	0	0	0	0
GH	4.464	1.289	3.175	0.475	0.771	0.488	0.806	2.851	1.193	0.927	2.946	1.548	0.943	2.621	1.662	0.959	0.456	2.148	1.518	0.630	3.554	0.909	2.921	0.78	224	1	0	0	0	1	-	1	1	0
GI	5.214	1.690	3.524	0.539	0.821	0.512	0.890	2.780	1.098	0.835	3.098	1.430	0.919	2.986	2.036	0.950	0.705	2.792	2.079	0.713	4.183	1.031	3.393	0.909	238	1	1	0	0	0	0	0	0	0
GJ	4.389	1.467	2.922	0.529	0.764	0.535	0.803	2.558	1.066	0.874	2.794	1.427	0.886	3.071	1.993	1.077	0.489	2.399	1.918	0.481	3.495	0.894	2.811	0.955	371	0	0	1	1	0	1	1	1	0
GK	5.112	1.541	3.571	0.694	0.848	0.565	0.893	2.962	1.210	0.930	3.222	1.660	1.060	3.208	2.084	1.124	0.463	2.808	2.095	0.713	3.950	1.163	3.106	0.939	420	1	1	0	1	1	1	1	-	0
GM	4.285	1.413	2.872	0.446	0.709	0.514	0.892	3.156	1.282	1.027	3.193	1.487	1.039	3.241	2.048	1.193	0.636	3.108	2.388	0.720	3.222	1.063	3.018	0.961	224	1	1	0	0	0	-	1	1	0
GN	5.258	1.553	3.705	0.565	0.859	0.603	0.948	2.984	1.185	1.036	3.260	1.813	1.143	3.366	2.258	1.108	0.575	2.825	2.125	0.700	4.286	0.972	3.364	0.866	238	1	1	0	1	1	1	1	0	0
GO	3.414	1.136	2.278	0.459	0.601	0.520	0.776	2.439	1.071	0.928	2.702	1.415	0.974	2.520	1.648	0.872	0.450	2.181	1.706	0.475	2.596	0.818	2.310	0.66	301	1	0	1	0	1	-	1	1	0
GP	5.404	1.572	3.832	0.589	0.865	0.481	0.825	2.628	1.223	0.989	2.898	1.473	0.984	3.256	2.352	0.904	0.529	2.653	1.922	0.731	4.508	0.896	3.434	0.959	364	1	1	1	0	0	-	1	1	0
GQ	4.384	1.400	2.985	0.455	0.747	0.430	0.872	2.181	0.887	0.777	2.890	1.466	1.145	2.722	1.709	1.013	0.529	2.253	1.725	0.528	4.354	0.866	2.811	0.894	455	1	0	0	1	1	1	1	0	0
GR	4.807	1.430	3.377	0.502	0.823	0.528	0.830	2.829	1.108	0.859	3.109	1.423	0.894	2.951	1.997	0.954	0.493	2.936	2.178	0.758	3.750	1.057	3.434	0.975	280	1	1	0	1	1	1	1	1	0
GS	4.694	1.398	3.296	0.489	0.789	0.558	0.919	2.932	1.118	0.895	3.179	1.505	0.918	3.234	2.296	0.938	0.449	3.043	2.241	0.802	3.695	0.999	3.267	1.217	245	1	0	1	1	1	1	1	1	0
GU	3.826	1.295	2.531	0.411	0.661	0.419	0.689	1.919	0.889	0.691	2.072	1.113	0.695	2.956	1.991	0.965	0.440	2.474	1.835	0.639	3.057	0.769	2.777	0.8	364	1	1	0	1	1	1	1	1	0
GV	5.109	1.565	3.544	0.561	0.798	0.635	1.004	3.244	1.205	0.923	3.678	1.790	1.159	2.833	1.930	0.903	0.665	2.533	1.941	0.592	4.020	1.089	3.346	1.036	210	1	0	0	0	0	-	0	1	0
GW	4.242	1.325	2.917	0.376	0.655	0.437	0.798	2.461	1.026	0.851	2.746	1.311	0.854	2.782	1.825	0.957	0.402	2.568	1.861	0.707	3.266	0.976	2.708	0.857	266	1	0	0	1	1	1	1	0	0
GX	4.337	1.373	2.964	0.553	0.856	0.528	0.924	2.702	1.142	0.948	2.879	1.410	0.856	2.875	1.894	0.981	0.416	2.725	2.066	0.659	3.420	0.917	2.914	0.901	238	1	0	0	1	1	1	1	0	1
GY	4.600	1.438	3.163	0.520	0.853	0.594	1.083	2.933	-	-	3.139	1.536	-	3.287	2.145	1.142	0.412	2.840	2.103	0.737	3.540	1.060	3.158	1.064	266	1	1	0	0	0	0	0	1	0
GZ	5.170	1.629	3.541	0.503	0.947	0.443	0.825	2.218	1.064	0.914	2.441	1.199	0.763	3.044	2.016	1.028	0.483	2.739	2.064	0.675	4.032	0.775	3.534	1.129	392	1	1	1	1	1	1	1	1	0
HA	4.660	1.542	3.118	0.516	0.719	0.566	1.030	3.058	1.220	0.930	3.304	1.647	0.995	2.946	1.901	1.045	0.539	2.385	2.041	0.344	3.602	1.058	3.357	0.846	252	0	0	1	0	0	0	0	0	0
HB	4.947	1.393	3.554	0.624	0.927	0.696	1.098	3.044	1.239	1.006	3.284	1.593	1.004	2.946	1.978	0.968	0.496	2.720	1.887	0.833	3.910	1.037	3.087	1.091	280	1	0	1	1	1	1	1	1	0
HC	4.875	1.490	3.385	0.761	0.769	0.542	0.930	3.218	1.196	0.954	3.372	1.483	0.944	3.130	2.032	1.098	0.432	3.127	2.397	0.730	3.762	1.113	3.333	0.858	245	0	0	1	1	1	1	1	1	0
HD	5.523	1.654	3.869	0.491	0.897	0.714	1.073	3.541	1.264	0.979	3.666	1.822	1.196	3.102	2.124	0.978	0.800	2.823	1.982	0.841	4.284	1.239	3.408	1.14	238	0	0	0	1	1	1	1	0	0
HE	4.101	1.183	2.918	0.495	0.689	0.596	0.873	2.917	1.091	0.899	2.946	1.396	0.933	3.098	1.944	1.154	0.422	2.616	1.900	0.716	2.971	1.130	2.671	0.865	266	1	1	0	1	1	1	1	0	0
HF	5.428	1.583	3.845	0.477	0.881	0.637	1.032	3.252	1.171	0.926	3.860	1.725	1.076	3.347	2.293	1.054	0.716	2.972	2.326	0.646	4.220	1.208	3.452	1.175	203	0	0	1	0	1	-	1	1	0

	Trait numbers																																		
Qual.	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**	**25**	**26**	**27**	**28**	**29.1**	**29.2**	**29.3**	**29.4**	**30**	**31**
HG	4.229	1.340	2.889	0.422	0.766	0.436	0.749	2.252	0.997	0.840	2.378	1.285	0.821	3.078	2.132	0.946	0.350	2.799	2.123	0.676	3.411	0.818	2.886	0.764	217	1	1	1	0	0	-	0	1	0
HH	4.285	1.265	3.020	0.478	0.792	0.446	0.856	2.386	1.114	0.910	2.647	1.387	0.866	2.869	1.863	1.006	0.398	2.298	1.787	0.511	3.386	0.899	2.881	0.779	217	1	1	0	0	0	0	0	0	0
HI	5.121	1.707	3.414	0.646	0.808	0.735	1.233	3.135	1.357	1.060	3.395	1.833	1.162	3.063	2.120	0.943	0.637	2.615	1.669	0.946	3.961	1.160	3.022	1.035	168	1	1	1	0	1	-	1	1	0
HJ	5.182	1.587	3.595	0.450	0.790	0.471	0.835	2.337	1.117	0.917	2.858	1.309	0.806	3.369	2.384	0.985	0.622	3.165	2.378	0.787	4.190	0.992	3.244	1.063	154	1	0	1	0	0	-	0	0	0
HK	4.599	1.424	3.175	0.462	0.749	0.532	0.865	2.685	1.112	0.907	3.089	1.357	0.901	2.959	1.871	1.088	0.511	2.671	1.900	0.771	3.491	1.108	2.849	1.112	182	1	0	1	0	1	-	1	1	0
HM	5.468	1.540	3.928	0.579	0.865	0.530	0.830	2.792	1.255	0.928	3.455	1.784	1.117	3.261	2.161	1.100	0.483	2.843	2.188	0.655	4.233	1.235	3.442	1.113	294	1	1	0	0	1	1	1	1	0
HP	5.097	1.708	3.389	0.514	0.758	0.562	0.901	3.062	1.342	1.104	3.611	1.668	0.967	2.732	1.961	0.771	0.467	3.010	2.259	0.751	4.153	0.944	3.560	0.969	189	1	1	1	0	0	-	0	1	0
HR	4.979	1.436	3.544	0.542	0.877	0.553	0.907	2.536	1.179	0.873	2.908	1.524	0.954	2.941	1.853	1.088	0.339	2.727	1.959	0.768	3.998	0.981	3.112	1.045	343	1	1	1	1	1	1	1	1	0
HS	4.659	1.549	3.110	0.448	0.773	0.665	0.899	2.896	1.222	1.025	3.084	1.551	0.894	3.077	2.238	0.839	0.439	2.859	2.098	0.761	3.784	0.875	3.438	0.968	210	1	0	0	0	0	0	0	0	0
HT	4.271	1.540	2.731	0.560	0.824	0.576	0.888	2.735	1.116	0.895	2.771	1.329	0.849	2.964	1.948	1.016	0.607	2.880	2.185	0.695	3.265	1.006	2.975	0.882	168	0	0	1	0	0	-	0	1	0
HU	4.579	1.493	3.086	0.477	0.722	0.465	0.777	2.624	1.089	0.889	2.803	1.402	0.844	2.333	1.524	0.809	0.381	2.074	1.592	0.482	3.893	0.686	2.953	0.876	364	1	0	1	1	1	1	1	0	0
HV	4.839	1.604	3.235	0.514	0.819	0.562	0.857	2.556	1.099	0.859	2.992	1.443	0.788	2.932	1.998	0.934	0.457	2.915	2.212	0.703	3.822	1.017	3.283	1.002	385	1	0	1	1	1	1	1	1	0
HW	5.587	1.759	3.828	0.769	0.959	0.735	1.088	2.582	1.131	0.900	3.049	1.520	0.862	2.977	2.156	0.821	0.548	2.842	2.094	0.748	4.310	1.277	3.491	1.229	294	1	0	1	1	1	1	1	1	0
HX	4.454	1.433	3.021	0.541	0.729	0.543	0.783	1.841	0.970	0.756	2.321	1.154	0.754	3.008	1.871	1.136	0.420	2.538	1.888	0.650	3.541	0.913	2.983	0.78	294	1	1	1	1	1	1	1	1	0
HY	4.820	1.577	3.243	0.471	0.827	0.480	0.765	2.167	1.001	0.815	2.745	1.364	0.837	2.859	1.960	0.899	0.508	2.692	2.041	0.652	3.839	0.981	3.254	0.793	385	1	0	0	1	1	1	1	0	0
HZ	5.230	1.792	3.438	0.662	0.958	0.696	1.014	2.910	1.294	0.946	3.339	1.660	1.053	3.150	2.036	1.114	0.757	2.633	1.979	0.654	4.111	1.119	3.080	1.017	168	1	0	0	0	0	-	1	1	0
IA	4.166	1.454	2.712	0.438	0.737	0.488	0.852	2.693	1.132	0.940	2.762	1.430	0.954	2.905	2.150	0.755	0.396	3.004	2.231	0.773	3.289	0.876	3.086	0.873	175	1	0	0	1	1	1	1	0	0
IB	4.530	1.552	2.978	0.451	0.722	0.471	0.794	2.694	1.192	0.937	3.050	1.474	0.963	2.996	1.969	1.027	0.366	2.647	1.937	0.710	3.565	0.965	2.828	0.998	210	1	0	0	0	0	-	0	0	0
ID	5.049	1.706	3.343	0.640	0.924	0.696	1.020	3.717	1.438	1.114	4.061	1.884	1.317	3.226	2.098	1.128	0.753	2.872	2.027	0.845	3.839	1.210	3.396	1.101	168	1	0	0	0	0	-	0	0	0
IF	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1	1	1	1	0	0
IG	4.686	1.456	3.230	0.530	0.784	0.559	0.810	1.969	0.815	0.706	2.369	1.168	0.759	2.952	2.070	0.883	0.437				3.869	0.817	3.132	0.933	294	1	0	1	1	1	1	1	1	0
IH	4.870	1.496	3.374	0.549	0.883	0.513	0.869	1.972	0.916	0.776	2.405	1.238	0.772	2.931	1.855	1.076	0.464	2.623	1.990	0.633	3.832	1.038	3.058	0.948	385	1	1	0	1	1	1	1	1	0
IJ	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1	1	1	1	0	0
IK	6.432	1.856	4.576	0.728	1.100	0.718	1.089	3.433	1.163	0.896	3.279	1.843	1.297	3.392	2.288	1.104	0.861	3.313	2.426	0.887	5.117	1.315	3.944	1.367	154	1	0	1	0	1	-	0	1	0
IL	4.749	1.567	3.182	0.493	0.768	0.493	0.895	2.531	1.070	0.838	3.032	1.427	0.870	2.779	1.982	0.797	0.560	2.582	1.981	0.601	3.693	1.056	3.122	0.723	147	1	0	0	0	0	-	0	1	0
IP	3.699	1.132	2.567	0.416	0.664	0.513	0.751	1.913	0.929	0.769	2.232	1.149	0.718	2.319	1.594	0.725	0.452	-	-	-	3.008	0.691	2.497	0.573	252	1	1	1	0	1	-	1	1	0
IQ	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1	1	1	-	0	0
IR	4.379	1.297	3.082	0.512	0.773	0.461	0.771	2.468	1.098	0.839	2.757	1.381	0.826	2.698	1.717	0.981	0.447	2.492	1.782	0.710	3.490	0.889	2.834	0.787	168	1	1	1	0	1	-	0	1	0
IS	4.558	1.337	3.221	0.562	0.760	0.565	0.883	2.968	1.282	1.046	3.091	1.591	0.954	2.877	1.795	1.082	0.498	2.529	1.871	0.658	3.587	0.971	2.913	0.754	140	1	1	0	0	1	-	1	-	0

Statistics of the quantitative traits

AVG	4.750	1.461	3.289	0.527	0.809	0.560	0.905	2.657	1.200	0.977	2.999	1.504	1.016	3.030	1.997	1.034	0.546	2.708	2.004	0.704	3.762	0.991	3.079	0.918	274.48
STD	0.561	0.195	0.432	0.079	0.098	0.081	0.117	0.479	0.204	0.182	0.467	0.228	0.177	0.250	0.205	0.139	0.132	0.254	0.211	0.111	0.468	0.165	0.332	0.166	77.36

Appendix 6.6

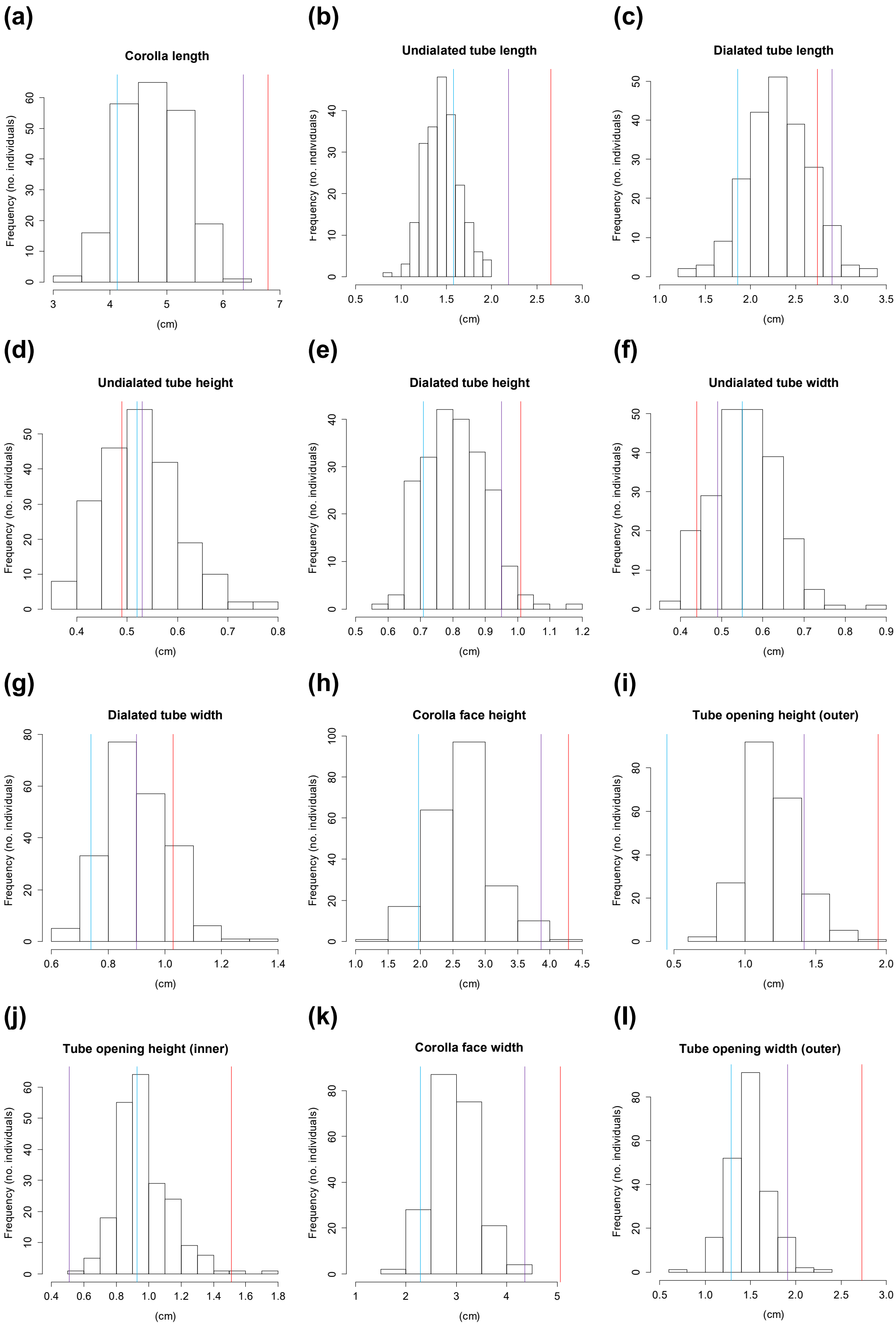
Table. Summary of the Shapiro-Wilk normality test results of the quantitative traits measured for the BC population (*N* = 200)

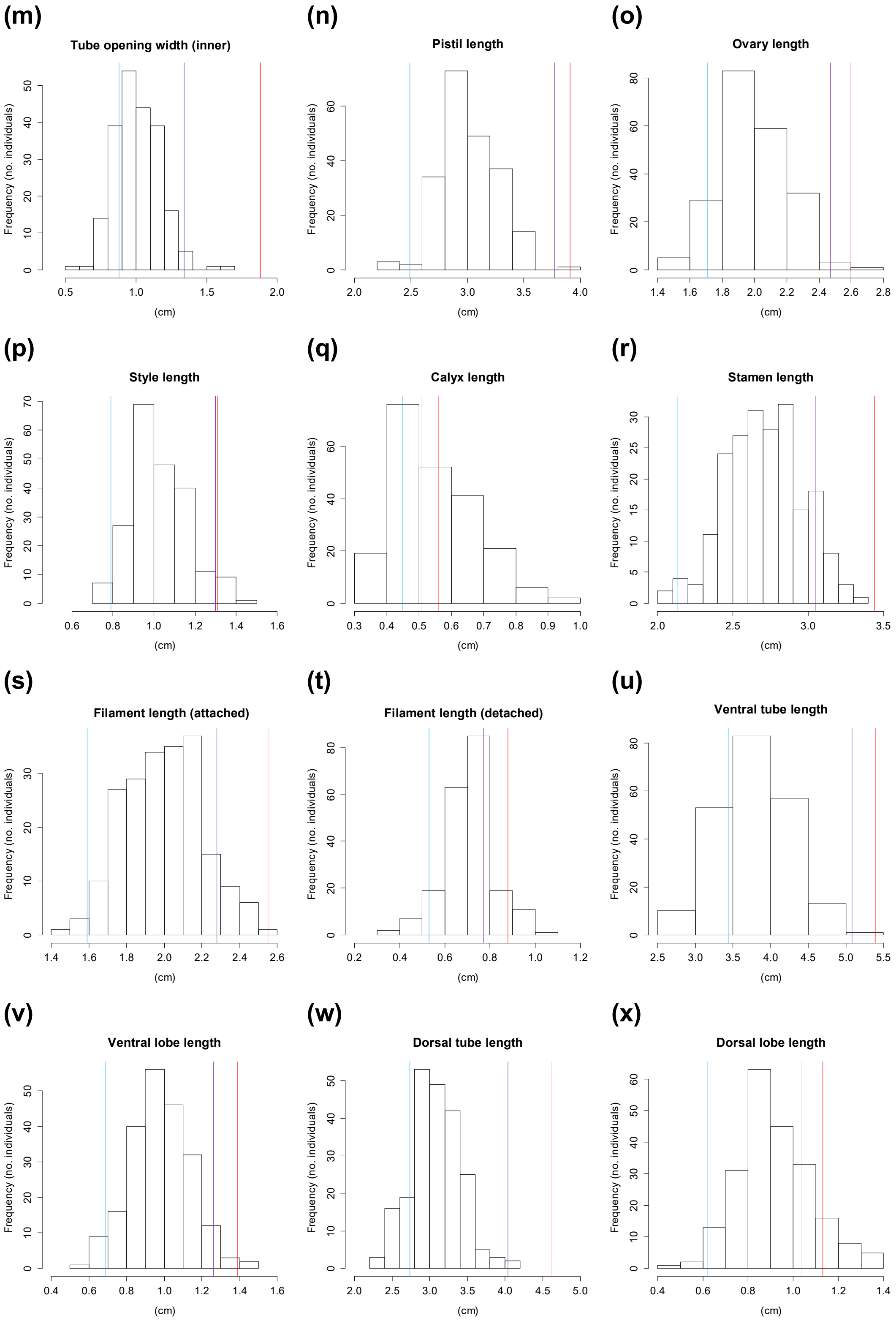
Char. No.	Trait	*P*-value	Type of distribution
1	Corolla length (cm)	0.64	Normal distribution
2	Undilated tube length (cm)	0.16	Normal distribution
3	Dilated tube length (cm)	0.89	Normal distribution
4	Undilated tube height (cm)	0.02	Nonparametric distribution
5	Dilated tube height (cm)	0.05	Normal distribution
6	Undilated tube width (cm)	0.04	Nonparametric distribution
7	Dilated tube width (cm)	0.01	Nonparametric distribution
8	Corolla face height (cm)	0.11	Normal distribution
9	Tube opening height (outer) (cm)	0.01	Nonparametric distribution
10	Tube opening height (inner) (cm)	< 0.01	Nonparametric distribution
11	Corolla face width (cm)	0.25	Normal distribution
12	Tube opening width (outer) (cm)	0.03	Nonparametric distribution
13	Tube opening width (inner) (cm)	< 0.01	Nonparametric distribution
14	Pistil length (cm)	0.26	Normal distribution
15	Ovary length (cm)	0.39	Normal distribution
16	Style length (cm)	0.01	Nonparametric distribution
17	Calyx length (cm)	< 0.01	Nonparametric distribution
18	Stamen length (cm)	0.78	Normal distribution
19	Filament length (attached part) (cm)	0.71	Normal distribution
20	Filament length (detached part) (cm)	< 0.01	Nonparametric distribution
21	Ventral tube length (cm)	0.80	Normal distribution
22	Ventral lobe length (cm)	0.92	Normal distribution
23	Dorsal tube length (cm)	0.28	Normal distribution
24	Dorsal lobe length (cm)	0.06	Normal distribution
25	Time to flowering[a] (DAS)	< 0.01	Nonparametric distribution
31	Time to 1st leaf initiation[b] (DAS)	< 0.01	Nonparametric distribution

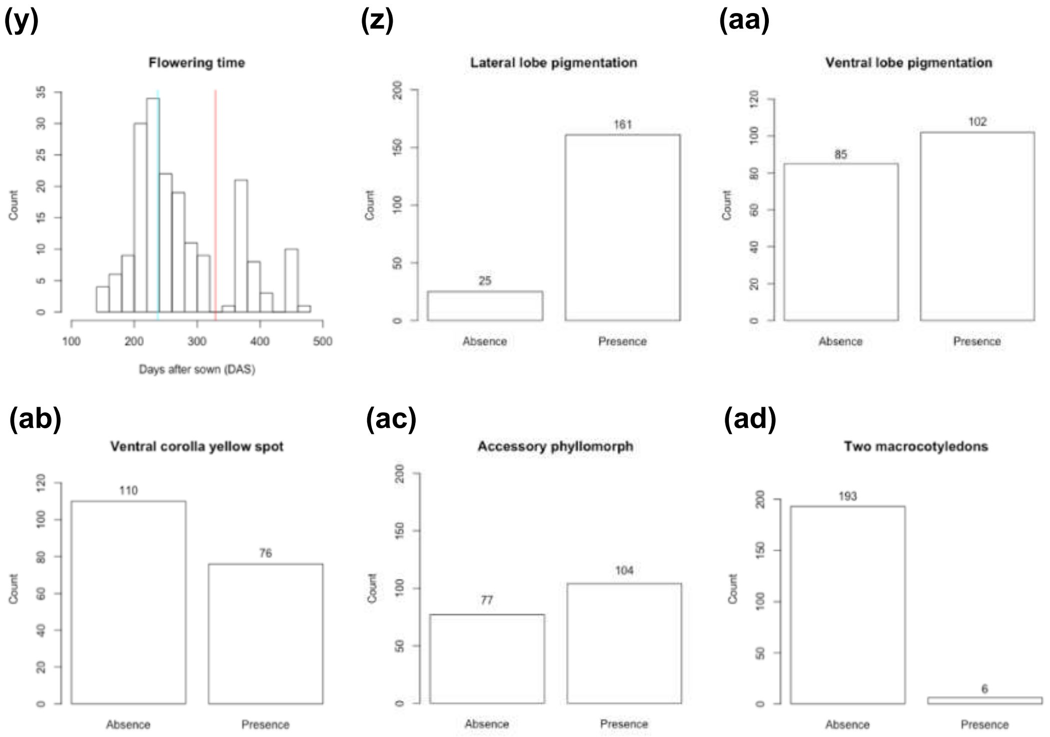
a Days to flowering (DAS, days after sowing)

b Days to first phyllomorph initiation (DAS, days after sowing)

Appendix 6.7


(a)
Corolla length
Frequency (no. individuals)
0
10
20
30
40
50
60
3
4
5
6
7
(cm)
(b)
Undialated tube length
Frequency (no. individuals)
0
10
20
30
40
0.5
1.0
1.5
2.0
2.5
3.0
(cm)
(c)
Dialated tube length
Frequency (no. individuals)
0
10
20
30
40
50
1.0
1.5
2.0
2.5
3.0
3.5
(cm)
(d)
Undialated tube height
Frequency (no. individuals)
0
10
20
30
40
50
0.4
0.5
0.6
0.7
0.8
(cm)
(e)
Dialated tube height
Frequency (no. individuals)
0
10
20
30
40
0.5
0.6
0.7
0.8
0.9
1.0
1.1
1.2
(cm)
(f)
Undialated tube width
Frequency (no. individuals)
0
10
20
30
40
50
0.4
0.5
0.6
0.7
0.8
0.9
(cm)
(g)
Dialated tube width
Frequency (no. individuals)
0
20
40
60
80
0.6
0.8
1.0
1.2
1.4
(cm)
(h)
Corolla face height
Frequency (no. individuals)
0
20
40
60
80
100
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
(cm)
(i)
Tube opening height (outer)
Frequency (no. individuals)
0
20
40
60
80
0.5
1.0
1.5
2.0
(cm)
(j)
Tube opening height (inner)
Frequency (no. individuals)
0
10
20
30
40
50
60
0.4
0.6
0.8
1.0
1.2
1.4
1.6
1.8
(cm)
(k)
Corolla face width
Frequency (no. individuals)
0
20
40
60
80
1
2
3
4
5
(cm)
(l)
Tube opening width (outer)
Frequency (no. individuals)
0
20
40
60
80
0.5
1.0
1.5
2.0
2.5
3.0
(cm)

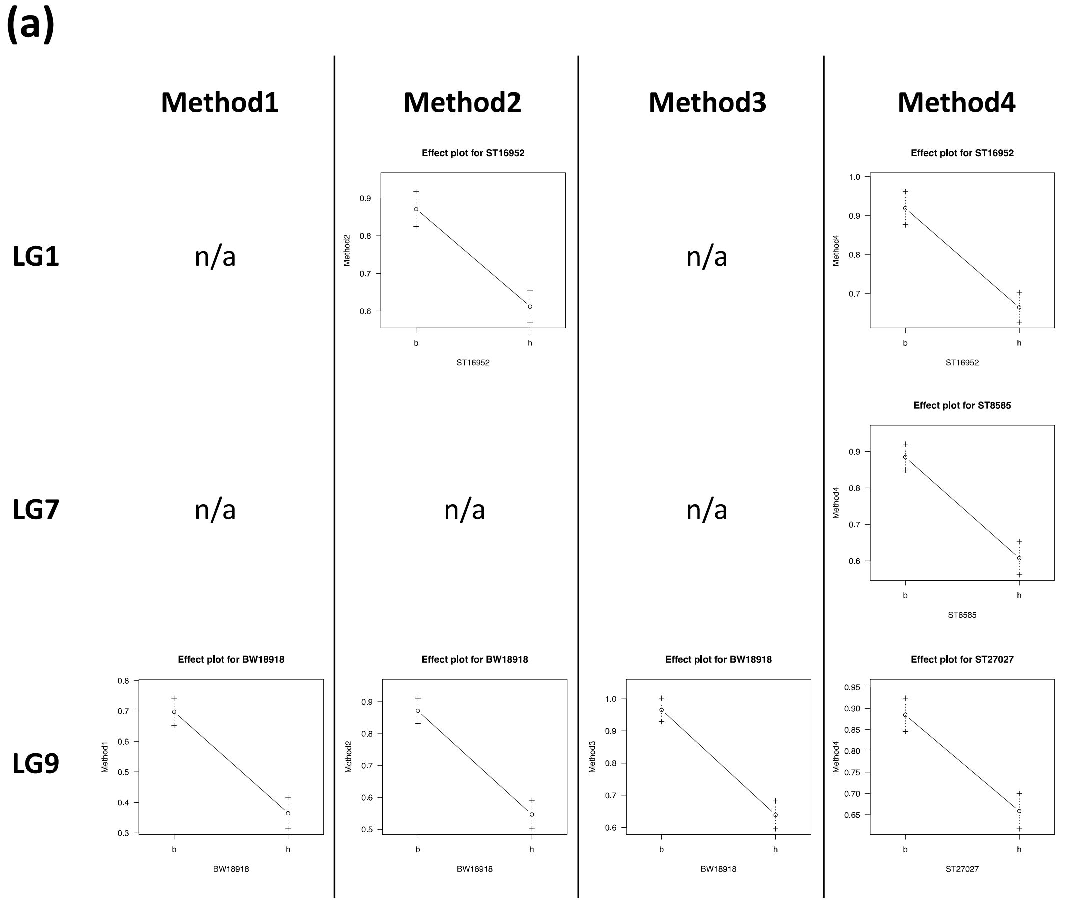

(m)
Tube opening width (inner)
Frequency (no. individuals)
0 10 20 30 40 50
0.5 1.0 1.5 2.0
(cm)
(n)
Pistil length
Frequency (no. individuals)
0 20 40 60
2.0 2.5 3.0 3.5 4.0
(cm)
(o)
Ovary length
Frequency (no. individuals)
0 20 40 60 80
1.4 1.6 1.8 2.0 2.2 2.4 2.6 2.8
(cm)
(p)
Style length
Frequency (no. individuals)
0 10 20 30 40 50 60 70
0.6 0.8 1.0 1.2 1.4 1.6
(cm)
(q)
Calyx length
Frequency (no. individuals)
0 20 40 60
0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0
(cm)
(r)
Stamen length
Frequency (no. individuals)
0 5 10 15 20 25 30
2.0 2.5 3.0 3.5
(cm)
(s)
Filament length (attached)
Frequency (no. individuals)
0 10 20 30
1.4 1.6 1.8 2.0 2.2 2.4 2.6
(cm)
(t)
Filament length (detached)
Frequency (no. individuals)
0 20 40 60 80
0.2 0.4 0.6 0.8 1.0 1.2
(cm)
(u)
Ventral tube length
Frequency (no. individuals)
0 20 40 60 80
2.5 3.0 3.5 4.0 4.5 5.0 5.5
(cm)
(v)
Ventral lobe length
Frequency (no. individuals)
0 10 20 30 40 50
0.4 0.6 0.8 1.0 1.2 1.4 1.6
(cm)
(w)
Dorsal tube length
Frequency (no. individuals)
0 10 20 30 40 50
2.0 2.5 3.0 3.5 4.0 4.5 5.0
(cm)
(x)
Dorsal lobe length
Frequency (no. individuals)
0 10 20 30 40 50 60
0.4 0.6 0.8 1.0 1.2 1.4
(cm)



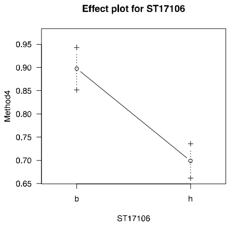
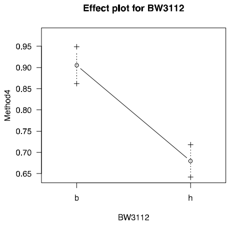
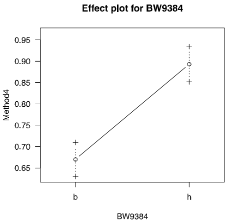
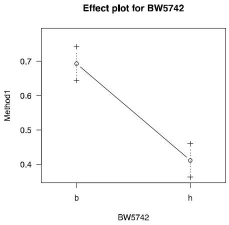
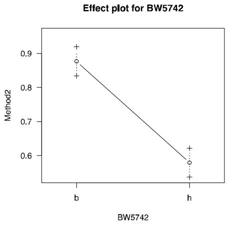
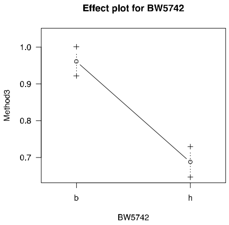
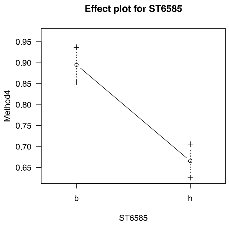
Appendix Figure 6.7 Phenotypic distribution of the traits measured in the BC population (*N* = 200). **(a)** Corolla length. **(b)** Undilated tube length. **(c)** Dilated tube length. **(d)** Undilated tube height. **(e)** Dilated tube height. **(f)** Undilated tube width. **(g)** Dilated tube width. **(h)** Corolla face height. **(i)** Tuber opening height, outer. **(j)** Tube opening height, inner. **(k)** Corolla face width. **(l)** Tube opening width, outer. **(m)** Tube opening width, inner. **(n)** Pistil length. **(o)** Ovary length. **(p)** Style length. **(q)** Calyx length. **(r)** Stamen length. **(s)** Filament length, attached part. **(t)** Filament length, free part. **(u)** Ventral tube length. **(v)** Ventral lobe length. **(w)** Dorsal tube length. **(x)** Dorsal lobe length. **(y)** Flowering time. **(z)** Pigmentation on lateral lobe. **(aa)** Pigmentation on ventral lobe. **(ab)** Yellow spot on ventral corolla tube. **(ac)** Accessory phyllomorph. **(ad)** Two macrocotyledons. Blue vertical lines: average trait value of *S. rexii*. Red vertical lines: average trait value of *S. grandis*. Purple vertical lines: average trait value of F1 hybrid.

Appendix 6.8

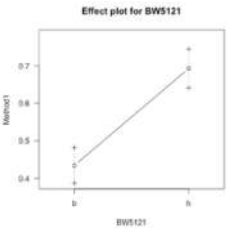
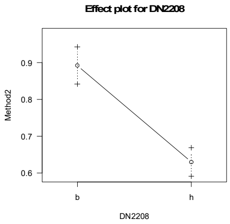
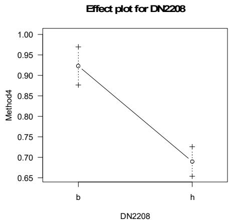
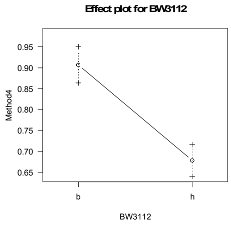
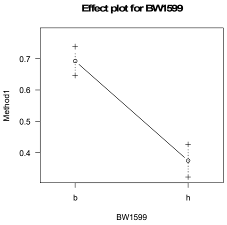
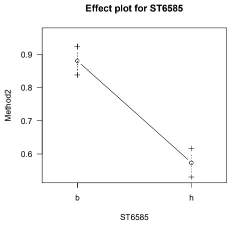
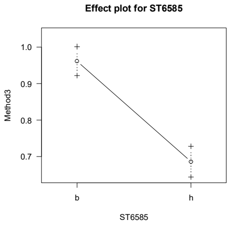
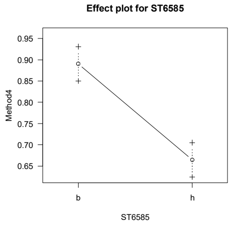
Effect plots of the BTL detected in the mapping rosulate / unifoliate trait. (a) Loci detected in MapA. (b) Loci detected in MapB-1. (c) Loci detected in MapB-3.



(b)

	Method1	Method2	Method3	Method4
LG2	n/a	n/a	n/a	
LG4	n/a	n/a	n/a	
LG10	n/a	n/a	n/a	
LG14				

(c)

	Method1	Method2	Method3	Method4
LG1		n/a	n/a	n/a
LG2	n/a		n/a	
LG4	n/a	n/a	n/a	
LG14				

Appendix 6.9

Scanning electron microscopy (SEM) of the parental plants. Used to produce the images in Figure 1.6 and 1.7

Seedlings of *S. rexii*, *S. grandis*^{BC} and F1 hybrids were collected at 5 DAU (days after cotyledon unfolding), 20 DAU, 30 DAU, 40 DAU, 50 DAU, 60 DAU, 65 DAU, 70 DAU, 80 DAU, and 90 DAU. In addition, for *S. grandis*^{BC} the stages 140 DAU and 150 DAU were also collected. The 5 DAU material represented the isocotylous stage; the 20 DAU and 30 DAU material represented the onset of anisocotylous development; the 60 DAU and 65 DAU materials represented the phyllomorph initiation stage in *S. rexii* and the F1 hybrid plant. Finally, the 140 DAU and 150 DAU samples of *S. grandis*^{BC} represented the initiation of the inflorescence meristem. The seedling materials of *S. grandis*^{F1} were not available at the time of this study. The collected samples were fixed in FAA (50% ethanol, 5% acetic acid, 3.7% formaldehyde). The samples submerged in FAA and in infiltrated in a vacuum chamber overnight, and later transferred to 70% ethanol for long term storage.

The samples stored in 70% ethanol were first dehydrated through the liquid substitution process in an ethanol and acetone series (Table 6.2). The samples were then transferred to a K850 critical point drier (Quorum, Lewes, United Kingdom) followed by liquid CO^2 exchange, to replace the acetone with liquid CO^2. The fluid exchange was repeated 10 times, each lasting for 1 minute. The heating system of the K850 machine was then turned on and the temperature inside the CPD raised until the critical point of CO^2 was reached, i.e. at 31°C and 1,071 psi. The chamber was then depressurised at a rate of ~1000 cm^3 per minute, which took about 20 minutes for complete depressurisation.

Table Ethanol and acetone dehydration series of samples prior to critical point drying

Solution	Incubation time
70% Ethanol	Long term storage
95% Ethanol	1 hr
100% Ethanol	1 hr
100% Ethanol	1 hr
100% Acetone (in molecular sieve)	5 min
100% Acetone (in molecular sieve)	5 min

The critical point dried samples were transferred to SEM stubs covered with carbon conductive tape. The stubs were then sputter coated with platinum particles using the K575X sputter coater (Quorum, Lewes, UK). The following settings for the sputter coater were used: sputter current 25 mA, sputter time 2 min, pump hold disabled. Finally, the sputter coated samples were observed using a Carl Zeiss SUPRA-55VP SEM machine (Carl Zeiss AG, Oberkochen, Germany). Photos were taken at 5 kV and a working distance of approximately 10 – 12 mm via the SmartSEM software (Carl Zeiss AG) integrated in the SEM machine.