

Systematic and Genomic Studies in the Plant Genus *Zygophyllum*

by
Pieter de Wet van der Merwe

*Thesis presented in fulfilment of the requirements for the degree of
Master of Science (Biochemistry) in the Faculty of Natural Science at
Stellenbosch University*



Supervisor: Prof. Dirk Uwe Bellstedt
Co-supervisor: Dr Michael David Pirie

March 2015

Declaration

By submitting this thesis/dissertation electronically, I declare that the entirety of the work contained therein is my own, original work, that I am the sole author thereof (save to the extent explicitly otherwise stated), that reproduction and publication thereof by Stellenbosch University will not infringe any third party rights and that I have not previously in its entirety or in part submitted it for obtaining any qualification.

Pieter de Wet van der Merwe

Date: March 2015

Abstract

Zygophyllum orbiculatum Welwitsch ex Oliv. from Angola and *Zygophyllum stapffii* Schinz from Namibia were described in the late 1800's. Recent comparisons of these two species revealed that they were morphologically very similar except that *Zygophyllum orbiculatum* has unifoliolate leaves and *Zygophyllum stapffii* has bifoliolate leaves. The similarity of these two species was investigated using nuclear ITS (Internal Transcribed Spacer, i.e. ITS1, 5.8SrDNA, ITS2) region sequence data as phylogenetic markers. Due to almost identical sequences and phylogenetic grouping, it was concluded that they were conspecific. However, the phylogenetic relationships of the major groups within the subfamily Zygophylloideae based on ITS sequences, were unresolved and unsupported, as was found in previous studies using chloroplast gene marker sequences.

To resolve the phylogenetic relationships of the major groups within the subfamily Zygophylloideae, a next generation sequencing (NGS) approach was taken. Chloroplasts of taxa representing the major groups within the subfamily were isolated and chloroplast genome sequence data were generated using the Ion Torrent™ sequencer. Additional nuclear ITS cassette data (18SrDNA, ITS1, 5.8SrDNA, ITS2, 26SrDNA) were generated as a by-product and used to produce a large combined aligned sequence matrix for phylogenetic analysis.

Model-based phylogenetic programs were able to retrieve strongly supported and resolved phylogenetic relationships of the major groups within Zygophylloideae. Two basal groupings were retrieved in the subfamily. The first grouping consisted of the genera *Tetraena*, *Fagonia* and *Melocarpum*. The second grouping consisted of the monotypic genus *Augea* and *Zygophyllum orbiculatum/stapffii* which were embedded within the genus *Roepera*. Using a gene duplication approach, the chloroplast marker data of genus *Zygophyllum sensu stricto* placed this genus basal to the *Augea*, *Zygophyllum orbiculatum/stapffii*, *Roepera* clade whilst the nuclear marker data of *Zygophyllum sensu stricto*, was found in a basal position to the entire subfamily. From this it is concluded that *Zygophyllum sensu stricto* shows evidence of incomplete lineage sorting. A revised taxonomy for the entire subfamily Zygophylloideae is proposed.

Abstrak

Zygophyllum orbiculatum Welwitsch ex Oliv. uit Angola en *Zygophyllum stapffii* Schinz van Namibië is in die laat 1800's beskryf. Onlangse vergelykings van hierdie twee spesies het getoon dat hulle morfologies baie anders is, behalwe dat *Zygophyllum orbiculatum* unifoliolate blare besit en dat *Zygophyllum stapffii* bifoliolate blare besit. Hierdie ooreenkoms is ondersoek, met behulp van die nukleêre “ITS” (Internal Transcribed Spacer d.w.s. ITS1, 5.8SrDNA, ITS2) DNS-strook volgordedata as filogenetiese merkers. As gevolg van feitlik identiese geenopeenvolgings is bevind dat die twee spesies konspesifiek is. Die filogenetiese verwantskappe van die groot binnegroepe van die subfamilie Zygophylloideae, gebaseer op ITS geenopeenvolgings, was egter onopgelos en nie ondersteun nie, net soos in vorige studies waarin chloroplast geenmerkervolgordes gebruik was.

Om die filogenetiese verwantskappe van die groot binnegroepe van die subfamilie Zygophylloideae op te los, was ‘n betreklik nuwe DNS volgordebepalingstegniek, naamlik “Next Generation Sequencing” (NGS), gebruik. Chloroplaste van taksa, wat die groot groepe binne-in die subfamilie verteenwoordig, is geïsoleer en chloroplast genoomdata is gegenereer met behulp van die Ion Torrent TM (NGS) DNS-volgordebepaler. Bykomend was die nukleêre “ITS”-kasset volgordedata (18SrDNS, ITS1, 5.8SrDNS, ITS2, 26SrDNS) ook as 'n by-produk gegenereer en ook gebruik om 'n groot gesamentlike DNS oplyningmatriks vir filogenetiese doeleindes.

Model-gebaseerde filogenetiese programme was in staat was om sterk ondersteuning en opgeloste filogenetiese verwantskappe van die groot groepe binne-in Zygophylloideae te ontravel. Die subfamilier toon twee basale groeperinge. Die eerste groepering bestaan uit die genera *Tetraena*, *Fagonia* en *Melocarpum*. Die tweede groepering bestaan uit die monotipiese genus *Augea* en *Zygophyllum orbiculatum/stapffii*, wat ingebed is binne-in die genus *Roepera*. Deur ‘n geendupliseringsbenadering te gebruik op die DNS geenopeenvolgings van die verteenwoordigende takson van *Zygophyllum sensu stricto*, is bevind dat die chloroplast DNS volgordes hierdie groep basaal aan ‘n *Roepera/Augea/Zygophyllum orbiculatum/stapffii* klade plaas, terwyl die nukleêre DNS volgordes hierdie groep basaal aan die hele subfamilie Zygophylloideae plaas. Hieruit is die gevolgtrekking gemaak dat *Zygophyllum sensu stricto* bewyse van onvolledige afstammelingsortering toon. ‘n Gewysigde taksonomie vir die hele subfamilie Zygophylloideae word voorgestel.

Acknowledgements

Supervisor: Professor Dirk Uwe Bellstedt for guidance, support in all facets of the study, as well as for funding.

Co-supervisor: Dr Michael (Mike) David Pirie for suggestions and guidance with NGS analyses and interpretation of results.

Parents: Hendrik Naudé van der Merwe (1959-2010) and Malinda van der Merwe

Siblings: Bosman Botha van der Merwe and Hendrik Naudé van der Merwe

My grandparents, uncles, aunts, nephews, nieces, cousins and family friends.

Dr Chris Visser for initial help and guidance with the parsimony analyses.

Mrs Coral de Villiers, friends and colleagues in the Bellstedt-Botes laboratories.

The staff at the Central Analytical Facility (CAF) for all sequencing analyses, as well as read assemblies for the NGS studies.

National Research Foundation, for funding.

Abbreviations

§	section
3'	Three prime
5'	Five prime
AIC	Akaike Information Criterion
AICc	Akaike Information Criterion corrected
APG	Angiosperm Phylogeny Group
APS	Adenosine 5'-phosphosulfate
ATP	Adenosine triphosphate
BEAST	Bayesian Evolutionary Analysis Sampling Trees
BIC	Bayesian Information Criterion
BLAST	Basic Local Alignment Search Tool
bp	Base pairs
BS	Bootstrap
C ¹⁴	Carbon ¹⁴ isotope
C ₂	2-carbon (Phosphoglycolate)
C ₃	3-carbon (3-Phosphoglyceric acid)
C ₄	4-carbon (Malate/Aspartate)
CAF	Central Analytical Facility
CAM	Crassulacean Acid Metabolism
CI	Consistency Index
CIPRES	Cyberinfrastructure for Phylogenetic Research
CO ₂	Carbon dioxide
COM	Celestrales-Oxalidales-Malpighiales
contigs	Contiguous sequences
CsCl	Cesium chloride
dATPS	Alpha-thio triphosphate
DNA	Deoxyribonucleic acid
dNTPs	Deoxyribonucleotide triphosphates
emPCR	Emulsion polymerase chain reaction
ETS	External transcribed spacer
Gb	Gigabases
IGS	Intergenic spacer
IR	Inverted repeat
ISFET	Ion-sensitive field-effect transistor
ITS	Internal transcribed spacer
ITS cassette	18S rRNA-ITS 1-5.8S rRNA-ITS 2-26S rRNA
ITS region	ITS 1-5.8S rRNA-ITS 2
LSC	Large single copy
MRCA	Most Recent Common Ancestor
NCBI	The National Center for Biotechnology Information
NGS	Next-generation sequencing
NOR	Nucleolus Organizer Regions
PAUP	Phylogenetic Analysis Using Parsimony
PCR	Polymerase Chain Reaction
PGM	Personal Genome Machine
polony	Polymerase colony
PP	Posterior Probability
PP _i	Inorganic pyrophosphate
RAxML	Randomized Axelerated Maximum Likelihood

rDNA	Ribosomal DNA
RI	Retention Index
RNA	Ribonucleic acid
rRNA	Ribosomal ribonucleic acid
RT-bases	Reversible terminator bases
Rubisco	Ribulose-1,5-bisphosphate carboxylase/oxygenase
SANBI	South African National Biodiversity Institute
<i>sensu lato</i>	Latin: "in the wide" or "broad sense"
<i>sensu stricto</i>	Latin "in the strict sense"
SMRT	Single molecule real-time
SOLiD	Sequencing by Oligonucleotide Ligation and Detection
SSC	Small single copy
ssDNA	Single stranded deoxyribonucleic acid
sstDNA	Single stranded template deoxyribonucleic acid
TIS	Transcription initiation site
tRNA	Transfer RNA
UCT	University of Cape Town
USD	United States Dollars
ZMW	Zero-mode waveguide

1 Table of Contents

1	Introduction and literature review	11
1.1	Introduction.....	11
1.2	Plant systematics	12
1.2.1	Computational phylogenetics.....	12
1.2.2	Plant molecular systematics	16
1.2.3	Phylogenetic relationships of the Angiosperms based on a molecular systematics approach.....	26
1.3	The Next-Generation Sequencing platforms.....	27
1.3.1	Pyrosequencing (454 sequencing)	28
1.3.2	Polymerase-based sequence-by-synthesis (Illumina®)	32
1.3.3	Ligation-based sequencing (SOLiD)	33
1.3.4	Ion semiconductor sequencing (Ion Torrent™).....	37
1.3.5	Single molecule Real-Time Sequencing (SMRT Sequencing).....	39
1.4	The chloroplast genome	41
1.4.1	Structure of the chloroplast genome	41
1.5	Phylogenomics - The use of NGS data for phylogenetic inference.....	42
1.6	Systematics of <i>Zygophyllum</i>	44
1.6.1	Systematic classification of <i>Zygophyllum</i>	44
1.6.2	Molecular systematics of <i>Zygophyllum</i>	49
1.7	Objectives and thesis structure.....	56
2	The <i>Zygophyllum orbiculatum</i> and <i>Zygophyllum stapffii</i> conundrum and the phylogenetic relationships of <i>Zygophyllum</i> based on the nuclear ITS region.	57
2.1	Introduction.....	57
2.2	Materials and Methods.....	60
2.2.1	Taxon sampling.....	60
2.2.2	DNA extraction	60
2.2.3	Amplification of the ITS gene region	60

2.2.4	Parsimony analysis.....	62
2.2.5	Maximum likelihood.....	63
2.2.6	Bayesian inference	63
2.3	Results.....	64
2.3.1	Molecular data.....	64
2.4	Discussion	68
3	Next-Generation Sequencing and Combined Trees.....	71
3.1	Introduction.....	71
3.2	Materials and Methods.....	72
3.2.1	Species that were investigated using the Next-Generation Sequencing approach.....	72
3.2.2	The chloroplast genome isolation and DNA purification procedures.....	74
3.2.3	Contiguous sequence assembly and bioinformatic strategies	74
3.2.4	Initial genetic information processing.....	75
3.2.5	Aligning contigs to the chloroplast reference genome in CodonCode Aligner.	76
3.2.6	Genetic markers used after the contig alignment to the reference genomes.....	78
3.2.7	Phylogenetic inference methods used	79
3.2.8	Tree editing	82
3.3	Results.....	83
3.3.1	Phylogenetic analyses of the genes involved in photosynthesis	83
3.3.2	Phylogenetic analyses of the genes not involved in photosynthesis.....	84
3.3.3	Phylogenetic analyses of the genes of the nuclear ITS cassette	86
3.3.4	Phylogenetic analyses of of combined chloroplast genes.....	87
3.3.5	Phylogenetic analyses of the combined chloroplast genes and the nuclear ITS cassette	89
3.3.6	Conflicting nuclear and chloroplast signal in Asian <i>Zygophyllum</i>	92
3.4	Discussion	92
4	Conclusions and Future Perspectives.....	96
5	References.....	98

6	Appendices.....	113
---	-----------------	-----

1 Introduction and literature review

1.1 Introduction

The first *Zygophyllum* species was identified by the Swedish naturalist Linnaeus in 1753 who named it *Zygophyllum fabago* (102). The family this plant would eventually be classified in was named Zygophyllaceae by Robert Brown in 1814 (24). Over the years several studies, using morphological characteristics, failed to resolve the phylogenetic relationships of the species that were subsequently described in this group. These studies had disputed placements of several taxa, most importantly for this study, of *Zygophyllum orbiculatum* and *Zygophyllum stapffii*. A phylogenetic study on the family, using morphological, anatomical and chloroplast *rbcL* gene sequence data was published in 1996, by Sheahan and Chase. They divided family Zygophyllaceae into five subfamilies, namely Morkillioideae, Tribuloideae, Seetzenioideae, Larreoideae and Zygophylloideae (179). Van Zyl in her PhD study analysed most of the *Zygophyllum* species found in southern Africa based on morphology (207). She, as several authors before her, had disputed placements of the species *Zygophyllum orbiculatum* and *Zygophyllum stapffii* within the genus *Zygophyllum*. *Zygophyllum orbiculatum* is found Angola while *Zygophyllum stapffii* is found in Namibia. These two species appear very similar except that *Zygophyllum stapffii* has bifoliolate leaves and *Zygophyllum orbiculatum* has unifoliolate leaves. A study by Beier *et al.*, in 2003, using morphological and chloroplast molecular *trnL* intron data divided the subfamily Zygophylloideae into the genera *Fagonia*, *Melocarpum*, *Augea* (monotypic), *Roepera*, *Tetraena* and *Zygophyllum* (13). The support of the some of the nodes separating the proposed genera was low. Bellstedt *et al.*, in 2008, disputed some of the taxonomic changes made by Beier *et al.* as some key taxa, e.g. *Zygophyllum orbiculatum* and *Zygophyllum stapffii* were not included in their investigation (17). The first comparison of *Zygophyllum orbiculatum* and *Zygophyllum stapffii* was also published in the study by Bellstedt *et al.* (2008). The *trnLF* and *rbcL* data suggested that these two species may be conspecific (17). However, all phylogenetic analyses on the subfamily Zygophylloideae have failed to resolve the relationships between the major groups and genera within the subfamily (13, 17, 179, 180).

This literature review begins with a brief description of plant systematics based on molecular characters, as well as the programs used to determine phylogenetic relationships from these characters and presenting them in phylogenetic trees. Since this study utilized a fairly new DNA sequencing science, which is named Next Generation Sequencing (NGS), a section is allocated to five of the different NGS platforms describing their sequencing principles. These are 454 sequencing, Illumina, SOLiD, Ion Torrent and SMRT sequencing. Following this section, a brief

discussion of the chloroplast genome as well as a section discussing the use of near complete chloroplast genomes for phylogenetic inference (phylogenomics) is discussed. The literature review is concluded with history of the plant group being studied, i.e. subfamily Zygothylloideae, as well as the problems of the classification within the group.

1.2 Plant systematics

As described by Michener *et al.* (1970) - “*Systematic biology (hereafter called simply systematics) is the field that (a) provides scientific names for organisms, (b) describes them, (c) preserves collections of them, (d) provides classifications for the organisms, Keys for their identification, and data on their distributions, (e) investigates their evolutionary histories, and (f) considers their environmental adaptations. This is a field with a long history that in recent years has experienced a notable renaissance, principally with respect to theoretical content. Part of the theoretical material has to do with evolutionary areas (topics e and f above), the rest relates especially to the problem of classification. Taxonomy is that part of Systematics concerned with topics (a) to (d) above.*”(120). This statement can be condensed as follows: Taxonomy is the science of defining groups of biological organisms based on shared characteristics and giving names to those groups. Systematics is the scientific study of the diversification of living organisms, both past and present, and the relationships among these living organisms through time. Since Michener *et al.* (1970) wrote this definition, the investigation into evolutionary histories, part (e), in particular the use of DNA sequence data, has allowed major advances.

1.2.1 Computational phylogenetics

Computational phylogenetics is defined as the use of computational algorithms, programs and methods in order to reconstruct phylogenetic histories. These phylogenetic histories are represented in a phylogenetic tree. The data that is used in a phylogenetic reconstruction can be based on morphological character states or it can contain molecular data such as DNA, RNA or amino acid sequences of proteins. The data is displayed in a data matrix in which the lines represent the species or taxa being investigated. The columns represent the data under investigation.

1.2.1.1 Phylogenetic trees

A phylogenetic tree or phylogeny is a representation of relationships between species which are inferred from their shared, as well as their unique characteristics. It is a representation of the evolutionary changes. “Branching points”, also known as nodes, within the tree represent a

common ancestor, while the “trunk” or “root” is representative of the common ancestor of the whole group under investigation. The individual species that are investigated are found on the terminals of the phylogenetic tree and are referred to as operational taxonomic units (OTU’s). Taxa that are the closest related to one another are known as sister taxa, and represent a monophyletic group, as opposed to para- or polyphyletic groups. The choice of characters used to construct a tree affects the shape or topology. Homologous traits are favoured, while convergent traits are not. In classical systematics, the characters used in the phylogenetic reconstructions are the morphological character states whilst in molecular systematics the characters used in phylogenetic reconstructions are base changes in DNA or RNA sequences or in the case of proteins, amino acid changes.

Producing phylogenetic trees from datasets requires homology within the data being analysed. In studies based on morphological characters it requires a choice by the botanist of which characters to use and how to measure and encode the different states of each character under investigation. In studies based on molecular characters, the main problem is to produce a multiple sequence alignment, i.e. for each taxon being investigated there needs to be a molecular character string to compare to each other molecular character string within the multiple sequence alignment. Each character string needs to be in the right orientation and each base pair or amino acid needs to be aligned to the corresponding base pair or amino acid of all other taxa in the alignment matrix.

There are several aspects that need to be considered when constructing a phylogenetic tree. These are that all species share a common ancestor and that no two species are identical. The more homologous characters species share, the more closely they are related (temporally). Another aspect which needs to be considered is that of irreversibility. This was proposed by Louis Dollo in 1893 (Dollo’s Law). It states that species cannot wholly return to a previous state that was once achieved in its evolutionary history (41, 132). There are studies that indicate that there might be exceptions to this statement (33, 91, 109, 167, 212), but this is a topic of contention (35, 62, 64).

There are several aspects that relate to the topologies that can arise within phylogenetic trees that will be described in the following paragraph that are important to understand. These are mono-, para- and polyphyletic lineages as well as polytomies.

In Figure 1-1 is a diagram of a phylogenetic tree. Species A, B, C and D are a monophyletic group as they share a common ancestor at node 2. Species C, D and E are paraphyletic as they cannot form a monophyletic group without including other taxa. Species D and E are defined as being polyphyletic if they were previously grouped together as a taxonomic group, but they do

not derive from a single common ancestor. Species F, G and H are located on a polytomy as they originated from a single common ancestor, but the phylogenetic relationships between them are unresolved. Polytomies can be soft or hard. Soft polytomies occur when not enough information is available to resolve the phylogenetic relationships of the taxa being studied, or due to conflict in sequence data which can arise due to recombination or due to different phylogenetic signals from different genes (61, 182). Hard polytomies occur when due to very rapid speciation more than two lineages simultaneously diverged from a single ancestor (85, 182, 194, 205).

The phylogenetic trees generated by computational phylogenetics can be unrooted or rooted depending on the algorithm and input datasets. A rooted tree is a phylogenetic tree in which the most recent common ancestor is explicitly defined. Genetic distances are then measured in relation to the defined “most recent common ancestor” and the taxa are plotted on the tree proportional to their genetic distance from the root taxon. A rooted tree necessitates that the data matrix also contains the information of the root taxon. An unrooted tree plots the genetic distances and relationships between the investigated taxa without making assumptions regarding their descent.

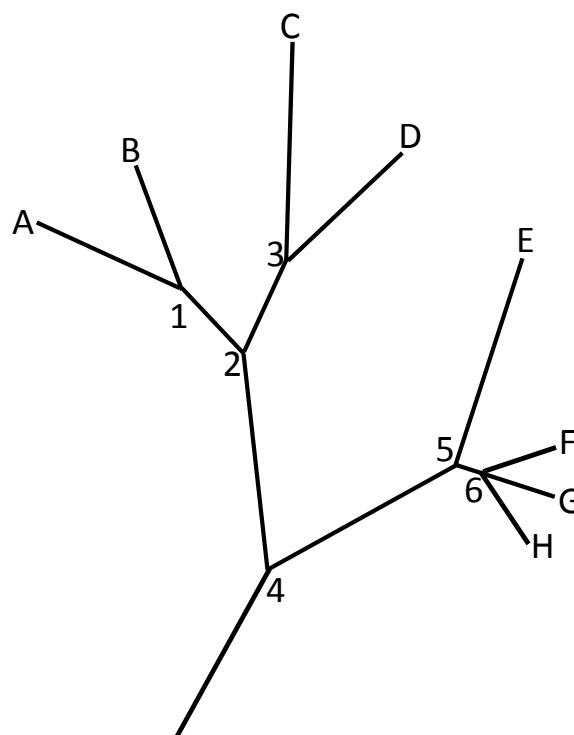


Figure 1-1: A diagram indicating the possible relationship situations that can arise in a phylogenetic tree.

1.2.1.2 Phylogenetic analyses

There are four main types of analyses that are used for phylogenetic inference, namely Distance Analysis, Parsimony, Bayesian Inference and Maximum Likelihood.

1.2.1.2.1 Distance analyses

In phylogenetic inference studies using distance analysis, pairwise distances are used to infer phylogenetic relationships from character state data. Distance methods attempt to map a tree to a data matrix of pairwise genetic distances (59). This means that for every two taxa, the corresponding distance is a single value based on the fraction of positions in which the two taxa differ, which is defined as the p -distance. This p -distance is an underestimation of the actual genetic distance as some character state positions are inclined to have had more than one substitution event. This means that in distance analyses the aim is to identify the number of substitution events that have occurred by applying a specific evolutionary model that makes certain assumptions about the nature of the changes in the matrix data. When all these pairwise distances have been calculated for a specific matrix dataset, a tree topology can be inferred by several methods. A more detailed description, however, falls beyond the scope of this thesis (96).

1.2.1.2.2 Parsimony analyses

Although distance analyses were initially the predominant method to reconstruct phylogenetic histories, parsimony analyses have been preferentially used since the early seventies. The parsimony analysis, explained in its simplest form, is based on what is known as the “Occam’s Razor” approach, that the simplest, most direct route is probably the correct one. A textbook definition is as follows: Maximum parsimony can be described as a particular non-parametric statistical method for constructing phylogenies. Through this application, the preferred constructed phylogenetic tree is the tree that supposes the least amount of character-state changes to explain the observed data. A phylogenetic program that utilizes the parsimony principle is PAUP (Phylogenetic Analyses Using Parsimony) (199).

1.2.1.2.3 Bayesian inference

Bayesian inference utilizes a number of statistical algorithms for sampling from a probability distribution, which is based on constructing a Markov chain which has the property that it has the desired distribution at its equilibrium distribution. It is defined in “The Phylogenetic Handbook” as follows: “A *statistical technique for integrating a function by drawing samples at random (“Monte Carlo”) from the function, basing each sample on the previous one (“Markov chain”). This stochastic technique is useful when the function cannot be integrated directly, but can fail if the sample drawn is not big enough or does not explore all important regions of the function.*”, (96). Apart from parsimony, this type of analysis has been the most widely adopted form of analysis. This type of analysis differs from parsimony in that it is a model-based

program, meaning that it can test several models of evolution on character-states. Models for gene markers can be analysed in programs, e.g. jmodeltest or PartitionFinder. The phylogenetic program that uses Bayesian Inference is MrBayes (69, 158)

1.2.1.2.4 Maximum Likelihood

Maximum likelihood is a method for calculating the parameters of a statistical model. It is defined in *“The Phylogenetic Handbook”* as follows: *“A principle of statistical inference developed by R. A. Fisher in the 1920s. Essentially, it is a generalization of least-squares to non-normal data, and can be shown to lead to optimal estimators, at least for large sample size. Moreover, it is fully automatic once a model is specified, and allows computing confidence bands by means of the so-called Fisher information.”*, (96). This method is computationally very demanding and has only in recent been adopted in mainstream science as computers have become more advanced. Similarly to Bayesian inference this is also a model-based analysis program. Models for these analyses can also be analysed in jmodeltest or PartitionFinder. The program performs an exhaustive search on the data and only gives the most likely tree with bootstrap support on the branches. RAxML (Randomized Accelerated Maximum Likelihood) is the phylogenetic program used to perform maximum likelihood analyses (186).

1.2.1.3 Computational phylogenetics based on molecular characteristics

In molecular phylogenetics character encoding is very different than in morphological phylogenetics. The data that is generated is discretely defined and immediate, albeit nucleotides in RNA or DNA) or amino acids in proteins. Defining homology can be troublesome due to the inherent nature of multiple sequence alignments. For a given gapped alignment several rooted trees can be derived that vary in the interpretation of the gaps. The question is whether the gaps are mutations or ancestral traits, or if the gaps are insertions in the taxa that contain them or whether they have been deleted in the taxa in which they do not occur. In order to circumvent this problem gaps are commonly excluded during phylogenetic analyses. Phylogenetic inferences can subsequently be performed on these data matrices using the methods described before and a phylogenetic tree can subsequently be obtained (43, 147, 176).

1.2.2 Plant molecular systematics

Molecular phylogenetics is a division of phylogenetics that, by definition, analyses hereditary differences in molecules of different organisms, e.g. the amino acid sequences of proteins or the nucleic acid sequences of RNA, but mainly the nucleic acid sequences of DNA, to enable scientists to determine their evolutionary relationships. Molecular phylogenetics, however, is

just one aspect of molecular systematics, which is a much broader field that uses the generated molecular data in taxonomy, sometimes in combination with morphology, as well as in biogeography.

There are several key technological and biochemical advances that were made that led to the recent advancements in molecular systematics. The most important are listed below in Table 1-1.

Table 1-1: Technological and biochemical breakthroughs made in the 20th century that led to the development of the scientific field of molecular systematics.

Timeline	Discovery
Before, during and after WWII (1930's-1940's)	The invention of electronic mechanical computers, e.g. the Z1 designed by Konrad Zuse from 1935 to 1936 and built by him from 1936 to 1938, paved the way for the invention of the first modern computers in the 1950's.
1940's	The Nobel Prize in chemistry was awarded to Swedish biochemist Arne Tiselius, in 1948, for his work regarding electrophoresis and adsorption analysis (The father of electrophoresis).
1950's	Discovery of the double helix nature of DNA.
1960's	Elucidation of the triplet DNA codons that encode for specific aminoacids.
1970's	The development of Sanger sequencing.
1980's	Invention of the Polymerase Chain Reaction (PCR).

One of the largest benefits to using molecular characters as opposed to morphological characters is the sheer amount of data available (Many thousands of molecular characters as opposed to tens to hundreds of morphological characters). Given modern sequencers, sequence data can be generated rapidly and with relative ease. Inheritance patterns are more easily distinguishable in molecular studies than in morphological studies, especially in closely related organisms (122). Much of the genomes of higher organisms (eukaryotes) are non-coding and are not under selective pressure and therefore do not distort the true phylogenetic relationships. This means that even when there are independent identical point mutations (which could lead to homoplasy), these are outweighed by the number of dissimilar mutations. This would imply that divergence outweighs convergence. Lastly the number of mutations is directly proportional to divergence times since the rate of mutation can be calculated. This means even in the absence of fossil data divergence events can be dated (209).

There are also disadvantages to using molecular data for phylogenetic inference. Different character states are easily identifiable, but an alignment of identical sequences in different taxa can be problematic, e.g. introns, indels and intergenic spacers. A potential problem inherent to DNA and RNA sequences is that they only possess four character states due to the fact that there are only four possible bases in these nucleic acids. This is more likely to lead to homoplasy

which can distort true phylogenetic relationships and cause long-branch attraction in phylogenetic studies.

1.2.2.1 Genetic markers used in phylogenetic reconstruction

The majority of molecular phylogenetic studies in botany currently utilize gene regions of chloroplast genomes. Chloroplasts are small organelles within the cells of photosynthesizing organisms including plants. There can be up to hundreds of chloroplasts per cell depending on where in the plant they are found or at which developmental stage the tissue is. Younger tissue tends to have many more chloroplasts than older cells. Chloroplasts have their own small circular genomes typically around 120 000 to 160 000 bp in size. Chloroplasts only have one copy of the genome per organelle, making them haploid. This is important to note as this means that phylogenetic studies based on chloroplast information can only describe the evolution/lineage of one of the parents and not the other. In most of the studied Gymnosperms the inheritance of chloroplasts has been found to be paternal. Within the Angiosperms in most cases the chloroplasts are only inherited maternally, but there are some cases of biparental or even paternal inheritance of the chloroplasts. This means that in some cases, when biparental inheritance of chloroplasts is known or suspected in the plants under investigation, there might be instances of recombination between the different chloroplast lineages which must then be taken into account. Within the scope of this thesis it is of importance to note that there have been reports of biparental and paternal inheritance of chloroplasts within Zygophyllaceae in the subfamily Larreoideae (*Larrea tridentata* or Creosote Bush) (36, 67, 215).

There are many advantages to using chloroplast markers for phylogenetic inference. Chloroplast genome concentrations are tens to hundreds of times that of the nuclear genome. This means that very little DNA of a plant sample is needed to generate genetic information, which might be beneficial if a limited supply is available.

Besides the inheritance problems outlined above there are other drawbacks when using chloroplast genetic information such as the photosynthetic pathways of the plants in question as some of the genes for the photosynthetic pathway enzymes are found in the chloroplast genome. Since there are several different known photosynthetic mechanisms, i.e. C₃, C₄ and CAM photosynthesis, genes of the different subunits of the enzymes involved in photosynthesis might be under a selective pressure which can distort phylogenetic inferences. It has been shown in the Poaceae that certain key aminoacids mutation in the the *rbcL* gene are associated with taxa that are known to have C₄ photosynthesis (34).

1.2.2.2 *The use of chloroplast gene region sequences for phylogenetic inference*

The *rbcL* gene encodes for the Ribulose-1,5-bisphosphate carboxylase/oxygenase large subunit of the Rubisco Ribulose-1,5-bisphosphate carboxylase/oxygenase enzyme. This enzyme is responsible for the incorporation of atmospheric CO₂ into organic carbohydrates within the photosynthetic process. This enzyme has a very low turnover and also is not specific only to CO₂, but can also bind and incorporate atmospheric O₂ in the process known as photorespiration.

RbcL is a gene that is approximately between 1 400 - 1 500 bp in length and encodes for an enzyme subunit that is approximately 500 amino acids in length. The gene is found on the Large Single Copy (LSC) region of the chloroplast genome. This gene is a powerful tool for elucidation of ancient divergent events and can be used in the analyses of all photosynthetic organisms, not just plants. Since the Rubisco enzyme is one the most important enzymes on the planet it is one the most extensively studied, either for structure/function/activity or, of importance of this thesis, in phylogenetic analyses starting in the 1980's and gaining momentum in the early 1990's (21, 22, 32, 63, 80, 119, 128, 190, 213). Typing the letters *rbcL* into The National Center for Biotechnology Information's nucleotide portal and focussing only on green plants retrieves 129 059 results (search performed on 8 October 2014). Another coding gene that has been commonly used for phylogenetic inference is the gene for the maturase K enzyme (*matK*) which is located within the intron of the *trnK* transfer RNA (82)(195).

Another well-studied region of the chloroplast genome in early studies of molecular systematics was the *trnL* intron within the *trnL* gene adjacent to the *trnF* gene, as well as the intergenic spacer between the two mentioned genes. Due to their high mutation rates these two regions are ideally suited for the phylogenetic study of closely related species (200). The *trnLF* marker was first used for phylogenetic analyses by Taberlet *et al.* in 1991 (200). They developed primers for use in PCR to amplify and sequence the marker of various gymnosperms, angiosperms, algae, bryophytes and pteridophytes (200). The authors specifically designed six primers, indicated in Figure 1-2, that show the locations where the primers bind. Two of the primers, "a" and "b", are used to amplify the adjacent region which includes the intergenic spacer between the *trnT* and *trnL* exons.

There are three copies of the *trnL* gene on the chloroplast. The gene for *trnL* in the *trnL-F* marker, named *trnL* (UAA) is unusual as it contains an intron, whereas the other two copies do not. The gene is roughly 85 bp long and the intron varies between approximately 500-1600 bp (186).

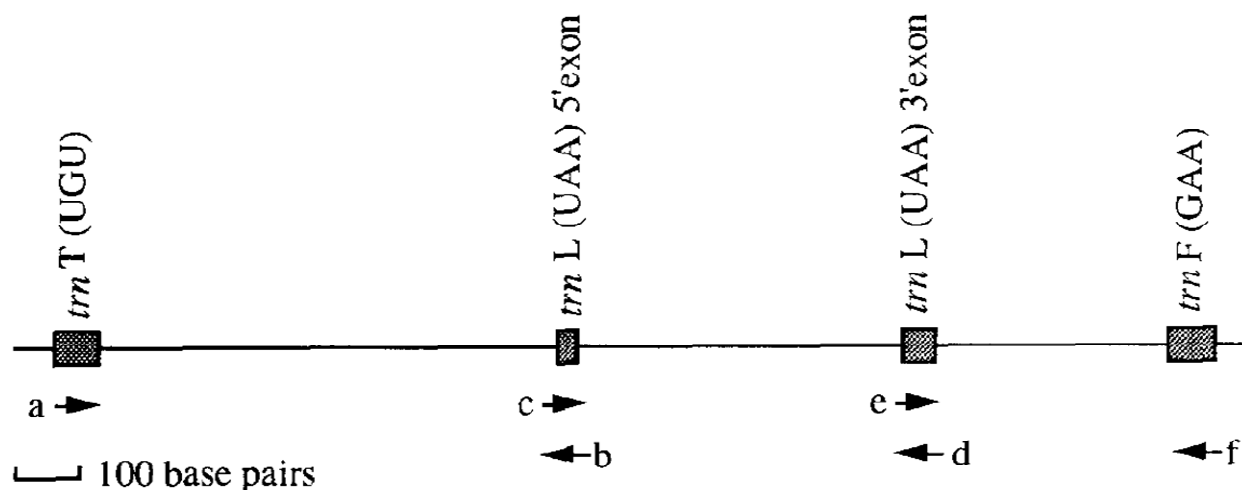


Figure 1-2: A diagram indicating the conserved regions of the *trnT*, *trnL* and *trnF* genes. The primers sequences are located on the genes of the specific tRNA molecules, which are highly conserved (200).

Since the initial use of these genetic markers, numerous other chloroplast gene sequences have been used for phylogenetic inference. A study initially published in 1998, identified several non-coding regions, as well as two nuclear encoded alcohol dehydrogenase enzymes for use in phylogenetic studies to resolve intraspecific relationships within five closely related cotton species (*Gossypium*) (186). The authors named the publication “The tortoise and the hare” as they attempted to find slowly and faster mutating regions within plant genomes that could be in the phylogenetic studies throughout the plant kingdom. The slower mutating regions could be used in studies to elucidate ancient speciation events, e.g. between plant orders, while the faster mutating regions could be used for the elucidation of the phylogeny of closely related species, e.g. within families or genera.

In 2005, the second publication in this series appeared and is referred to as “The Tortoise and the Hare II” (177). In this publication several areas (21 intergenic spacers or introns) found within the Large Single Copy regions of the chloroplast genome, were identified. The aim of the study was to identify regions with a high net mutation rate, which in turn meant more phylogenetically informative characters which could resolve phylogenies of closely related species. Most genetic markers up to this time were unable to resolve the phylogeny of closely related species since they are genes which encode for functional amino acid sequences of proteins and which therefore constrained due to selective pressure (e.g. *rbcL*) (177). In Table 1-2 is a list indicating the regions that were investigated in the 2005 study by Shaw *et al.* (177).

Table 1-2: The list of 21 chloroplast marker regions used by Shaw *et al.* in 2005 (177).

Count	Marker Name	Marker Type
1	<i>trn H-psb A</i>	Intergenic spacer
2	<i>psb A-3'trn K</i>	Intergenic spacer
3	<i>3'trn K-mat K</i>	Intergenic spacer
4	<i>mat K-5'trn K</i>	Intergenic spacer
5	<i>rpS 16</i>	Intron
6	<i>trn S-trn G</i>	Intergenic spacer
7	<i>trn G intron</i>	Intron
8	<i>rpo B-trn C</i>	Intergenic spacer
9	<i>trn C-ycf 6</i>	Intergenic spacer
10	<i>ycf 6-psb M</i>	Intergenic spacer
11	<i>psb M-trn D</i>	Intergenic spacer
12	<i>trn D-trn T</i>	Intergenic spacer
13	<i>trn S-trnf M</i>	Intergenic spacer
14	<i>trn S-rps 4</i>	Intergenic spacer
15	<i>rps 4-trn T</i>	Intergenic spacer
16	<i>trn T-trn L</i>	Intergenic spacer
17	<i>trn L</i>	Intron
18	<i>trn L-trn F</i>	Intron & Intergenic spacer
19	<i>5'rps 12-rpl 20</i>	Intergenic spacer
20	<i>psb B-psb H</i>	Intergenic spacer
21	<i>rpl 16</i>	Intron

The latest publication in the “Tortoise and the Hare” series appeared in 2007 (178), in which additional marker regions were identified that showed even more variation than the markers identified in the study of 2005 (177). New primers were also designed for the *trnS-trnG* region and these were also assessed in the study (see Table 1-3).

Table 1-3: The regions investigated in the Tortoise and Hare III (178).

Count	Marker Name	Marker Type
1	<i>rpl 14-rps 8-inf A-rpl 36</i>	Coding & intergenic spacer
2	<i>pet L-psb E</i>	Intergenic spacer
3	<i>psb J-pet A</i>	Intergenic spacer
4	<i>psa I-acc D</i>	Intergenic spacer
5	<i>3'trn V-ndh C</i>	Intergenic spacer
6	<i>ndh J-trn F</i>	Intergenic spacer
7	<i>psb D-trn T</i>	Intergenic spacer
8	<i>atp I-atp H</i>	Intergenic spacer
9	<i>trn Q-5'rps 16</i>	Intergenic spacer
10	<i>3'rps 16-5'trn K</i>	Intergenic spacer
11	<i>ndh A</i>	Intron
12	<i>ndh F-rpl 32</i>	Intergenic spacer
13	<i>rpl 32-trn L</i>	Intergenic spacer
14	<i>trn S-trn G-trn G</i>	Intergenic spacer

A publication based on a similar approach was published by Dong *et al.*, in 2012 (42). In this publication the entire chloroplast genomes of 12 genera were investigated for the highly variable regions that could be applied in phylogenetic studies to resolve resolution at low/terminal levels/branches. They identified 23 regions (see Table 1-4) that were highly variable and listed them from the most to least variable (42).

Table 1-4: The regions investigated in a 2012 publication (42).

Count	Marker	Marker Type
1	<i>acc D-psa I</i>	Intergenic spacer
2	<i>atp H-atp I</i>	Intergenic spacer
3	<i>clp P</i>	Intron
4	<i>ndh A</i>	Intron
5	<i>ndh C-trn V</i>	Intergenic spacer
6	<i>ndh F</i>	Intron
7	<i>pet A-psb J</i>	Intergenic spacer
8	<i>pet B-pet D</i>	Intergenic spacer
9	<i>pet N-psb M</i>	Intergenic spacer
10	<i>psb E-pet L</i>	Intergenic spacer
11	<i>psb M-trn D</i>	Intergenic spacer
12	<i>rbc L-acc D</i>	Intergenic spacer
13	<i>rpl 32-trn L</i>	Intergenic spacer
14	<i>rpo B-trn C</i>	Intergenic spacer
15	<i>rps 16-trn Q</i>	Intergenic spacer
16	<i>trn H-psb A</i>	Intergenic spacer
17	<i>trn K</i>	Intron
18	<i>trn S^{GCU}-trn G^{GCC}</i>	Intergenic spacer
19	<i>trn S^{UGA}-trn G^{GCC}</i>	Intergenic spacer
20	<i>trn T-psb D</i>	Intergenic spacer
21	<i>trn W-psa J</i>	Intergenic spacer
22	<i>ycf 1-a</i>	Intron
23	<i>ycf 1-b</i>	Intron

1.2.2.3 The use of nuclear gene region sequences for phylogenetic inference

Although phylogenetic studies utilizing chloroplasts markers are far more prevalent in botany, nuclear encoded genes are equally important as they are biparentally inherited and hence such genes are also used in plant phylogenetic inferences. The nuclear region most often used is the Internal Transcribed Spacer region or ITS regions separating the DNA that encodes for the rRNA molecules.

Many organisms encode their nuclear genetic information on several chromosomes. Higher eukaryotic organisms receive half of their nuclear genetic information from the paternal and half from the maternal line. This means in the case of a diploid organism it has two copies of each chromosome. Polyploids can contain many copies of each chromosome. In the context of this thesis it is of interest to note that most *Zygophyllum* species that have been karyotyped are found to be diploid in nature (1, 27, 75, 104, 114, 155–157), although in the case of the Creosote bush (*Larrea tridentata*), depending on the aridness of the environment it can be diploid, tetraploid or

in the most extreme environment, hexaploid (95). It is important when working with nuclear genetic information to reconstruct the phylogenetic relationships that recombination between the identical copies of the chromosomes (in diploid organisms) can occur. There are also slightly varied versions of the identical genes on the chromosomes named alleles, which might result in ambiguous base callings in sequence determinations.

The ITS region is found in a cassette of multiple copies in the somatic DNA of photosynthetic organisms like algae, plants and fungi and is a widely used region for phylogenetic investigations since it has a high mutation rate (Figure 1-3). The cassette includes three conserved coding but untranslated ribosomal genes and two transcribed, but exised, non-coding introns (Figure 1-4). The three coding genes are named 18S, 5.8S and 26S ribosomal RNA genes and the two non-coding spacers are named ITS1 and ITS2. This genetic area is flanked by what is known as an IGS region between the 18S and 26S rRNA genes (9)(100). The IGS can be further divided into a NTS non-transcribed spacer and external transcribed spacer (ETS) regions. Transcription is initiated at the TIS (transcription initiation site) and runs through to the terminal end of the 26S rRNA. The ETS, ITS1 and ITS2 are the spliced and the rRNA mature and fulfill their functions in ribosomal protein synthesis.

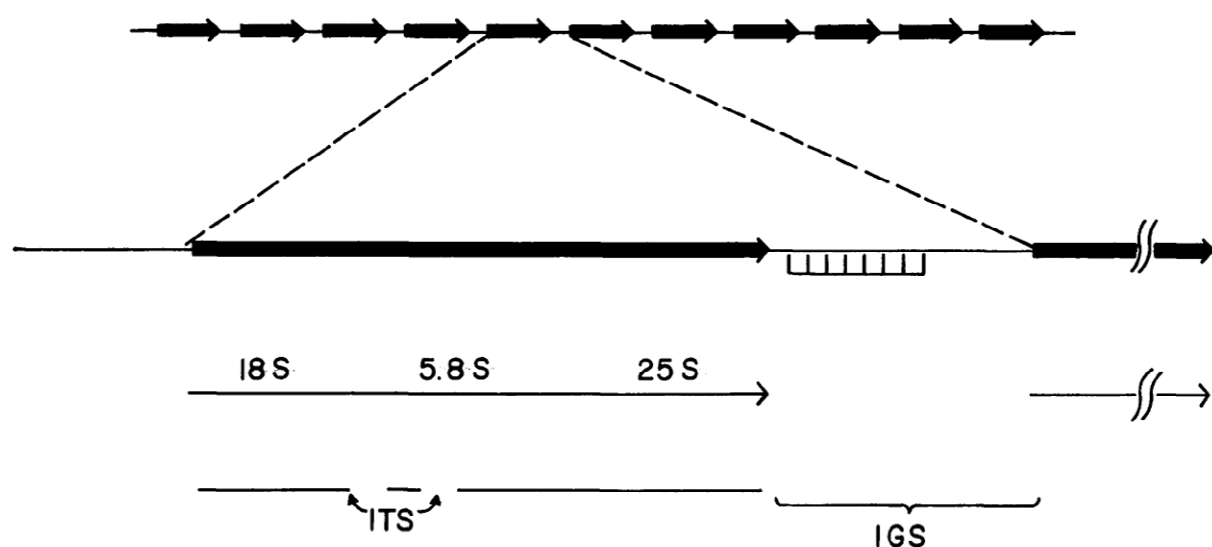


Figure 1-3: A diagram of the rDNA region containing the ITS region and indicating the IGS downstream. Upstream of the region is the ETS, or external transcribed spacer. Multiple copies, referred to as cassettes are found in repeat regions in the nuclear genome (83).

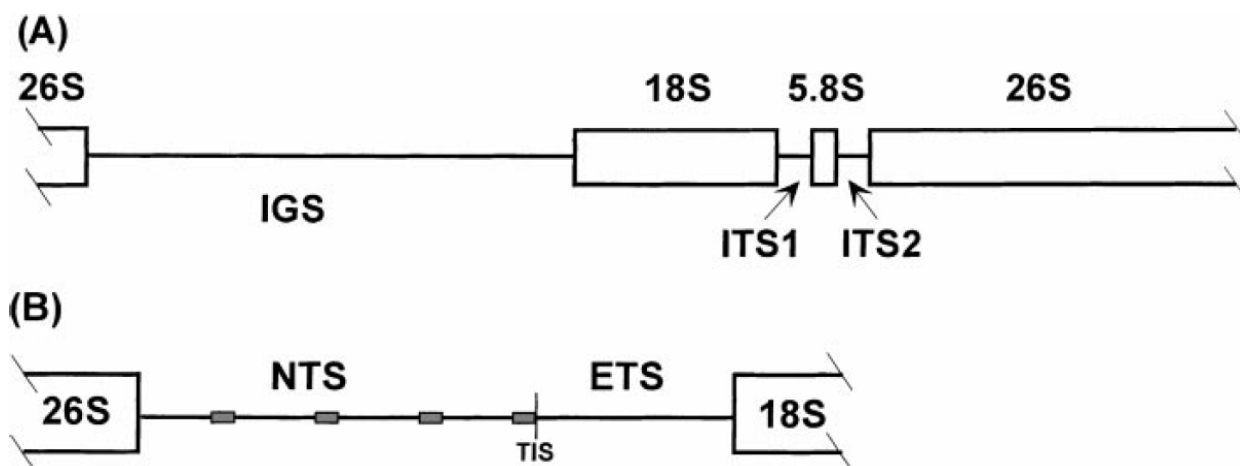


Figure 1-4: (A) A schematic of the IGS region between the 18S and 26S ribosomal RNA genes. (B) The IGS can be further divided into a NTS nontranscribed spacer and external transcribed spacer (ETS) regions. The transcription initiation site (TIS) is also indicated as being between NTS and ETS.

The cassette has a total base pair length of about 12 500–13 000 bp, but can vary considerably in length due to the length variability of the IGS (83)(100). The ITS region consists of two introns named ITS1 and ITS2 and a ribosomal RNA named 5.8S (9). As was previously mentioned, the ITS region is widely used in phylogenetic investigations as it consists of both translated/conserved and excised/variable regions. The conserved regions are suitable for investigating ancient diversification as they need to be conserved to maintain function and the introns are ideal for elucidating recent diversification as they are not under selective pressure and can thus mutate at higher rates (83),(9). Figure 1-5 indicates a schematic representation of the internal transcribed spacer region flanked by the 18S and 26S rRNA genes.

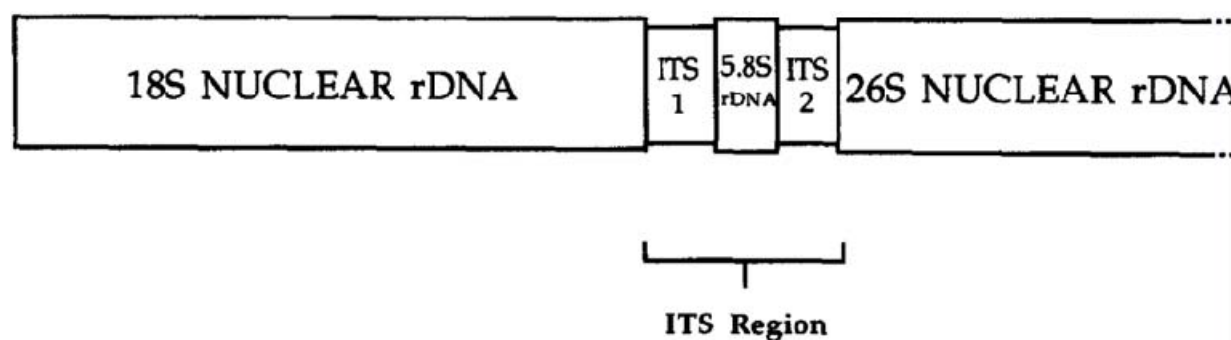


Figure 1-5: A diagram of the internal transcribed spacer regions marked as ITS1 and ITS2 (9).

Even though the ITS region is widely used, there are several advantages and drawbacks to using it. These are listed in Table 1-5.

Table 1-5: A list of benefits and drawback to using the Internal Transcribed Spacer regions of the nuclear plant genome (2, 9, 10, 46).

Advantages	Availability of almost universal primers working across large taxonomic groups is very appealing.
	The high copy number makes amplification very easy, even from herbarium specimens.
	A moderate size that facilitates amplification and sequencing without the need for internal primers.
	High mutation rates in non-coding regions usually provides enough phylogenetically useful information to perform evolutionary studies.
Postulated Advantages	Biparental inheritance as opposed to mostly uniparental inheritance of organellar genomes.
	Intragenomic uniformity as a result of active homogenization of repeat copies via concerted evolution.
	In hybrids and introgressants where concerted evolution has not progressed completely the ITS region may help identify progenitors of lineages under investigation
Disadvantages	Nucleolus Organizer Regions (NOR) with which the ITS is closely positioned to have been found to be able to change position in the nuclear genome. This violates the supposed orthology of the ITS region and could have implications for phylogenetic inferences.
	The sequencing products are the consensus sequence of all the loci in the genome that share identical priming sites. These loci might have different origins and might not be identical.
	Concerted evolution might not happen or might be incomplete across all loci which can distort phylogenetic inference.
	The DNA of rRNA molecules is under a selective pressure as the secondary structures which are created upon transcription need to be conserved which means one mutation might cause a secondary mutation event to allow the formation of the secondary structures. (Not all mutation are independent from one another)
	Due to the high mutation rate alignment can be very difficult when taxa are not closely related, leading to incorrect homology determinations. (Homoplasy)
Other reasons for widespread implementation	An initial lack of alternative highly variable and useful regions within any of the three main plant genomes.
	Most of the scientists studying the ITS region did not have a molecular background, but rather a background in taxonomy which were attracted to the very easy and universal protocol. (A bandwagon effect)

Since ITS might have these significant drawbacks as a phylogenetic marker other nuclear markers have also been identified and have been increasingly used.

Some of the more prominent genes are the phytochrome photoreceptors, PPR, TPI, LEAFY, ACCase, PGK, GBSSI, GPAT, ncpGS, GAP3DH, GIGANTEA, GPA1, AGB1, ADH, RBP2 and others (5, 8, 54, 60, 66, 70, 98, 105, 110, 111, 137, 216).

There are several benefits, as well as drawbacks to using nuclear information in phylogenetic reconstructions (185). The fact that nuclear information is inherited bi-parentally is of major importance as the evolutionary relationships retrieved from such studies will be derived from both parental lineages and not just one, as in the case of the chloroplasts or mitochondrial genomes. The nuclear genes that might be investigated can be from any of the chromosomes, meaning that they are unlinked.

There are disadvantages that might be encountered when using nuclear genes in that it might be difficult to distinguish orthologous loci from paralogous loci in analyses. It must also be kept in mind that complications might arise from concerted evolution, as well recombination within paralogous loci. There might also be intraspecific, intra-population and intra-individual polymorphisms within the gene region being studied (185).

1.2.2.4 The use of mitochondrial gene region sequences for phylogenetic inference

The plant genome that is the least studied for the purposes of phylogenetic inference is the mitochondrial genome. The reason for this that the mitochondrial genome in plants is known to undergo from minor to major horizontal gene transfer events with the nuclear genome (19, 90). There are several studies in the literature which have utilized mitochondrial genetic sequences, either separate, or in conjunction with nuclear and chloroplast markers (47, 159, 188). Typical genes are *atp1*, *matR*, *nad5* and *rps3*.

1.2.3 Phylogenetic relationships of the Angiosperms based on a molecular systematics approach

One of the most ambitious projects involved in the systematic classification of plants was started in the early 1990's. It seeks to classify all angiosperms (colloquially referred as flowering plants) in a complete phylogenetic tree. The group of scientists that undertook this enormous task called themselves the Angiosperm Phylogeny Group, or APG for short. The first phylogenetic tree published by this group appeared in 1998 and was called APG I and set out to identify an ordinal classification of angiosperm families. Their second publication, published in 2003, and retained the name of APG II. This publication expanded on the first to classify angiosperm families within orders. The latest publication by the Angiosperm Phylogeny Group appeared in 2009 and was named APG III. This is currently one of the most comprehensive phylogenetic trees of its kind, but additional genetic markers were also used in a publication in 2011 (188). APG III analyses 413 of the 415 recognized plant families which all fall into 59 currently recognized orders. Through their initial publication the Angiosperm Phylogeny Group showed the immense power of molecular phylogenetics as an important tool at the disposal of botanists to classify flowering plants. Relatedness could, for the first time be measured at its most fundamental state, namely the molecules that make up the genetic code of life.

It is important to note that in APG III the family Zygophyllaceae was placed sister to Krameriaceae in the order Zygophyllales for the first time. Other authors, e.g. Takhtajan were already using the name Zygophyllales for this order as early as 1997 based only on morphological characters only (201). Other earlier publications based on molecular, as well as anatomical characters also suggested the designation of a new order containing both these families (23, 29, 172).

1.2.3.1 Recent advances in molecular systematics

Science and technology has advanced far enough to make it possible to sequence hundreds and in some cases thousands of base pairs per organism and draw conclusions based on this information. Since 2005, when the first Next-Generation Sequencing platform (454 Pyrosequencing) became available, the technology for the generation of significantly larger amounts of sequence data became available and botanists were no longer limited to the analyses of singular genetic markers. Whole chloroplast and mitochondrial genomes could be sequenced with ease and even whole plant genomes could be sequenced at unprecedented speed and cost.

With the ability to generate vast amounts of sequence data, tens to hundreds of genes can be sequenced and discovered. This also means that these genes are at the disposal of botanists who attempt to determine phylogenies of the plant kingdom. More data means that the conclusions based on the data become more reliable and better informed. However, the vast amounts of molecular data need appropriate software and computing power to analyse it and therefore to be able to draw conclusions from it.

Several breakthroughs and advancements in technology have led to the development of new sequencing technologies able to sequence millions of DNA sequences at one time in parallel. All of these new sequencing technologies are based on a shotgun sequencing approach. This means that DNA is shredded to smaller manageable fragments and is then sequenced in a massive parallel sequencing array. These smaller fragments are then overlapped with one another to generate contiguous sequences, named “contigs” for short, of ever increasing sizes. Most of these technologies are termed as being of Second Generation sequencing platforms. There is also at the time of writing a platform available that is being termed as being a Third Generation sequencing platform. The main difference between second and third generation sequencing is that third generation sequencing has no need of a amplification procedure for any DNA fragments as the sequence is determined from a single DNA strand. This has major advantages over previous sequencing platforms as this eliminates potential errors during the DNA amplification steps used in second generation sequencing platforms. The next section will cover several second generation, as well as one third generation sequencing platform. The second generation sequencing platform from Life technologies, the Ion semiconductor sequencer named Ion Torrent, was use in the course of this thesis.

1.3 The Next-Generation Sequencing platforms

Five of the Next-Generation Sequencing platforms that are currently available are described in this section. These are the 2nd generation sequencing platforms such as the 454 Pyrosequencing

platform from Life Sciences (Roche), the polymerase-based platform from Illumina, the ligation-based SOLiD platform from ABI (Life Technologies), the ion semiconductor sequencing also from Life Technologies and the first commercially available 3rd generation sequencing platform from Pacific BiosciencesTM based on Single Molecule Real Time Sequencing or SMRT (called Smart sequencing) which requires no prior amplification step of DNA as the sequence is determined from a single DNA strand as the bases are incorporated.

1.3.1 Pyrosequencing (454 sequencing)

Pyrosequencing is based on the “sequencing by synthesis” principle. Rather than chain termination (Sanger sequencing), it relies on the measurement/detection of a pyrophosphate molecule which is released upon the incorporation of a nucleotide. One of the four nucleotides is added at a time and if it is incorporated it emits photons. After each round the nucleotides that were not incorporated are degraded before the next nucleotide is added. This approach is repeated for each of the four nucleotides until the sequence is determined. The technique was developed in Stockholm in 1996 by Pål Nyrén and Mostafa Ronaghi at the Royal Institute of Technology (136, 163, 164). Roche Diagnostics, in October 2013, confirmed that it was discontinuing the platform (162, 184).

1.3.1.1 Library building and emPCR

Library building involves the degradation of the DNA in question via nebulization and the attachment of adaptor molecules to the terminals of the DNA fragments. One of the adaptors (adaptor B) is a 5'-biotin tag which binds a streptavidin molecule on the surface of the bead. A single one of these modified fragmented DNA strands is bound to a magnetic bead. These beads are encapsulated in individual emulsion droplets and the unique DNA copy is amplified on the surface of the bead up to 10^7 times (44, 108). In Figure 1-6 and Figure 1-7 the first steps for pyrosequencing are indicated as well as the approximate time involved for each step.

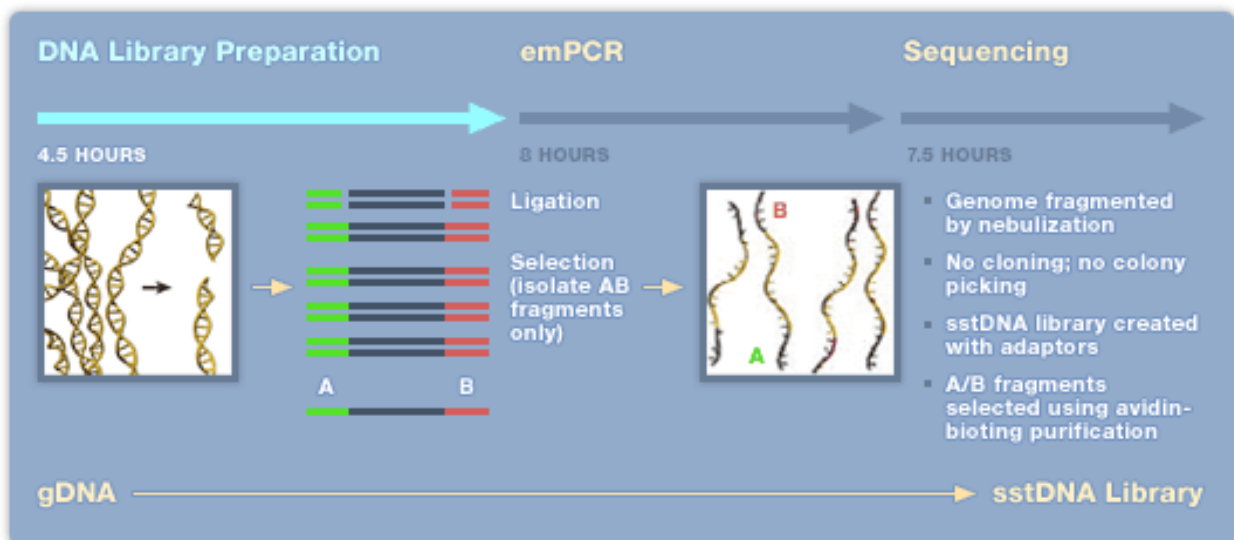


Figure 1-6: Above is diagram depicting the workflow of a typical pyrosequencing run. The library preparation step takes ~4 hours, the emulsion PCR step takes ~8 hours and the sequence determination step takes 7.5 hours. The images above only focus on the DNA library preparation step, indicated by the light blue arrow. DNA is fragmented using a nebulisation technique. A single stranded DNA library (sstDNA) is generated by adding A and B adapters. Only fragments containing one “A” adapter and one “B” (5’-biotin tag) adapter are selected by using an avidin-biotin purification step. Beads contain streptavidin molecules on their surface (44).

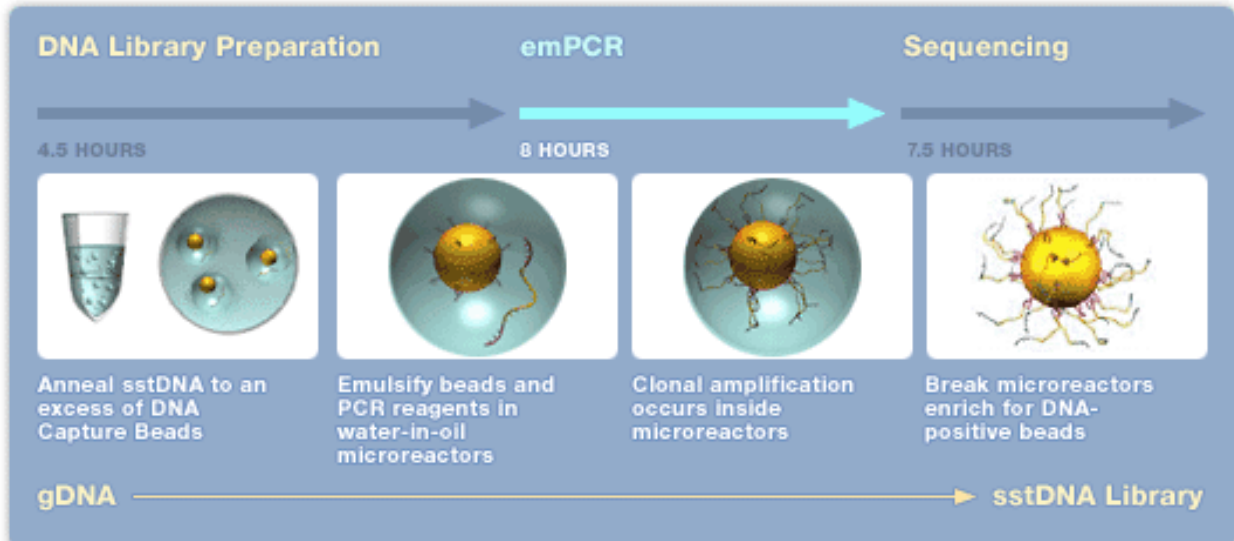


Figure 1-7: The images above show the emulsion PCR (emPCR) indicated with a light blue arrow. After the avidin-biotin purification step that attaches a sstDNA molecule to a bead, the beads are encapsulated in an emulsion that contains PCR reagents that clonally amplify the unique DNA fragment of each bead on the surface of the bead (44).

1.3.1.2 Sequencing

The beads are subsequently transferred to a microtitre plate and are fixed in wells. The sequences of the fragments are analyzed via pyrosequencing. Pyrosequencing relies on the

measurement of a released inorganic pyrophosphate molecule upon integration in the DNA sequence. Measurement is facilitated by means of a light reaction or chemiluminescence (44, 108). In Figure 1-8 the diagram shows how the sequencing beads are placed in wells on the microtitre plate.

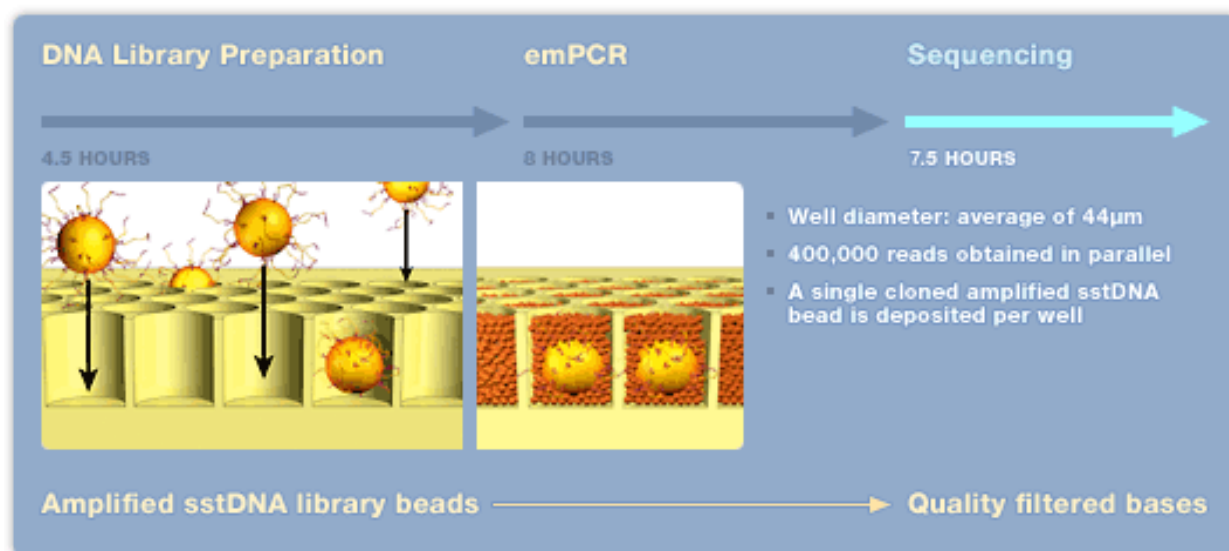


Figure 1-8: This image shows the first step in the last sequence of events which forms part of sequencing, indicated by the light blue line. The wells in the microtitre plate are just big enough to allow one bead to adhere. This technique makes it possible to obtain massive amounts of sequence information in parallel (44).

The chemiluminescent chemistry involves the enzymes DNA polymerase, ATP-sulfurylase, luciferase and apyrase (44, 108). Two substrates are also required, namely adenosine 5'-phosphosulfate (APS) and luciferin.

Deoxyribonucleotides triphosphates (dNTPs) are added one at a time to the reaction (44, 108). The enzyme DNA polymerase is responsible for the incorporation of dNTPs. If a nucleotide is complementary to the DNA base on the template strand a pyrophosphate molecule is released. The amount of pyrophosphate released is directly proportional to amount of nucleotides incorporated.

APS and PPi are converted into ATP via ATP sulfurylase, which drives the conversion of luciferin into oxyluciferin via luciferase (44, 108). This reaction produces the light which is detected by a charge coupled device chip and which is used to monitor nucleotide incorporation. The amount of light generated is directly proportional to the amount of nucleotides incorporated. The nucleotide incorporation reactions are visualized by means of appropriate software packages.

Before the next nucleotide in the sequence can be added, the remaining unincorporated nucleotides and ATP from the previous round need to be removed so as not to interfere with next round (44, 108). Apyrase is the enzyme that can degrade both of the molecules. Interesting to note is the use of alpha-thio triphosphate (dATPS) as a substitute molecule for the natural molecule dATP, as it can be used by DNA polymerase, but is not specific to luciferase which would recognize dATP and give false positives. In Figure 1-9 and Figure 1-10 the sequencing chemistry of pyrosequencing is discussed.

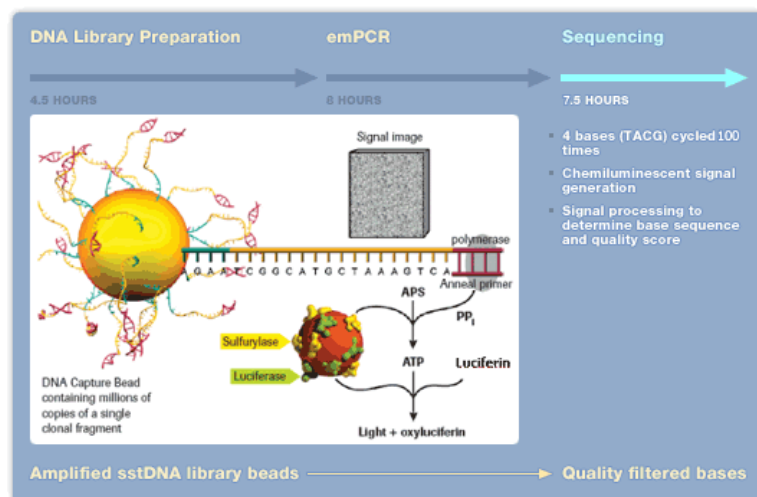


Figure 1-9: The image shows part of the sequencing step above in light blue. The four bases of the genetic code are cycled 100 times. If a base is incorporated an inorganic pyrophosphate is released by the incorporation reaction (APS = Adenosine-5'-phosphosulfate) (44).

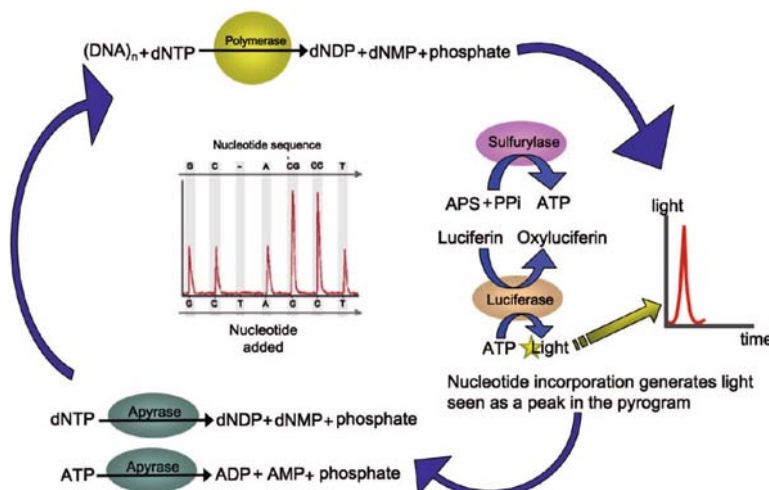


Figure 1-10: A diagram of the mechanism of pyrosequencing. If a nucleotide is successfully incorporated into the DNA strand a pyrophosphate molecule is released. ATP-sulfurylase combines this pyrophosphate molecule of one APS molecule to produce ATP. The ATP molecule is used by luciferase to produce oxyluciferin and light. The light is detected by the charge coupled device chip. If a nucleotide is not incorporated there will not be a peak on for that base that was added to the system. In both cases apyrase degrades all nucleotides that were not incorporated, as well as all ATP before and after each new nucleotide is added to the reaction mixture (150).

1.3.1.3 Limitations of the Technology

One of the biggest drawbacks to the technology is read error in homopolymer stretches, e.g. in poly-A regions in which there might be six adenosine residues, but the instrument records seven. This can result in indels in the DNA sequences that are not necessarily correct. Another limitation is short read lengths, which are currently between 200-500 bp in length, where standard Sanger sequencing reads are between 500 and 1000 bp (72, 89, 103, 129, 133).

1.3.2 Polymerase-based sequence-by-synthesis (Illumina[®])

Polymerase-based sequence-by-synthesis sequencing or colloquially known as Solexa-sequencing, named after the company which pioneered the technology, is one of the first commercially available next-generation sequencing technologies available. The principle behind this method involves attachment of single stranded DNA molecules and primers to a slide where after amplification occurs via DNA polymerase creating clusters or clonal colonies. To determine the sequence of each colony, four reversible terminator bases (RT-bases) are used. Non-incorporated bases are washed away. Images are taken by a camera visualizing the fluorescently labeled nucleotides. After visualizing the bases, the fluorescent labels, as well as terminal 3'- blocker are chemically removed allowing for the next base to be incorporated. Unlike pyrosequencing or ion semiconductor sequencing, only one base is incorporated at a time (77).

1.3.2.1 Library building

The DNA to be sequenced is sheared with a nebulizer. These fragments are polished or refined and two unique adapters are added to the terminals. These ligated fragments are separated on a agarose gel and the fragments between 150 and 200 bp in length are isolated and amplified by means of PCR (20, 117).

1.3.2.2 Cluster generation by means of Bridge Amplification

Repeated denaturation and extension cycles results in localized amplification of single molecules in millions of unique locations across the flow cell surface (7, 20, 76, 77, 103, 115, 118, 160). This process occurs in what is referred to as Illumina's "cluster station", an automated flow cell processor. The Illumina system uses a unique amplification step that is different from 454 and ABI systems where beads are encased in emulsions to generate polymerase colonies, called "colonies" for short. The "Bridge Amplification" occurs on the surface of the flow cell itself. The surface of the flow cell is covered in short single stranded oligonucleotides that correspond to the ligated adapters of DNA samples. These ssDNA-adapters complexes bind to the

oligonucleotides on the surface of the flowcell and are exposed to the reagents necessary for the polymer-based extension. “Priming” of the DNA strand occurs when the free end of the ligated strand bridges to a complementary oligonucleotide sequence in close proximity to the bound DNA strand. Repeated cycles of denaturation and extension allow for the formation of millions of small clusters of unique single stranded DNA molecules over the surface of the flow cell surface.

1.3.2.3 Sequence determination

The flow cell containing the millions of unique colonies “polymerase colonies” is loaded into an appropriate sequencer for automated extension and imaging cycles (7, 20, 76, 77, 103, 115, 118, 160). Each cycle consists of the incorporation of a single fluorescently labeled nucleotide and a subsequent high resolution imaging step of the entire flow cell. The generated image is the information of all the first bases for every single colony on the entire flow cell, the fluorescent label identifying the specific base pair in question.

The fluorescent label is removed and the next round of base pair incorporation and measurement is continued. Base calling is achieved with an algorithm that measures colour emission over time. Reads that can be used to assemble contigs are usually between 26 and 50 bp in length. In Figure 1-11 the principles of the Illumina sequencing platform is discussed.

1.3.2.4 Limitations of the technology

One of the main concerns of the technology is the possibility of faulty base incorporation, since the technology relies on modified polymerases. Another slight drawback is the length of the reads generated, which are relatively short and can hamper contiguous sequence assembly (103).

1.3.3 Ligation-based sequencing (SOLiD)

The third type of sequencing technology that will be discussed is the ligation-based sequencing developed by Life technologies, named the SOLiD system (Sequencing by Oligonucleotide Ligation and Detection). This technique is based on the colony sequencing method. The construction of the DNA library is similar to that of the other technologies in that a single sheared DNA fragment is attached to a magnetic bead and amplified on that bead via an emulsion PCR. These beads are then attached to a glass surface where sequencing occurs through various cycles of hybridization and ligation with 16 different dinucleotide combinations.

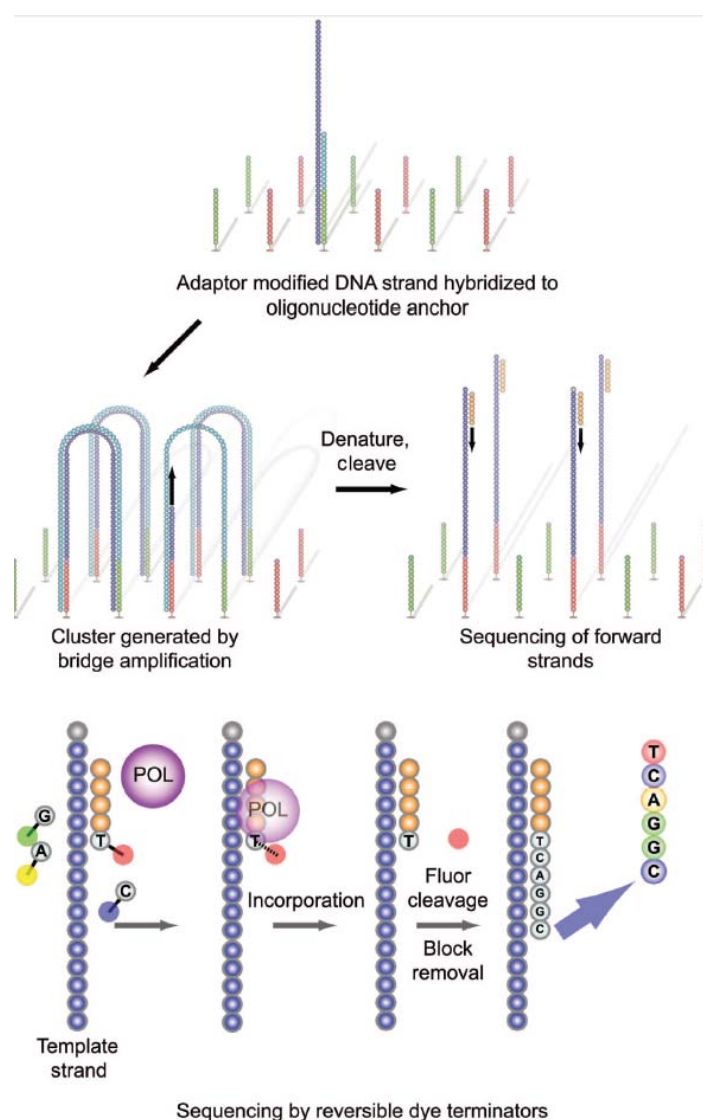


Figure 1-11: A diagram indicating the basic principles of the Illumina system. Single stranded DNA modified with adapters is added to the flow cell after which they are immobilized via hybridization. Clonally amplified clusters are generated by bridge amplification. These clusters are denatured and cleaved; initiation of sequencing is achieved with the addition of primer, polymerase and 4 reversible dye terminators. Fluorescence is measured after the reversible dye terminator is incorporated. The fluor and block are removed and the next synthesis cycle starts (208).

1.3.3.1 Library building

As before the DNA is sheared via nebulization to appropriate sizes. A single strand of sheared DNA is attached to the surface of a magnetic bead. Each unique DNA fragment has an identical adapter attached to it named the P1 adapter. These unique fragments are clonally amplified via an emulsion PCR step. Beads are covalently bound to a derivitized-glass flow cell surface (3, 116, 117, 133, 134).

1.3.3.2 *Sequence determination*

Although the initial steps of the procedure are almost identical to other NGS systems currently on the market, the sequence determination is much more unique in that this technology utilizes DNA ligase and ligation instead of DNA polymerase and DNA synthesis.

As mentioned above all DNA strands contain a universal P1 adapter. The primers needed to perform the first round of sequencing binds to this adapter. In this case the primer does not provide a free 3' hydroxyl group that would be needed for extension with normal sequence determination, but rather a 5' phosphate group for ligation to the interrogation probes in the first "ligation-sequencing" step. The interrogation probes are octamers, which include two probe-specific bases (3'-5') and six degenerate bases and one of four fluorescent labels linked to the 5' end. There are 16 di-base probes (e.g. AT, GA etc.). For the first sequence-ligation step all 16 interrogation probes and thermo stable DNA ligase are used. These probes compete to bind/anneal to the DNA template strand next to the primer. The ligation is performed after the annealing of the probe to the template strand. The unbound probes are subsequently washed away. Each bound probe is measured for its fluorescent signal before being cleaved. Another wash step is performed to remove the fluor and regeneration of the 5' phosphate group. In subsequent ligation sequencing steps the interrogation probes are ligated to the 5' phosphate group of the preceding pentamer. Seven such ligation steps are referred to as a "round" and are used to extend the first primer. The newly synthesized strand is then denatured and a new sequencing primer offset by one base (n-1) is annealed to the adapter sequence. Five such "rounds" are performed for each subsequent offset primer used (n-2, n-3, etc.). Using this approach each template nucleotide is sequenced twice. A sequencing run takes approximately 6 days creating sequence read lengths of 35 bp each. The sequence is inferred by interpreting the fluorescent signals of the di-base probes (72, 103, 127, 129, 153, 169, 208) . In Figure 1-12 and Figure 1-13 the basic principles of the SOLiD sequencing system are indicated.

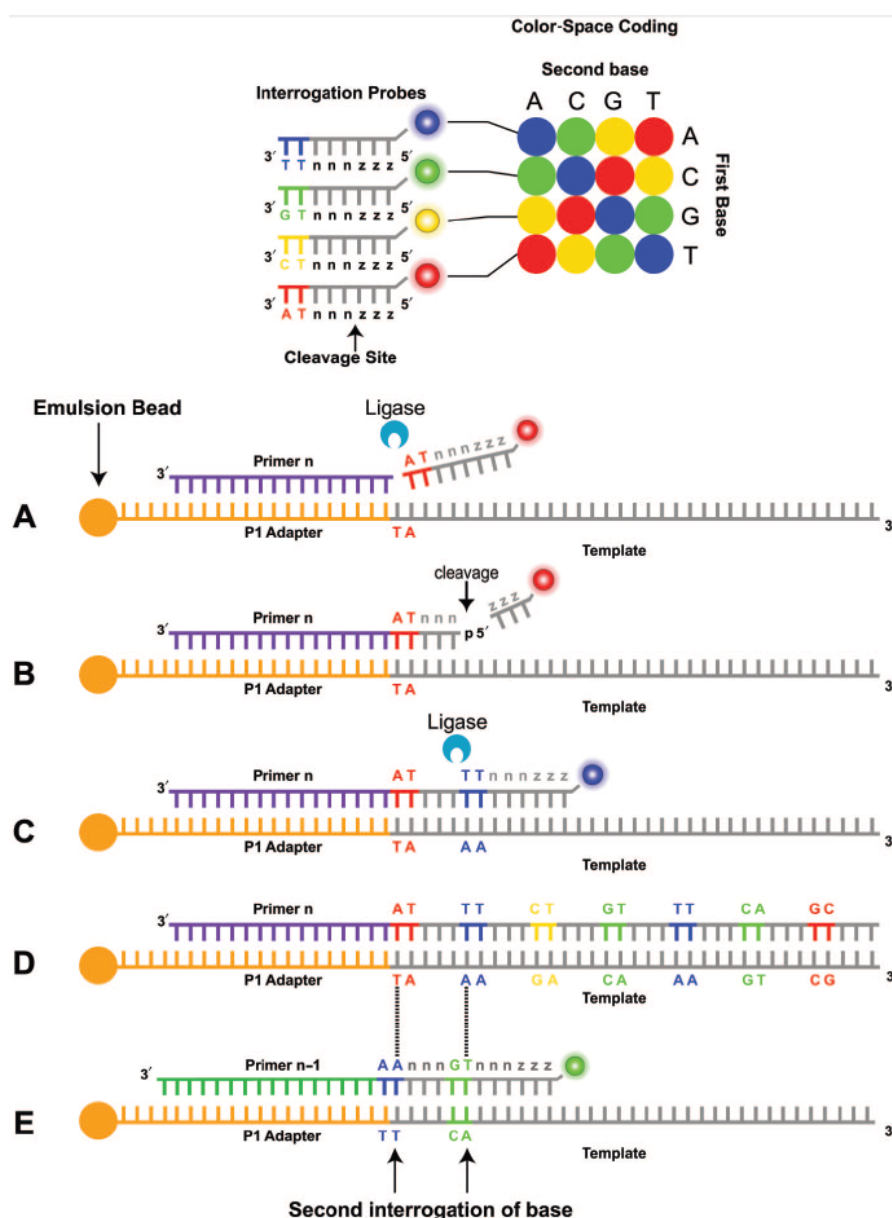


Figure 1-12: A diagram indicating the repeated cycles performed during SOLiD sequencing. At the top of the image is the colour key of the di-base probes. Each of these probes is an octamer which contains two probes specific bases (3'-5') and six degenerate bases, as well as one of four fluorescent labels at the 5' end. A di-base probe is one of the sixteen possible di-base combinations ($4^2 = 16$). Below the probe key is a diagram showing the basic principles of the SOLiD sequencing system. (A), A bead containing the P1 adapter and an attached DNA fragment. Primer n is bound to the adapter. The primer is "interrogated" by all sixteen di-base probes. In the diagram, the probe with the bases AT is complimentary to the DNA fragment. (B), Fluorescence is recorded after the primer has been annealed and ligated. The last three degenerate bases are removed and the newly generated 5' end is phosphorylated before subsequent sequencing steps. (C), A subsequent di-base probe molecule is annealed and ligated. (D), Completion of the first "round" of sequencing, consisting of seven ligation cycles. (E), The sequencing product of primer n is denatured from the adapter/template complex. A second primer is annealed offset by one base (n-1). Using progressive offset primers in this way, adapter bases are sequenced and this known sequence is used along with the colour-space coding to ascertain the DNA template sequence by means of deconvolution (208).

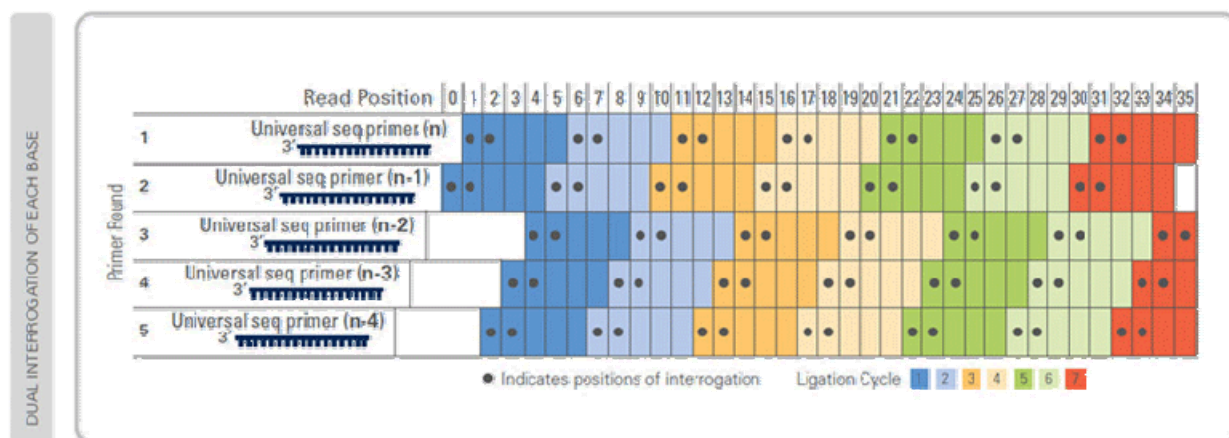


Figure 1-13: A total of five rounds of primer reset reaction are completed for each sequence tag. Through this process virtually all bases are interrogated in two independent ligation reactions by two different primers (141).

1.3.4 Ion semiconductor sequencing (Ion Torrent™)

This technology is also relatively new being released in late 2010. It is very similar to 454 pyrosequencing technology, but it measures a drop in pH when a proton is released as a base is incorporated into the DNA being synthesized instead of measuring the released pyrophosphate as the 454 system does. The change in pH is measured by an ISFET or ion-sensitive field-effect transistor. A single nucleotide is washed over the sequencing chip at a time and a change in pH is only measured when a base is incorporated into the DNA being synthesized. At time of writing there are two ion semiconductor sequence platforms available, the newer Ion Proton and the Personal Genome Machine (PGM). The main differences between the two are the distance between the wells and the size of sample-deposition surface. These two attributes combined determine the amount of data that each system can provide. Both of these systems also have different capacities depending on the type of chip used. The PGM platform currently has a maximum output of 1 Gb (1 000 000 000 bp) if using the largest chip. The sequence reaction takes 2-3 hours to complete and generates approximately 5 500 000 reads with an average length of 200 bp. The Ion Proton platform has a maximal output of 100 Gb (100 000 000 000 bp) in 2-4 hours generating more than 300 000 000 reads with an average read length of 200 bp. The PGM is thus more suited to targeted resequencing of gene panels or smaller genomes, i.e. chloroplasts, mitochondria or bacteria, while the Ion Proton is more suited for whole genome and exome sequencing (127). In Figure 1-14 the basic principles of ion semiconductor sequencing are explained.

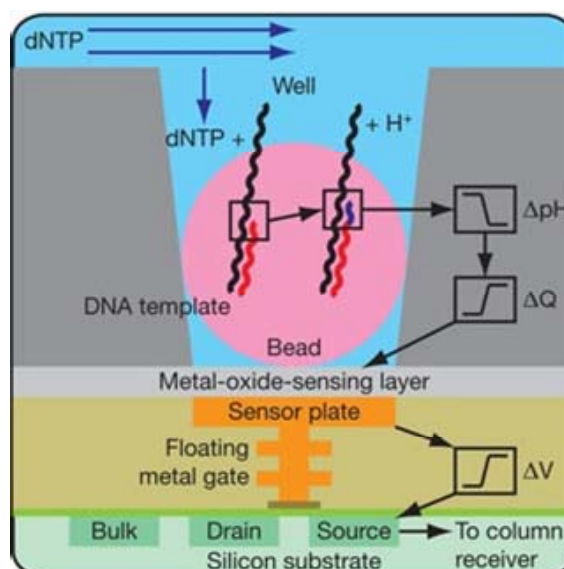


Figure 1-14: The diagram shows the chemical reaction when a nucleotide base is incorporated. Unlike pyrosequencing which measures the released pyrophosphate molecule, the ion semiconductor sequencer measures the released proton. When a nucleotide base is washed over a specific well and does not match the complementary base, no signal is detected. When the nucleotide base is complementary it is incorporated and the drop in pH is measured as a proton is released. When more than one base is incorporated the signal will be proportionally stronger (166).

1.3.4.1 Library building

Library building for ion semiconductor sequencing is similar to that of other 2nd generation sequencing platforms. The DNA that will be sequenced is sheared to appropriate sizes using a nebulizer, sonication or enzyme-based degradation methods. This leaves short DNA fragments that are recessed, overhanging or blunt in nature. Blunt-ended fragments are generated by either digesting back or filling in the 5' and 3' overhangs. The 5' ends of fragments are subsequently phosphorylated to ensure that ligation of the fragments is possible. Adapter sequences are ligated to the blunt ends of the DNA fragments (134).

Singular fragments are attached to non-paramagnetic beads and an emulsion PCR step is performed. Each unique fragment is amplified on each bead. Beads which did not contain any attached DNA are removed and a single bead is deposited per well on the ion semiconductor chip. These wells are only large enough to contain a single bead. The well is positioned above a sensor plate (ISFET) which is sensitive to changes in pH (127, 166).

1.3.4.2 Sequence determination

Sequencing is achieved by flooding the entire chip surface with one of the four DNA bases. If a base is incorporated a proton is released. This released proton causes a drop in the local pH

above the ISFET sensor plate. The amount of ions released is proportional to the amount of bases that are incorporated (127, 166).

1.3.4.3 Advantages and disadvantages of the platform

One of the major advantages of the platform is the relative low upfront costs involved. This can be achieved, because there is no need for an imaging system which can be expensive and time consuming. This platform also does not use modified bases to perform the polymerization (modified bases can be relatively more expensive) (39, 149). The PGM sequencer itself is also priced very reasonable (\$50 000 USD at release) (87). Bases in repetitive regions can also be attached at the same step in the sequencing reaction. The signal generated, however is not perfectly linear and can lead to misinterpretations in homopolymer stretches longer than seven bases (127, 166). Another drawback is relatively short read lengths that can be obtained from the system. The read lengths are constantly being increased and have reached 400 bp in 2012 (78, 81). Short read lengths are detrimental to *de novo* assembly of genomes.

1.3.5 Single molecule Real-Time Sequencing (SMRT Sequencing)

The term Next-Generation Sequencing is a term usually reserved for techniques that are classified as being “second generation sequencing”. Single molecule Real-Time Sequencing (pronounced Smart Sequencing) is a revolutionary approach to sequencing DNA and is classified as being of the third generation. As the name suggests sequencing is achieved by analyzing a single strand of DNA in isolation from all other DNA strands being sequenced in a massive parallel sequencing array. This technology has been available since 2011 and in October 2013 reads of more than 30 000 bp in length had been reported.

1.3.5.1 Basic principles of SMRT sequencing

As this technology makes it possible to sequence single strands of DNA it eliminates the necessity of needing to amplify pieces of DNA to be able to achieve high enough concentrations of specific fragments in order to detect them. The sequencing reaction takes place in what is known as a zero-mode wave guide (ZMW) (92, 97, 103, 127, 130, 151, 160, 161, 183). There are several thousand ZMWs on what are known as SMRT cells. Each ZMW contains a single affixed DNA polymerase enzyme. Since each ZMW contains only one DNA polymerase molecule only one DNA molecule can interact with each ZMW. The ZMW has a volume that is so small that it is possible to measure the incorporation of a single nucleotide base by the DNA polymerase. A diagram illustrating the technique is shown in Figure 1-15.

The nucleotide bases that are used in SMRT sequencing are modified molecules of the ones found in nature. They contain a fluorescent dye tag that is cleaved upon incorporation of the base. Each of the four bases has a unique fluorescent colour to distinguish them from each other. The colour is measured when the tag is cleaved before it moves out of the detection area.

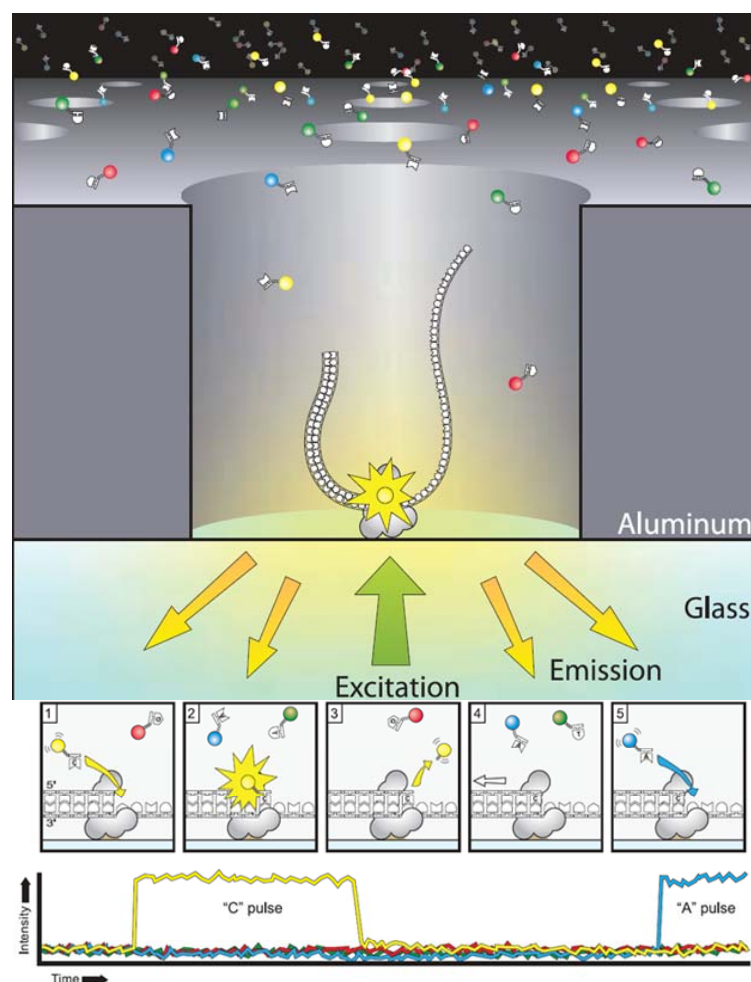


Figure 1-15: The diagram above shown the the basic principles of SMRT sequencing. A single DNA polymerase is attached at the bottom of each well. Only a single ssDNA molecule can interact with each DNA polymerase at a time. Fluorescently labelled nucleotide bases are used instead of naturally occurring bases. When a labelled base is incorporated a corresponding light signal is emitted and detected allowing for real-time sequence dtermination of the DNA strand (49).

This technology went into the public domain in 2011 (144) after beta testing in 2010 (143). Upon release the mean read length was around 1 100 bp, but as of 2014 Pacific Biosciences claim that depending on the quality of the DNA, that over half the reads generated are over 14 000 bp long with the longest over 40 000 bp in length (145).

The SMRT cell technology has also been updated. Prototypes had 3 000 ZMW wells per cell. At commercial release that was increased to 150 000 ZMW holes that were analysed as two sets of

75 000 (187). In April 2013, a new sequencer was released which utilized all 150 000 ZMW holes in a single analysis which doubled the throughput (135, 142).

1.3.5.2 Limitations of the technology

Even though this technology is a significant technological breakthrough, there are still some limitations. The single read error rate is still high, around 14%. There is no bias for the erroneous base calling though and a statistically a 8X genome coverage is recommended to correctly report a nucleotide as the odds of reporting the wrong base eight times with no bias in erroneous base calling is negligible (161). Another disadvantage is that the amount of data generated is less than other NGS systems, but improvements are being developed to address this issue. Another issue is the need for a very pure sample as the procedure is very sensitive to contaminants and damaged sample, i.e. damaged DNA (161). In Figure 1-15 the basic principles behind SMRT-sequencing are depicted.

1.4 The chloroplast genome

Chloroplasts are organelles within the cells of photosynthesizing plants, algae, bacteria and other organisms. As with mitochondria, chloroplasts have their own DNA separate from that of the nucleus. There can be tens to hundreds of chloroplasts per cell while the nuclear genome normally only has two copies of each chromosome, making the chloroplast gene concentration much higher than their nuclear counterparts, and thus much easier to sequence for use in phylogenetic inference.

Chloroplasts are mostly only inherited maternally in Angiosperms, but as mentioned in section 2.6.2.1 a study has shown that within Zygophyllaceae in the subfamily Larreiodeae these organelles can be inherited paternally or even biparentally (215).

A reason for studying chloroplast genomes rather than nuclear or mitochondrial genomes is that substantially more chloroplast genomes have been successfully sequenced and annotated than whole genomes, making it possible to choose a closely related species of plant to use as a reference genome.

1.4.1 Structure of the chloroplast genome

In most cases the chloroplast genome is a single structure, circular in shape, the exception being those of dinophyte algae, and varies roughly between 120 000 – 160 000 bp in size, although some parasitic plants have chloroplast genomes smaller than 60 000 base pairs. Certain algae also have smaller chloroplast genomes, e.g. *Bigeloviella natans* at around 69 200 bp. The largest chloroplast genome discovered so far is found in the genus *Pelargonium* in the order Geraniales

at around 218 000 bp. The circular chloroplast genome can be divided into four basic regions, namely a large single copy-, a small single copy-, and two inverted repeat regions (see Figure 1-16). The inverted repeat regions are two identical copies, but are in reverse orientations in the genome. There are, however, plants within the legume family which only have one copy of the inverted repeat region. These plants are named the inverted repeat lacking clade (IRLC).

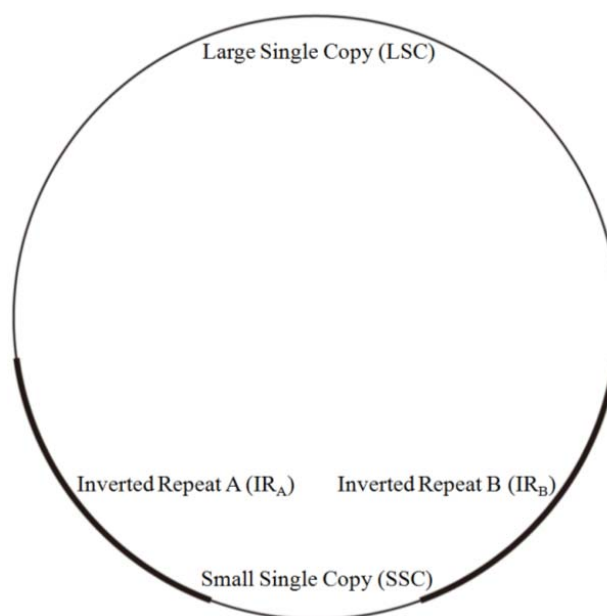


Figure 1-16: Above is a diagram indicating the structure of most chloroplast genomes. Firstly there is a large section known as the Large Single Copy (LSC) region, secondly a smaller section known as the Small Single Copy region (SSC), and lastly two identical Inverted Repeat Regions (IR_A and IR_B).

1.5 Phylogenomics - The use of NGS data for phylogenetic inference

Why would one want to be able to sequence an organism's entire genome? In most cases single genes are not sufficient to distinguish interordinal or interfamilial relationships, e.g. *rbcL* (32, 179). The ITS region is ideal for distinguishing interspecies relationships, but can fail to resolve deeper relationships. Even though studies like these have greatly advanced our understanding of the evolutionary history of plants, they still possess distinct shortcomings. These studies only focus on very small subsets of the total amount genetic data available. There are areas in plant history where ancient divergence and subsequent extinction events have left large gaps in extant lineages. The inverse problem also exists where ancient adaptive radiations have led to an explosive diversification over a very short time span (short internal branches) (11). In the latter case it requires thousands of informative characters to resolve rapid radiation/speciation events. Numerous phylogenetic informative characters (thousands) may be necessary to resolve these types of phylogenetic relationships.

The most effective manner to resolve these phylogenetic relationships is to sequence the entire genomes of the organisms under investigation, or at least be able to consistently sequence the same subsets of the genomes of different organisms in order to perform phylogenetic studies. This has led to the development of a new field in systematics, namely phylogenomics, in which genome-wide genetic information is used to resolve phylogenetic relationships (152).

Of the three genetic compartments found in plants, i.e. nuclear, mitochondrial and chloroplast, the most extensively studied is the chloroplast genome. With the advent of NGS technologies it is fairly easy to generate all the necessary data to assemble an entire genome of an organism in a single sequencing experiment. It is even easier to sequence the entire chloroplast genome of photosynthesizing organisms as there can be hundreds of chloroplast organelles in a single cell and the genome is much smaller than the nuclear genome. The first chloroplast genomes to be sequenced were those of tobacco (*Nicotiana tabacum*) and the liverwort (*Marchantia polymorpha*), as early as 1986 (138, 181). The sequences of these genomes were obtained by generating overlapping clone libraries of the chloroplast genomes using restriction enzymes and inserting these fragments into cosmids. Physical maps of these fragments were generated using either Maxam-Gilbert- and Sanger sequencing (113, 171). This basic approach remained the main procedure to obtain whole chloroplast genomes right up to the first commercial NGS platforms in the 2006. The first chloroplast genomes to be sequenced using a NGS platform were those of *Platanus occidentalis* (American Sycamore) and *Nandina domestica* (Sacred Bamboo), by Moore *et al.* in 2006 (125).

One of the first phylogenetic studies, using chloroplast genome sequences generated by NGS technologies, was published in 2007 by Moore *et al.* (124). This study utilized 61 coding genes for 45 taxa. In 2010, a paper was published by Moore *et al.* which contained 83 coding genes for 86 taxa. Importantly this paper also included NGS plastid data of *Bulnesia arborea* from the family Zygophyllaceae, subfamily Larreioideae, the closest subfamily related to Zygophylloideae (126). The most comprehensive phylogenetic analyses of plants to date were published by Ruhfel *et al.*, in early 2014 (see Figure 1-17). This study contained 78 plastid genes of 360 taxa spanning the entire Viridiplantae, i.e. green plants (168).

As mentioned above, single genes in some cases could not resolve phylogenetic relationships among rapidly diverging lineages. Several studies have emerged which attempt to resolve such relationships with complete or nearly complete chloroplast gene sets (11, 12, 71, 88, 197).

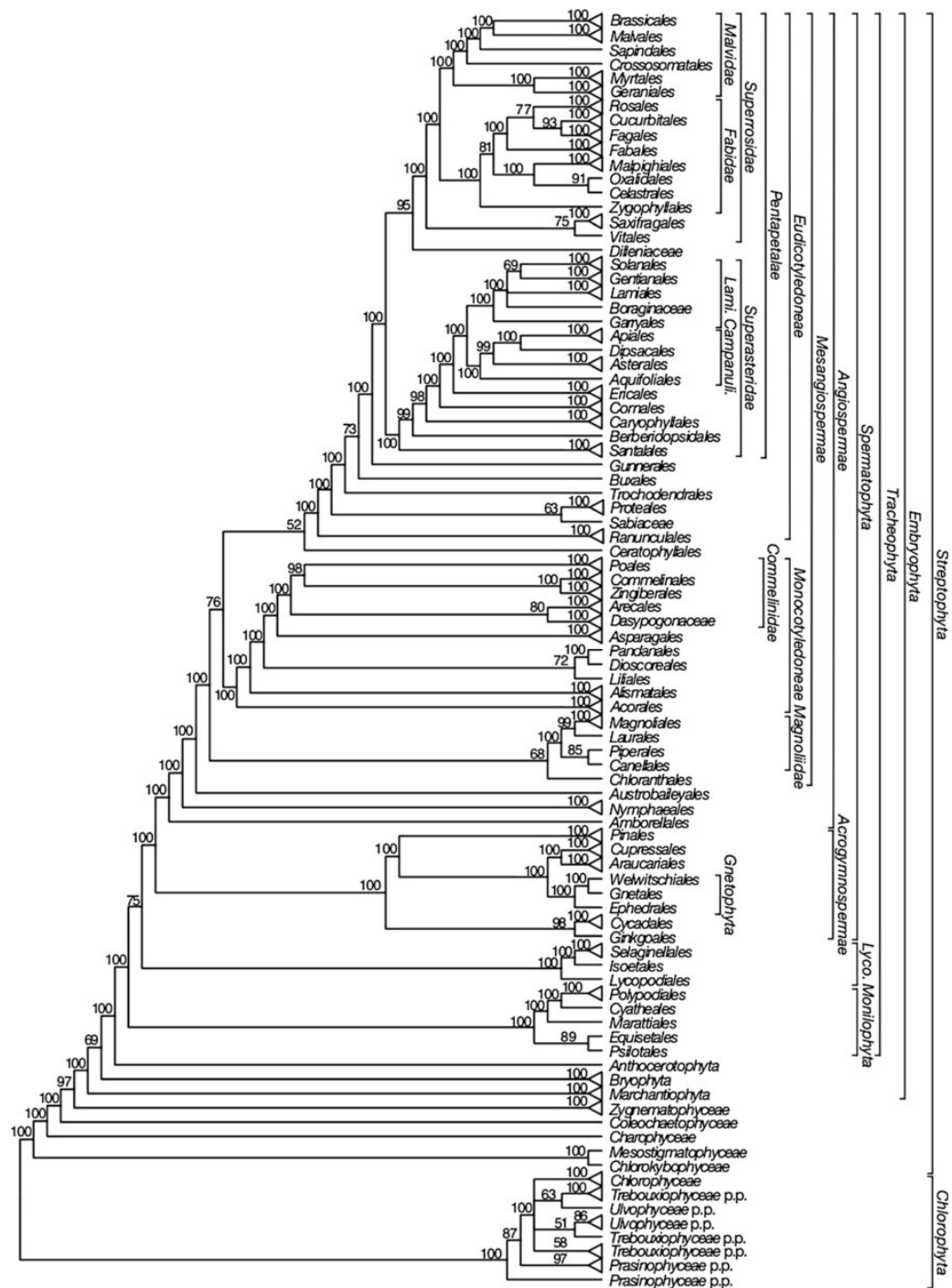


Figure 1-17: The phylogenetic tree retrieved from 78 chloroplast genes of 360 taxa. The tree has mostly been collapsed to show only the relationships at an ordinal level (168).

1.6 Systematics of *Zygophyllum*

1.6.1 Systematic classification of *Zygophyllum*

The name *Zygophyllum* derives from the Greek, “Zugon” meaning paired or union and “Phyllon” meaning leaf. Due to this characteristics they are commonly referred to as the “twinleaf” family.

The Zygophyllaceae is a widely-distributed family of heterogeneous plants consisting of approximately 250 species divided into 30 genera. Their diagnostic characters are listed in Table 1-6. They are very drought resistant and occur in arid to semi-arid regions in the subtropics and tropics worldwide (84, 93, 202). Several authors describe the Zygophyllaceae as perennial shrubs, herbs or in some instances trees. In Figure 1-18 the distribution of *Zygophyllum sensu lato* is shown.

Table 1-6: The typical botanical diagnostic characteristics of the family Zygophyllaceae.

Morphological entity	Typical characteristics
Leaves	Opposite, simple, petiolate, bifoliate, trifoliate or pinnate. Leaflets flat, fleshy, sometimes terete or stipulate.
Flowers	Usually solitary or in cymes, mostly bisexual, regular or rarely zygomorphic; calyx and petals usually free 3-5, stamens usually twice as many as petals, in 1-2 whorls, hypogynous and filaments free, usually terete and often with appendages at the base; ovary superior on an annular disc, furrowed, angled or winged, usually 3-5-locular with axile placentation; 2 to many pendulous ovules in each locule; style usually simple; stigma simple.
Fruit	Loculicidal capsule, a schizocarp dividing septicidally in mericarps which can be winged, tuberculate or spinescent or rarely a berry or drupe
Seeds	With or without endosperm, mucus producing or not; embryo straight or slightly curved.

The first *Zygophyllum* species, the Syrian bean caper (*Zygophyllum fabago*), was described by the Swedish naturalist Linneaus in 1753 (102). It was only in 1814, however, that Zygophyllaceae was classified as a distinct taxon, Zygophylleae, by Robert Brown (24).

After several additions by De Candolle (1824), Endlicher (1841), Lindley (1853), as well as Bentham and Hooker (1862), revisions were made to accommodate various geographical areas then known to contain Zygophyllaceae species (18, 40, 55, 101). In Table 1-7 is short list of notable work on the family during the latter half of the 19th century and the early 20th century.

A detailed classification of Zygophyllaceae, which contains 7 divisions of the 25 known genera within that family, was proposed by Engler in 1931, namely Tetradiclidoideae, Augeoideae, Zygophylloideae, Peganoideae, Chitinioidae, Nitrarioidae and Balanitoideae (57). Although the above-mentioned subfamily Zygophylloideae, as described by Engler, which also contains *Zygophyllum* (17 of the 25 genera), is also recognized by the later author El Hadidi, the

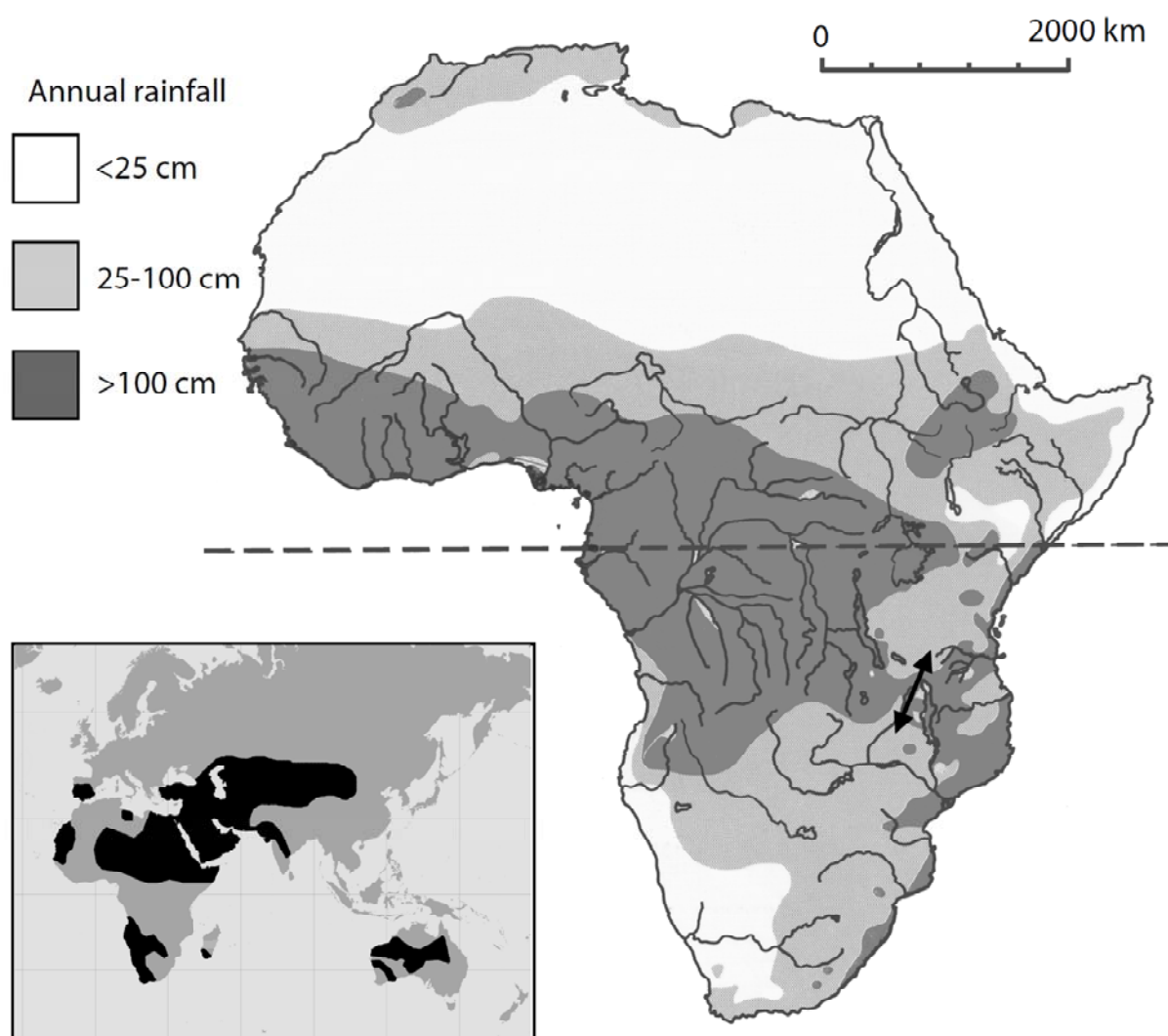


Figure 1-18: The worldwide distribution of the currently recognized genus *Zygophyllum*. Accompanying the distribution map is an annual rainfall map of the African continent showing the close association with lower rainfall areas on the continent. The double-sided arrow also indicates the possible migration route across the African arid corridor across which migration of species occurred (16).

Table 1-7: Below is list of important studies on Zygophyllaceae in the second half of the 19th centry and the early 20th century.

Publication	Year	Author
<i>Flora Capensis</i>	1860	Sonder
<i>Flora Australensis</i>	1863	Bentham, Mueller
<i>Flora Orientalis</i>	1867	Boisser
<i>Plantae lorentzianae</i>	1874	Grisebach
<i>Symbolae ad floram argentinam</i>	1879	Grisebach
<i>North American Flora</i>	1910	Vail, Rydberg

systematic classification of these subfamilies is still unresolved. Investigations into morphology, cytology, palynology and the biochemistry of the family all point to significant morphological diversity within the family (50, 52, 53).

Engler (1931) separated the subfamily Zygophylloideae into two tribes, namely Tribuleae and Zygophylleae. El Hadidi recognized Tribuleae as a distinct family (51). Engler further subdivided Zygophylleae into two subtribes, namely Zygophyllinae and Fagoniinae (57).

Dyer in 1975 described approximately 100 species as belonging to the genus *Zygophyllum* in Africa and Australia (48). El Hadidi in 1975 reduced this number to 80 species from Africa, Australia and Asia (50). Combining the species recognized by these authors with the species recognized on the Index Kewensis there are around 70 species in north-east Africa, the Middle East and Asia. There are 22 species in Australia and 35 species in south-western Africa. This brings the total number of species in the genus to around 150 (207).

Sonder (1860) described and classified 25 South African species known at the time (192). Although he recognized and quoted Endlicher (1841) he did not use Endlicher's division of *Zygophyllum* into two groups based on fruit dehiscence (55). Sonder mainly focused on leaf characteristics in combination with floral characteristics.

Although Engler's treatment of *Zygophyllum* was very comprehensive there was no proper basis for the sections he constructed. He also failed to use the most competent subdivisions of Endlicher based fruit dehiscence (57).

Van Huyssteen (1937) had a very comprehensive classification system that contained all known *Zygophyllum* species (206). She did, in contrast to Sonder and Engler, recognize the importance of both fruit dehiscence and floral morphology. She considered the leaf characteristics and habitat to be of lesser importance in her classification system. She reorganized many of Engler's 17 sections and described several new sections. Her classification system was a good foundation for subsequent studies. There are, however, some shortcomings. In spite of describing six previously unknown species she provided no descriptions of the 95 species listed by her and gave inadequate references for the nomenclature used. Only a single key was provided dealing with all the taxa irrespective of their geographical distribution. Van Huyssteen also did no fieldwork.

Schreiber (1963) focused mainly on 18 *Zygophyllum* species known from Namibia (175). Her work also included typification of names, as well as taxonomic nomenclature and distribution maps of all species listed. She, however, also did no field work and provided no descriptions of the alphabetically listed species of *Zygophyllum* (192). A list of notable authors and their

classification of Zygophyllaceae, based on morphological characteristics, up until 1992 is summarized in Table 1-8.

Table 1-8: The taxonomic treatments of the Zygophyllaceae from 1931-1992 (179).

Name	Order	Family
Engler, 1931	Geraniales	Zygophyllaceae (incl. <i>Balanites</i>)
Willis, 1931	Geraniales	Zygophyllaceae (incl. <i>Balanites</i>)
Lawrence, 1960	Geraniales	Zygophyllaceae (incl. <i>Balanites</i>)
Scholz, 1964 (Engler's <i>Syllabus</i>)	Geraniales	Zygophyllaceae (incl. <i>Balanites</i>)
Cronquist, 1968		Zygophyllaceae (<i>Balanites</i> in Simaroubaceae)
Takhtajan, 1969	Geraniales (but connected to Rurales)	Zygophyllaceae Peganaceae Nitrariaceae Balanitaceae
Hutchinson, 1973	Malpighiales (derived from Tiliales)	Zygophyllaceae Balanitaceae
El Hadidi, 1975		Zygophyllaceae Nitrariaceae Tribulaceae Balanitaceae (<i>Tetradiclis</i> & <i>Peganum</i> to be excluded)
Heywood, 1978	Sapindales } complex Geraniales } Polygalales }	Zygophyllaceae (incl. <i>Balanites</i>)
Dahlgren, 1980	Geraniales	Zygophyllaceae Nitrariaceae Peganaceae Balanitaceae
Cronquist, 1981, 1988	Sapindales (possibly with Balanitaceae/Nitrariaceae/Peganaceae as satellite families)	Zygophyllaceae
Takhtajan, 1980, 1983	Rurales	Zygophyllaceae (incl. Peganaceae) (<i>Tetradiclis</i> to be excluded) Nitrariaceae Balanitaceae
Takhtajan, 1986	Rurales	Zygophyllaceae Nitrariaceae Balanitaceae Peganaceae
Thorne, 1992	Linales	Tetradiclidaceae Zygophyllaceae Balanitaceae

In her doctoral thesis on southern African *Zygophyllum* Van Zyl (2000) recognized 54 species of which 17 were new (175). Her study was based on an extensive investigation into the morphological characteristics of the species. Based on her morphological characteristics she divided the species into two main groupings based on dehiscence of the fruit, seed attachments and spiral threads in the seed mucilage. These divisions are subgenus *Agrophyllum* and subgenus *Zygophyllum*. The most crucial changes that were made to Van Huyssteen's classification system was moving § *Grandifolia* from subgenus *Zygophyllum* to subgenus *Agrophyllum* and moving § *Morgsana* from subgenus *Agrophyllum* to subgenus *Zygophyllum* (see Table 1-9).

Table 1-9: The main changes that Van Zyl made to the classification system of Van Huyssteen (17).

Van Huyssteen (1937)	Van Zyl (2000)
Subgenus <i>Agrophyllum</i> Endl.	
§ <i>Alata</i> Huysst. subsection	6 species
§ <i>Alata</i> subsections	3 species (3 species reduced to synonymy)
§ <i>Morgsana</i> Huysst.	
§ <i>Bipartita</i> Huysst.	10 species including
§ <i>Cinerea</i> Huysst.	<i>Z. simplex</i> L.
	<i>Z. cinereum</i> Schinz =
	<i>Z. longicapsulare</i> Schinz
	§ <i>Annua</i> Engl.
	§ <i>Bipartita</i>
	§ <i>Cinerea</i>
	§ <i>Grandifolia</i> Engl.
	§ <i>Prismatica</i> Van Zyl
	<i>Z. simplex</i> L. (transferred from § <i>Alata</i>) and 2 new species
	10 species (3 species reduced to synonymy, <i>Z. simplex</i> L. (transferred to § <i>Alata</i>), <i>Z. chrysopteron</i> Retief and 3 new species
	<i>Z. giessii</i> Merxm. & A.Schreib., <i>Z. longicapsulare</i> Schinz
	<i>Z. stapffii</i> Schinz
	<i>Z. prismatocarpum</i> Sond. (transferred from section <i>Bipartita</i>) and 2 new species
Subgenus <i>Zygophyllotypus</i> (referred to as subgenus <i>Zygophyllum</i> by Van Zyl (2000))	
§ <i>Capensia</i> Engl.	24 species
§ <i>Paradoxa</i> Huysst.	<i>Z. paradoxum</i> Schinz, <i>Z. cordifolium</i> L.f.,
§ <i>Grandifolia</i>	<i>Z. orbiculatum</i> Welw. ex Oliv.
	<i>Z. stapffii</i> Schinz
	§ <i>Capensia</i>
	§ <i>Morgsana</i> (Huysst.) Van Zyl
	§ <i>Paradoxa</i>
	29 species
	<i>Z. morgsana</i> L.
	<i>Z. cordifolium</i> L.f., <i>Z. fusiforme</i> Van Zyl ined.,
	<i>Z. orbiculatum</i> Welw. ex Oliv.

Below, in Table 1-10 is an table of the subgenera of the southern African members of the genus *Zygophyllum* as proposed by Van Zyl (207).

Table 1-10: A sectional classification of the subgenus *Agrophyllum* and subgenus *Zygophyllum* in southern Africa, as proposed by Van Zyl (207).

Subgenus <i>Agrophyllum</i>	§ <i>Alata</i>	<i>Z. longistipulatum</i>	Subgenus <i>Zygophyllum</i>	§ <i>Capensia</i>	<i>Z. porphyrocaule</i>
		<i>Z. microcarpum</i>			<i>Z. incrustatum</i>
		<i>Z. rigidum</i>			<i>Z. lichtensteinianum</i>
	§ <i>Annua</i>	<i>Z. simplex</i>			<i>Z. cuneifolium</i>
		<i>Z. inflatum</i>			<i>Z. teretifolium</i>
		<i>Z. spongiosum</i>			<i>Z. botulifolium</i>
	§ <i>Bipartita</i>	<i>Z. clavatum</i>			<i>Z. hirticaule</i>
		<i>Z. decumbens</i>			<i>Z. schreiberanum</i>
		<i>Z. applanatum</i>			<i>Z. cretaceum</i>
		<i>Z. chrysopteron</i>			<i>Z. leucocladum</i>
		<i>Z. retrofractum</i>			<i>Z. calcicola</i>
		<i>Z. segmentatum</i>			<i>Z. divaricatum</i>
		<i>Z. turbinatum</i>			<i>Z. flexuosum</i>
		<i>Z. cylindrifolium</i>			<i>Z. maculatum</i>
		<i>Z. tenue</i>			<i>Z. aff. maritimum</i>
	§ <i>Cinerea</i>	<i>Z. giessii</i>			<i>Z. maritimum</i>
		<i>Z. longicapsulare</i>			<i>Z. namaquanum</i>
	§ <i>Prismatica</i>	<i>Z. patenticaule</i>			<i>Z. rogersii</i>
		<i>Z. prismatocarpum</i>			<i>Z. spinosum</i>
		<i>Z. pterocaula</i>			<i>Z. foetidum</i>
	§ <i>Grandifolia</i>	<i>Z. stapffii</i>			<i>Z. leptopetalum</i>
					<i>Z. debile</i>
					<i>Z. pubescens</i>
					<i>Z. swartbergense</i>
					<i>Z. fulvum</i>
					<i>Z. fuscatum</i>
					<i>Z. pygmaeum</i>
					<i>Z. sessilifolium</i>
					<i>Z. spitskopense</i>
				§ <i>Morgsana</i>	<i>Z. morgsana</i>
				§ <i>Paradoxa</i>	<i>Z. cordifolium</i>
					<i>Z. fusiforme</i>
					<i>Z. orbiculatum</i>

1.6.2 Molecular systematics of *Zygophyllum*

The overall systematic placement of the family Zygophyllaceae in the Angiosperm phylogeny has briefly been mentioned previously in section on molecular systematics. Several studies since

the early 1990's have considerably improved the resolution of the Angiosperm phylogeny and have placed the order Zygophyllales, at the base of the fabid clade (23, 30, 31, 65, 68, 86, 107, 126, 168, 173, 188, 189, 191, 210).

1.6.2.1 Molecular systematics focussing on the family Zygophyllaceae

The first molecular phylogenetic study involving species belonging to the family Zygophyllaceae was conducted in 1993 (32). This study utilized *rbcL* gene sequences and contained a specimen belonging to the New World Zygophyllaceae, namely, *Guaiacum*. In this study Zygophyllaceae appeared sister to the family Krameriaceae, a family of hemi-parasitic plants from the New World.

The first genetic study that focused on the family Zygophyllaceae including *Zygophyllum*, was by Sheahan and Chase in 1996 (179). The authors utilized the plastid gene *rbcL* of 20 Zygophyllaceae species and combined it with anatomical, as well as morphological data. This study also indicated that the genera *Malacocarpus*, *Nitraria* and *Peganum* were located in the order Sapindales and were not closely related to Zygophyllaceae as previously believed. In Figure 1-19 indicates a phylogenetic tree of several species including those of species previously thought to be from the family Zygophyllaceae.

Based on the phylogeny they proposed a recircumscription of Zygophyllaceae *sensu stricto* into five separate subfamilies, namely Morkillioideae, Tribuloideae, Seetzenioideae, Larreoideae and Zygophylloideae (see Figure 1-20).

In 2000 the same authors published an article in which they examined 36 species from the family Zygophyllaceae based on *rbcL*, as well as on the non-coding *trnL*-F spacer sequence data (180). The separate, as well as the combined analyses confirmed their initial classification of the five sub-families. Zygophylloideae was also further subdivided into five clades, namely a *Fagonia*, *Zygophyllum robecchii* and *Zygophyllum hildebrandtii* clade, a monotypic *Augea* clade, a southern African and Australian *Zygophyllum* clade, a Middle Eastern/Asian *Zygophyllum* clade and a *Agrophyllum* clade (see Figure 1-21). The connecting branches though were poorly supported and they recommended further examination.

1.6.2.2 Molecular phylogenetics focusing on the genus Zygophyllum

In 2003, Beier *et al.* focused on morphological and genetic information of the *trnL* intron region of 43 species of Zygophylloideae species encompassing numerous geographical and morphological variations in the subfamily (13). They found that the monotypic genera, *Fagonia*, *Augea* and *Tetraena* were embedded within *Zygophyllum*. Based on this phylogeny they

proposed a new taxonomic classification of Zygophylloideae into six genera based on morphological and monophyletic distinguishable entities (see Figure 1-22). The proposed genera were *Melocarpum*, *Fagonia*, *Zygophyllum*, *Augea*, *Tetraena* and *Roepera*.

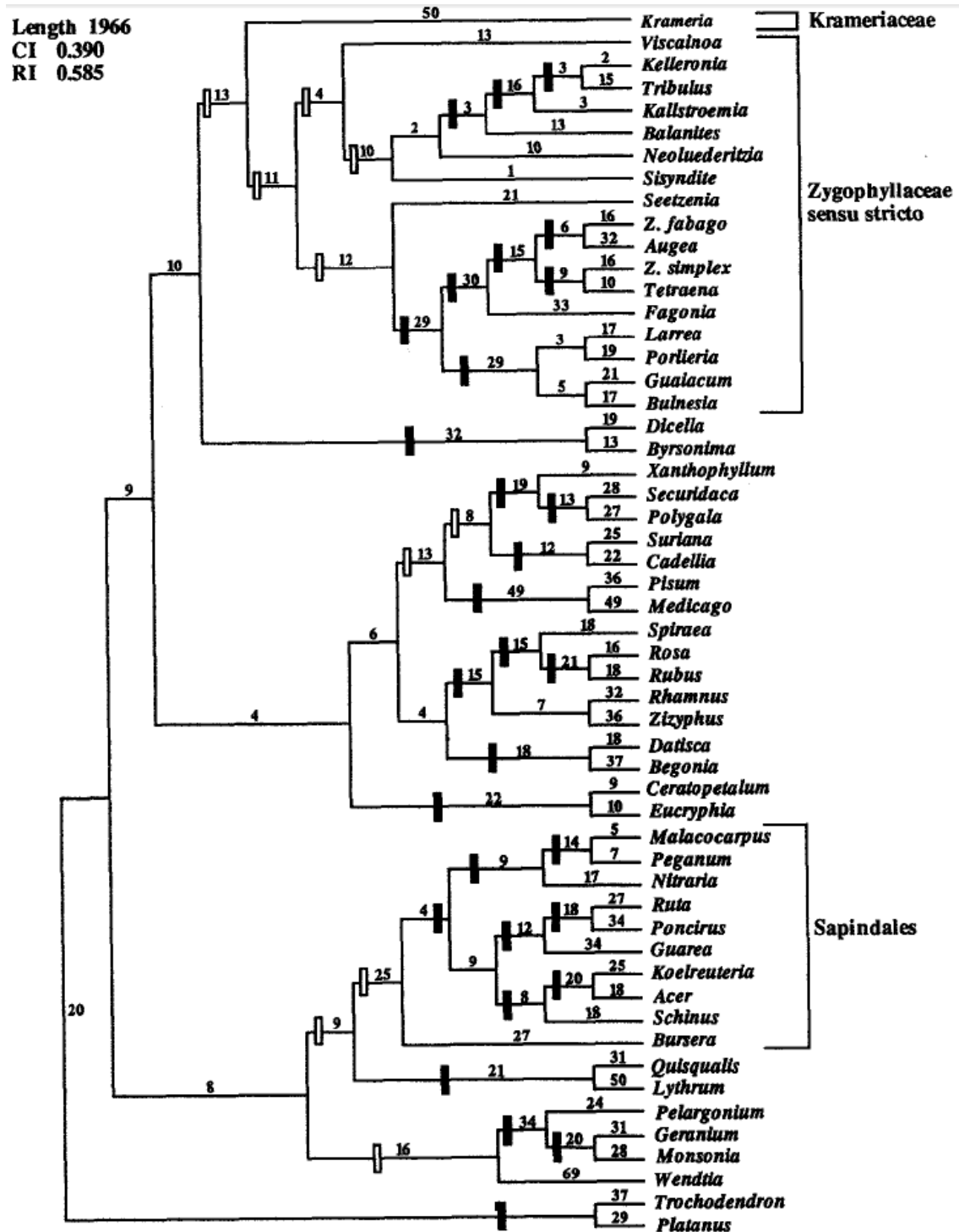


Figure 1-19: A phylogenetic tree based on *rbcL* data indicating the position of some of the genera previously associated with Zygophyllaceae, i.e. *Malocarpus*, *Nitraria* and *Peganum*. Zygophyllaceae also appears sister to Krameriaceae (179).

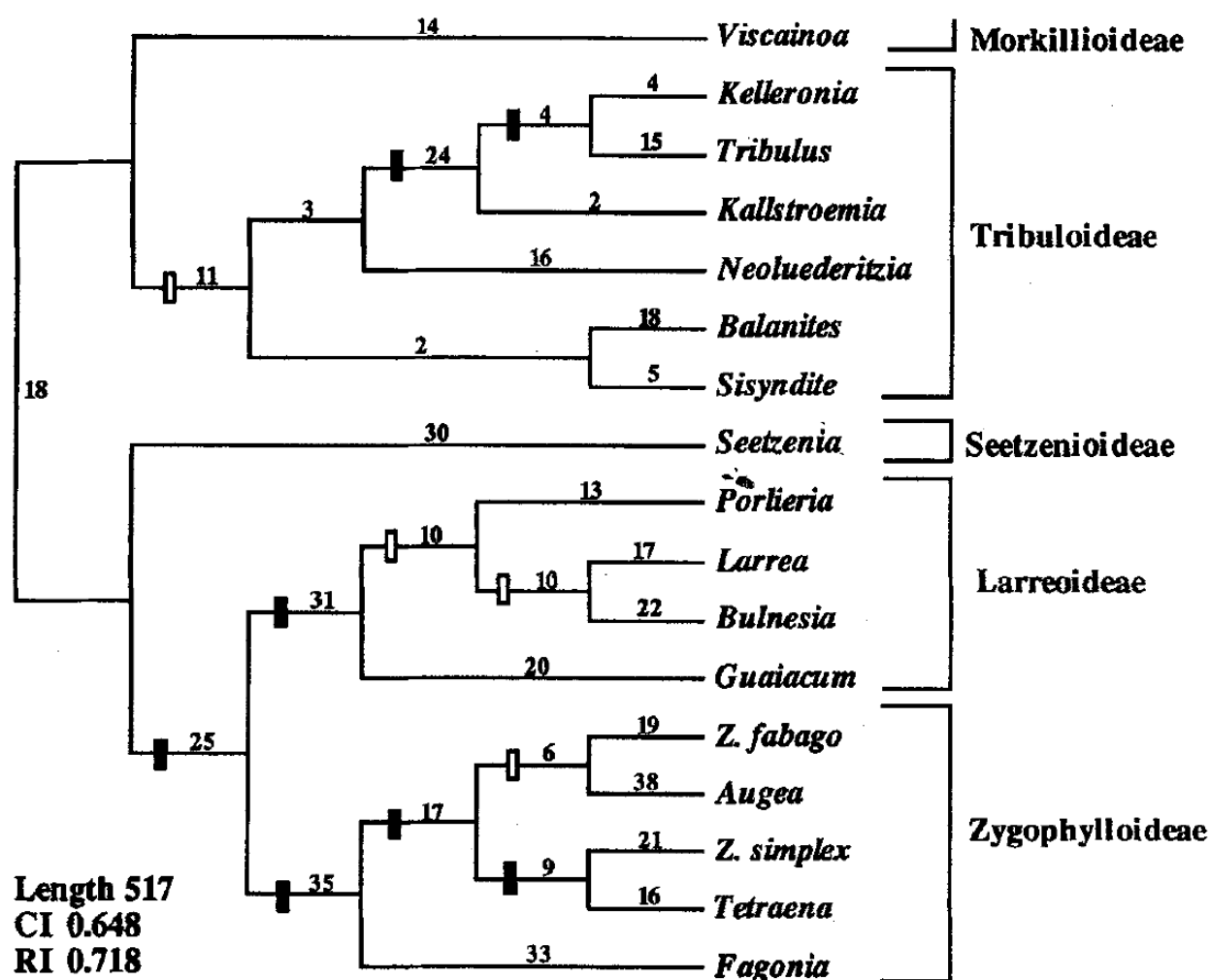


Figure 1-20: A phylogenetic tree retrieved from combined *rbcL* and morphological character data indicating the proposed recircumscription of the family Zygophyllaceae (179).

In 2004 the same authors published a publication in which they focused their analysis on 32 of 34 known *Fagonia* species based on the *trnL* intron and the nuclear marker ITS (internal transcribed spacer) (14). Using both parsimony and Bayesian model averaging they retrieved a tree in which all, except one of the Old World species, formed a weakly supported clade. All New World species excluding one taxon formed a well supported sister clade to the Old World clade. The two outlier species formed a well supported clade apart from the Old and New World clades.

The most recent study on the systematic classification of *Zygophyllum* was published in 2008 by Bellstedt *et al.* (17). They expanded on the *rbcL* and *trnL*-F data to include 53 of 55 southern African species and retrieved trees of single as well as combined *rbcL* and *trnL*-F data. The study on *trnL*-F by Bellstedt *et al.* in 2008 included 73 taxa from subgenus *Zygophyllum*. Eleven closely related outgroup taxa were also analysed, the furthest removed being *Tribulus*. The study on the *rbcL* region, included 42 taxa from the subgenus *Zygophyllum*. The same 11 outgroup species were used as in the *trnL*-F phylogenetic tree. Their results supported previous

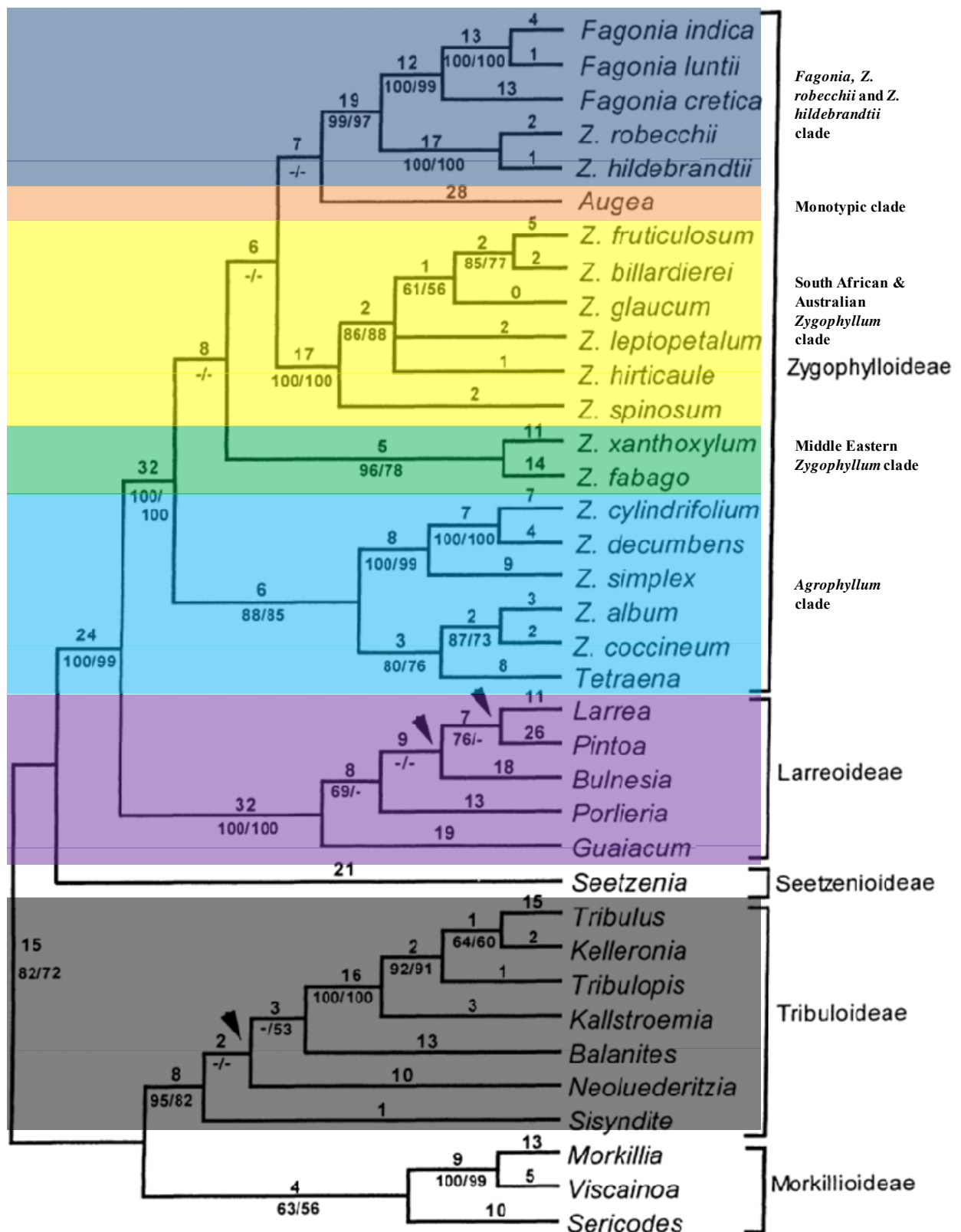


Figure 1-21: A phylogenetic tree based on combined *rbcL* and *trnLF* sequence data as determined by Sheehan and Chase, in 2000 (180).

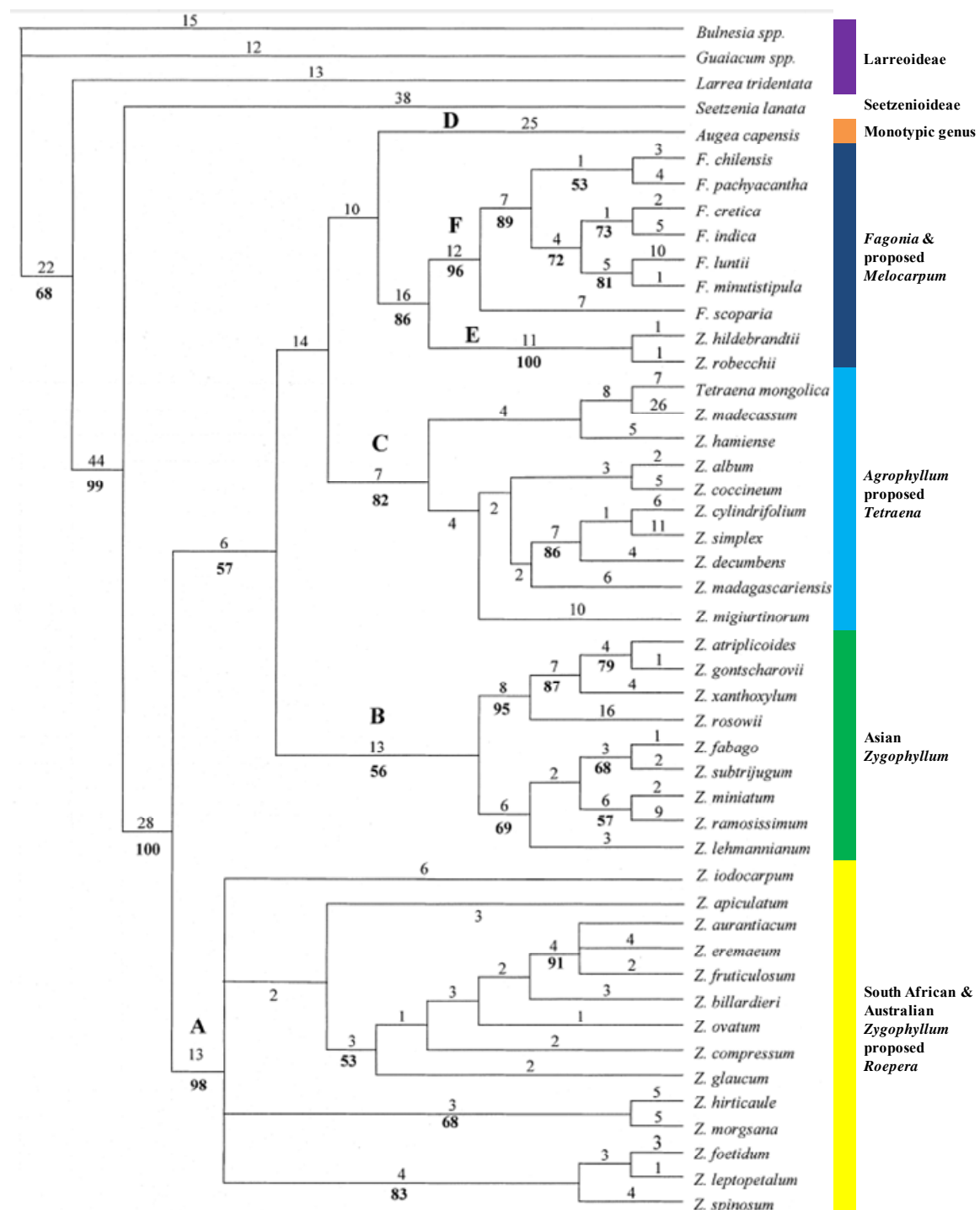


Figure 1-22: A phylogenetic tree indicating the taxonomic changes proposed by Beier *et al.* (13)

classifications of the genus into the subgenera *Zygophyllum* and *Agrophyllum* (see Figure 1-23). They analyzed the character evolution of seed mucilage, capsule dehiscence which also supported the division of the southern African species into these subgenera. This could however not be applied to species that occur elsewhere. Additional morphological characters were also investigated and unique combinations of these characters rather than unique characters were of importance in the systematic elucidation. Their study suggested repeated radiations from the horn of Africa to Asia and southern Africa and *vice versa*. These radiations led to the current

distribution of the subfamily Zygophylloideae. They also accepted some of the recent changes in the taxonomy by Beier *et al.* (7), but they also concluded that taxonomic changes may have been premature i.e. that *Tetraena* and *Roepera* could still be accommodated in a larger monophyletic *Zygophyllum*.

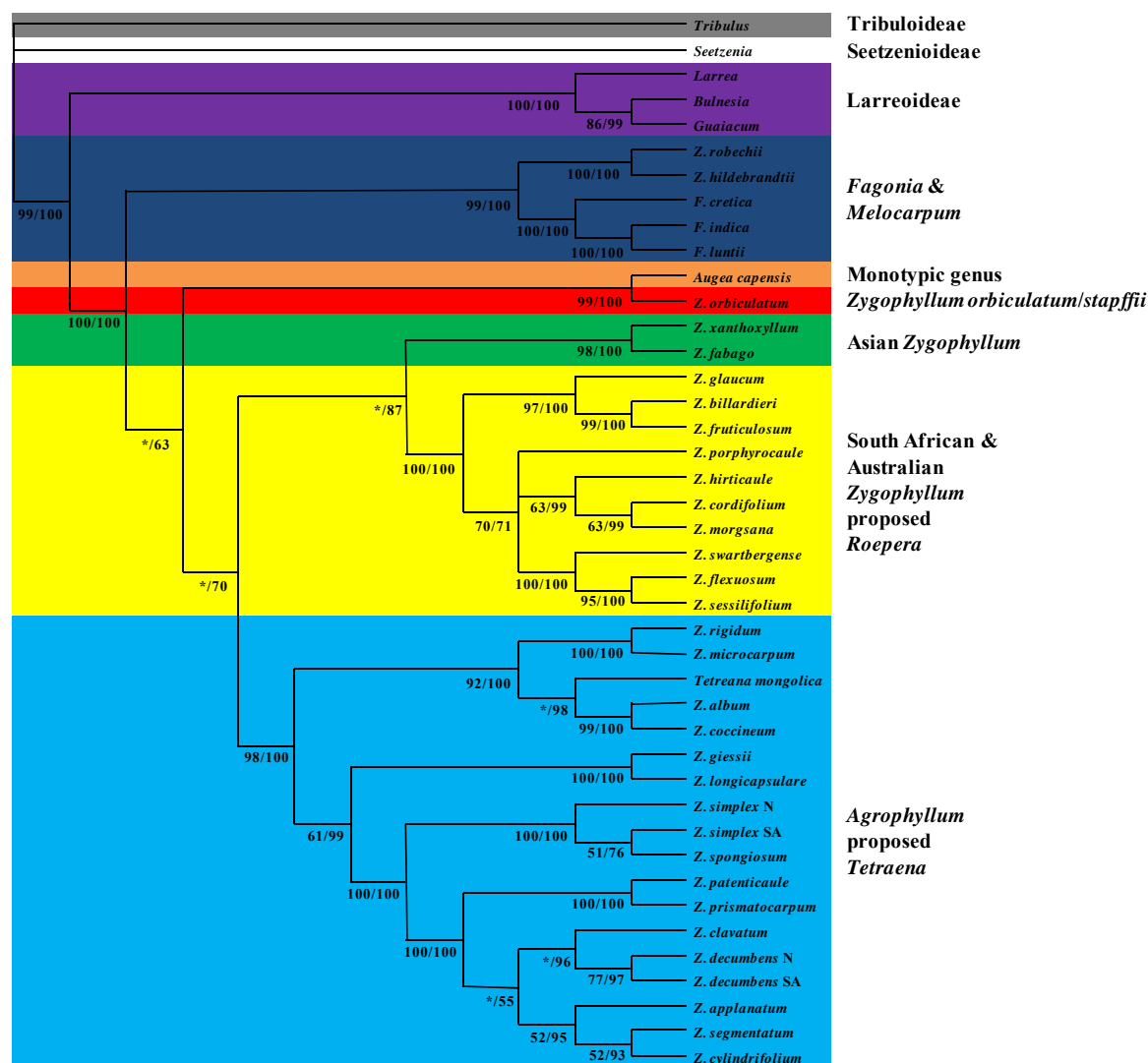


Figure 1-23: A phylogenetic tree of members of subfamily Zygophylloideae based on combined *rbcL* and *trnLF* data (17).

1.6.2.3 Carbon fixation strategies in Zygophyllaceae

Within the context of this thesis several taxa within the family Zygophyllaceae have been identified to have C_4 photosynthesis, i.e. some of the *Tribulus* and *Kallstroemia* species, as well as *Zygophyllum simplex*, *Zygophyllum inflatum* and *Zygophyllum spongiosum* in the § *Annua* in the subgenus *Agrophyllum* (93, 170, 207). It has also been reported that *Zygophyllum cordifolium* in the § *Paradoxa* is a mixed C_3 -CAM taxon within the southern African subgenus *Zygophyllum* (112, 207).

1.7 Objectives and thesis structure

The objectives of this study were therefore:

- To determine, using the ITS region as a phylogenetic marker region, if the two species *Zygophyllum orbiculatum* and *Zygophyllum stapffii* are conspecific.
- To determine, using the ITS region as a phylogenetic marker region, the phylogenetic relationships of the genera and groups within the subfamily Zygophylloideae proposed by Beier *et al.* in 2003 (13).
- To sequence the whole chloroplast genomes of representative taxa from the genera and groups for phylogenetic inference within the subfamily Zygophylloideae to infer their phylogenetic relationships, should the ITS region sequences fail to resolve them.

With a view to achieving these objectives, the comparison of *Zygophyllum orbiculatum* and *Zygophyllum stapffii* using the ITS region as a phylogenetic marker, was described in chapter two. Additionally an attempt was made to resolve the phylogenetic relationships of the genera and groups within the subfamily Zygophylloideae using the ITS region sequence data in this chapter.

In chapter three, a NGS approach is described to resolve the phylogenetic relationships of the genera and groups within the subfamily Zygophylloideae. This was achieved by selecting a single representative from each of the genera and groups within the subfamily Zygophylloideae, as well as from selected outgroups and sequencing most of the chloroplast coding genes and some non-coding chloroplast regions, as well as the complete nuclear ITS cassette sequence data. This was followed by subsequent phylogenetic inference from the aligned sequence data.

The conclusions of the ITS region phylogenies and the phylogenomic study, as well as future perspectives are presented in chapter four.

The thesis concludes with a reference list, followed by appendices.

2 The *Zygophyllum orbiculatum* and *Zygophyllum stapffii* conundrum and the phylogenetic relationships of *Zygophyllum* based on the nuclear ITS region.

2.1 Introduction

Daniel Oliver, in 1868, described *Zygophyllum orbiculatum* which was collected by Friederich Welwitsch (25 February 1806 - 20 October 1872), during his expedition to Angola on behalf of the Portuguese government from 1853-1860 (140). Hans Schinz (6 December 1858 - 30 October 1941) was a Swiss-born explorer and botanist, who described *Zygophyllum stapffii* from German South West Africa in 1888 (5). Due to the very different political environments of the colonies, as well as the mother countries of these colonies, *Zygophyllum orbiculatum* and *Zygophyllum stapffii* were never directly compared with each other in these early years.

Zygophyllum stapffii was placed in a monotypic subgenus § *Grandifolia* by Engler in 1915 (56). This placement was upheld by Van Huyssteen in 1937, placing the § *Grandifolia* in subgenus *Zygophyllum* (206). Van Zyl, in 2000, disputed this placement as she observed druse crystals in the mesophyll of the leaves. She also observed that the mucilage of the seed testa was structured with short spiral inclusions that unravelled at the apex. Both these anatomical characteristics are more closely associated with members of the subgenus *Agrophyllum* than subgenus *Zygophyllum*. Consequently she placed *Zygophyllum stapffii* in subgenus *Agrophyllum* (207).

Zygophyllum orbiculatum was placed in § *Paradoxa* by Schreiber, alongside *Zygophyllum cordifolium*, in 1963 (175). A thorough morphological investigation of *Zygophyllum orbiculatum*, by Van Zyl was not possible due to the poor condition of the original herbarium specimens made by Welwitsch, which had degraded over more than a century and contained no usable flowers or fruit capsules. However, in her thesis she mentions under the description of *Zygophyllum orbiculatum* that the placement close to *Zygophyllum cordifolium* was “dubious”, but due to a lack of fieldwork and lacking floral details she consequently retained the previous placement of this species within § *Paradoxa*, in the subgenus *Zygophyllum* (207).

Observations made by Dr Patricia Craven, an authority on the Namibian flora, during an expedition into southern Angola in 2007, resulted in the first comprehensive comparison of the two species since their original descriptions well over a century earlier. Two specimens of *Zygophyllum orbiculatum* collected by Craven and one specimen of *Zygophyllum stapffii* by Mannheimer, were analysed by Bellstedt *et al.*, 2008 (17) and were found to possess identical *rbcL* and *trnLF* sequences. Based on this, they concluded that these species were likely to be

conspecific, but this was based on three samples only. During the SANBI/Angolan joint biodiversity assessment expedition in 2009, Bellstedt was able to collect and photograph *Zygophyllum orbiculatum* in its extremely arid and northern distribution near the town of Namibe in Angola and *Zygophyllum stapffii* at its extreme southern distribution just south of Swakopmund in Namibia.

Below in Figure 2-1 and Figure 2-2 are images of *Zygophyllum orbiculatum* and *Zygophyllum stapffii*. The first image (Figure 2-1) shows *Zygophyllum orbiculatum*, which occurs in Angola. The second image (Figure 2-2) shows *Zygophyllum stapffii* from Namibia. At first glance they appear to be identical, i.e. the colour of the leaves, the shape and size of the plant and the leaves, as well as the similarity of the flowers. Upon closer inspection, however, the unifoliolate leaves of *Zygophyllum orbiculatum* as opposed to the bifoliolate leaves of *Zygophyllum stapffii* distinguish the two. In *Zygophyllum* it has been observed that in some species during times of drought one of the bifoliolate leaves is discarded leaving only unifoliolate leaves (17, 79). Closer inspection of the *Zygophyllum orbiculatum* leaf structure, however, indicates no evidence of leaf abscission and thus this hypothesis does not account for this morphological characteristic of *Zygophyllum orbiculatum*. It appears that the plant grows unifoliolate leaves rather than discarding one of an adult bifoliolate pair, although a microscopic analysis of leaf primordia was not conducted.

The first goal of this study was to evaluate if *Zygophyllum orbiculatum* and *Zygophyllum stapffii* are conspecific, meaning that they only represent one species using a molecular systematic approach and with a more extensive sampling. This was investigated using the highly variable nuclear ITS region. Due to the high mutation rates of these regions they are used as a DNA barcode to identify species, especially including Fungi and plants (69, 174, 219).

The second goal of this study was to attempt to resolve the phylogenetic relationships of the major groupings within the subfamily Zygophylloideae using the nuclear ITS region as a phylogenetic marker. None of the previous molecular systematic studies, all of which were based on chloroplast genetic markers, could with a high degree of certainty, resolve these phylogenetic relationships (13, 17). Furthermore, should the ITS region not be suitable for this, could the ITS sequence data be combined with the previously sequenced chloroplast marker sequence data in order to resolve the phylogenetic relationships of the major groupings in the subfamily?



Figure 2-1: *Zygophyllum orbiculatum* inland from the town of Namibe in the arid regions of Angola. Notice the unifoliate leaves as opposed to the bifoliate leaves of *Zygophyllum stapffii*. The photograph was taken by Dirk Bellstedt in 2009 on the SANBI/Angolan joint expedition in Angola.



Figure 2-2: *Zygophyllum stapffii* at the southern most limit of its distribution just south of Swakopmund, Namibia. Bifoliate leaves are visible and not unifoliate leaves as is the case with *Zygophyllum orbiculatum*. The photograph was taken by Dirk Bellstedt in 2009 after the SANBI/Angolan joint expedition.

2.2 Materials and Methods

2.2.1 Taxon sampling

Below in Table 2-1 is a list of all taxa of which the ITS region was sequenced. In cases where the ITS sequence information was retrieved from Genbank, an accession number is given. The rest of the samples were collected by several botanists including myself. The table also indicates the DNA samples of some of *Zygophyllum* species from the Middle East and Asia obtained from Kew, which were supplied by Dr Felix Forest.

In the context of this study, it is important to mention that the *Zygophyllum orbiculatum* specimens, *Zygophyllum orbiculatum* (1) DUB 1202, *Zygophyllum orbiculatum* (2) DUB 1213, *Zygophyllum orbiculatum* (5) Craven 5096 and *Zygophyllum orbiculatum* (6) Craven 5101 were collected in Angola. *Zygophyllum stapffii* DUB 1222 sample was collected at the southern most limit of the range of the species at Swakopmund in Namibia. The two *Zygophyllum orbiculatum* specimens *Zygophyllum orbiculatum* (3) HK 2753, *Zygophyllum orbiculatum* (4) HK 2757, although collected in Namibia were determined not to be *Zygophyllum stapffii* as could be expected, because they occur in Namibia, but as *Zygophyllum orbiculatum* based on the fact that they possessed unifoliolate leaves. These specimens were in fact the first descriptions of *Zygophyllum orbiculatum* in Namibia. Previously these specimens would have been identified as *Zygophyllum stapffii* because of the fact that they occur in Namibia.

2.2.2 DNA extraction

For most of the samples silica gel dried leaf material was used for DNA extraction. The extraction method used was a modified CTAB extraction published by Doyle and Doyle in 1987. Some samples were already pre-extracted as they were used by Bellstedt *et al.* in 2008 in previous studies (17). For the other samples silica dried leaf material was used. The samples obtained from Kew were pre-extracted and the DNA was used directly.

2.2.3 Amplification of the ITS gene region

Amplification of the ITS region was achieved by using several primers as listed in Table 2-2. For most of the species one set of primers was sufficient to successfully amplify the ITS region, namely AB101 and AB102 (214). For those species in which these primers did not yield amplification products, alternative primers were used to successfully amplify the gene region (158). All PCR reactions were performed in an Applied Biosystems Veriti 96 well Thermal Cycler in 25 µl reactions. Each tube contained 2.5 mM MgCl₂, 1x JMR-455 buffer (Southern

Cross Biotechnology, Cape Town, RSA), 1 U of Super-Therm Taq polymerase (Southern Cross Biotechnology, Cape Town, RSA), 200 µM of each of the dNTP's and 0.5 µM of each of the primers. Amplification profiles were 35 cycles with 1 min denaturation at 94 °C, 1 min annealing at 55 °C, 90 s extension time at 72 °C, followed by a final extension step of 6 min at 72 °C.

Table 2-1: Below is a detailed list of the species that were investigated. The outgroup species, as well as one internal species, *Zygophyllum billardierei*, were retrieved from Genbank and the accession numbers are included in table.

Species	Geographic distribution	Voucher details	Collection locality	ITS Database accession number
<i>Augea capensis</i>	South Africa	Bellstedt (STE)	Calitzdorp	
<i>Bulnesia arborea</i>	New World	Chase 641 (K)		
<i>Fagonia cretica</i>	Canary Islands	Beier 125 (UPS)		AY641623.1
<i>Fagonia indica</i>	United Arab Emirates	Thulin <i>et al.</i> 10024 (UPS)		AY641631.1
<i>Fagonia luntii</i>	Yemen	Thulin <i>et al.</i> 9881 (UPS)		AY641638.1
<i>Fagonia minutistipula</i>	Namibia	Giess and Muller 13952		AY641641.1
<i>Fagonia rangei</i>	South Africa	Leistner 3388		AY641647.1
<i>Guaiacum guatemalense</i>	New World	Chase 640 (K)		
<i>Larrea tridentata</i>	New World	Chase 636 (K)		
<i>Melocarpum hildebrandtii</i>	Somalia	Thulin <i>et al.</i> 9012 (UPS)		AY641615.1
<i>Melocarpum robecchii</i>	Yemen	Thulin <i>et al.</i> 9537 (UPS)		AY641616.1
<i>Seetzenia lanata</i>	South Africa	Bellstedt 938	Koue Bokkeveld, Western Cape	
<i>Tetraena mongolica</i>	Mongolia	Sheahan 1105	Dung Kou, Inner Mongolia	
<i>Tribulus</i>	Namibia	Bellstedt 1333		
<i>Tribulus lanuginosus</i>		Sathishkumar, R. <i>et al.</i>		HM236860.1
<i>Tribulus subramanyamii</i>		Sathishkumar, R. <i>et al.</i>		HM236858.1
<i>Tribulus terrestris</i>		B. B. Simpson 16-VI-00-1	USA: TX, Burnet Co	AY260972.1
<i>Zygophyllum hirticaule</i>	Namibia	Craven 2857		
<i>Zygophyllum album</i>		Thulin <i>et al.</i> 7977 (1697) (K)		
<i>Zygophyllum applanatum</i>	Namibia	Bellstedt 870 (STE)	Rosh Pinah	
<i>Zygophyllum billardierei</i> (1)	Australia	SR.417 (Adelaide BG.)	Adelaide	AY641613.1
<i>Zygophyllum billardierei</i> (2)	Australia	2287		
<i>Zygophyllum clavatum</i>	Namibia	Bellstedt 878 (STE)	Lüderitz peninsula	
<i>Zygophyllum cordifolium</i>	South Africa	Marais 446 (STE)	Koekenaap	
<i>Zygophyllum decumbens</i>	Namibia	Bellstedt 851 (STE)	Noordoewer	
<i>Zygophyllum fabago</i>	France	Chase 516 (K)	Lyons Bot. Gard.	
<i>Zygophyllum flexuosum</i>	South Africa	Bellstedt 794 (STE)	Langebaan, WC	
<i>Zygophyllum fruticosum</i>	W. Australia	Chase 2203 (K)	Port Gregory, North of Perth	
<i>Zygophyllum giessii</i>	Namibia	Bellstedt 1323		
<i>Zygophyllum giessii</i>	Namibia	Bellstedt 1326		
<i>Zygophyllum gontscharowari</i>	Tadzikistan	Ashnova 10191 (K)		
<i>Zygophyllum hamiense</i>	Yemen	Thulin and Berer 9840 (UPS)		
<i>Zygophyllum lehmanianum</i>	Turkmenistan	10191 (K)		
<i>Zygophyllum longicapsulare</i>	Namibia	Bellstedt 879 (STE)	Lüderitz peninsula	
<i>Zygophyllum madagascariensis</i> (1)	Madagascar	Bellstedt 1239		
<i>Zygophyllum madagascariensis</i> (2)	Madagascar	Bellstedt 1247		
<i>Zygophyllum madecassum</i> (1)	Madagascar	Bellstedt 1243		
<i>Zygophyllum madecassum</i> (2)	Madagascar	Bellstedt 1248		
<i>Zygophyllum microcarpum</i>	Namibia	Bellstedt 1315		
<i>Zygophyllum morgsana</i>	South Africa	Bellstedt 890 (STE)	Steinkopf, NC	
<i>Zygophyllum orbiculatum</i> (1)	S. Angola	Bellstedt 1202 (STE)	Namibe	
<i>Zygophyllum orbiculatum</i> (5)	S. Angola	Craven 5096 (WIND)	30 km west of Foz de Cunene	
<i>Zygophyllum orbiculatum</i> (6)	Angola	Craven 5101 (WIND)	River tributary, South of Tambor	
<i>Zygophyllum orbiculatum</i> (2)	S. Angola	Bellstedt 1213 (STE)	Lake Acoa	
<i>Zygophyllum orbiculatum</i> (4)	Namibia	HK 2757 (WIND)		
<i>Zygophyllum orbiculatum</i> (3)	Namibia	HK 2753 (WIND)		
<i>Zygophyllum patenticaule</i>	Namibia	Bellstedt 868 (STE)	Rosh Pinah	
<i>Zygophyllum prismatocarpum</i>	Namibia	Bellstedt 863 (STE)	Dumusbiriver, near Orangeriver	
<i>Zygophyllum ramosissimum</i>	Kazakhstan	Giamolov 10182 (K)		
<i>Zygophyllum rigidum</i> (1)	Namibia	Bellstedt 852 (STE)	Noordoewer	
<i>Zygophyllum rigidum</i> (2)	Namibia	Bellstedt 855 (STE)	Aussenkehr	
<i>Zygophyllum rosowii</i>	China	Chaney 57	Karakum	
<i>Zygophyllum segmentatum</i>	Namibia	Bellstedt 861 (STE)	Rosh Pinah	
<i>Zygophyllum sessilifolium</i>	South Africa	Marais 434 (STE)	Morreesburg, WC	
<i>Zygophyllum simplex</i> (1)	Namibia	Bellstedt 1226 (STE)	Grünaau	
<i>Zygophyllum simplex</i> (2)	Namibia	Bellstedt 1224 (STE)	Swakopriv mouth	
<i>Zygophyllum simplex</i> (3)	Africa and Asia	El Hadidi, Sheahan 806	Northern Hemisphere	
<i>Zygophyllum spongiosum</i>	Angola	Bellstedt 1214 (STE)	Praia Azul, Benguela	
<i>Zygophyllum stapfii</i>	Namibia	Bellstedt 1222 (STE)	Swakopmund	
<i>Zygophyllum swartbergense</i>	South Africa	Bellstedt 798 (STE)	Swartberg Pass, WC	
<i>Zygophyllum xanthoxylum</i>	China	Chase 1700 (K)		

Amplification was confirmed by agarose gel electrophoresis. PCR products were purified using either the Wizard PCR Prep kit (Promega Corp., Madison, USA) or Exonuclease-I and Shrimp Alkaline Phosphatase (ExoSAP) from Affymetrix (<http://www.affymetrix.com>). Cycle sequencing was used to generate nucleotide sequences of the ITS regions including the 5.8S rRNA gene. The reactions were performed in 10 µl reactions using the BigDye Terminator v3.1 Cycle Sequencing Kit (Applied Biosystems Inc., Foster City, USA) using the primers listed in Table 2-2. The reactions additionally contained approximately 100 ng of DNA, 2 µl 5 buffer

(400 mM Tris-HCl, 10 mM MgCl₂ at pH 9.0), 3.2 pmol primer, 2 µl of Terminator Ready Reaction Mix and water. The cycle sequencing profile was 35 cycles consisting of 10 s at 96 °C, 30 s at 52 °C and 4 min at 60 °C. Excess terminator dye was removed using gel filtration through Centri-Sep 96 Multi-well Filter Plates (Princeton Separation, Adelphia, USA). Sequencing reactions were subsequently analyzed on an ABI 377 sequencer (Applied Biosystems Inc., Foster City, USA).

Table 2-2: A list of all primers that were used to amplify the ITS region.

Primer name	Primer use	Primer sequence (5'-3')
AB101 (Forward)	Amplification & sequencing	ACG AAT TCA TGG TCC GGT GAA GTG TTC G
AB102 (Reverse)	Amplification	TAG AAT TCC CCG GTT CGC TCG CCG TTA C
P16 (Forward)	Amplification & sequencing	TCA CTG AAC CTT ATC ATT TAG AGG A
P17 (Forward)	Amplification & sequencing	CTA CCG ATT GAA TGG TCC GGT GAA
P25 (Reverse)	Amplification	GGG TAG TCC CGC CTG ACC TG
26S-82R (Reverse)	Amplification	TCC CGG TTC GCT CGC CGT TAC TA
2G (Reverse)	Sequencing	GTG ACG CCC AGG CAG ACG T

Sequence chromatograms were viewed and edited in ChromasPro v1.7.6 and subsequently exported into BioEdit for alignment. Initial alignment was done using the ClustalW algorithm in BioEdit, but due to the high variability of the region, the alignments were refined by eye. Once all taxa were incorporated into the alignment matrix, a NEXUS file was generated.

2.2.4 Parsimony analysis

The generated NEXUS file was subsequently used to perform a maximum parsimony phylogenetic analysis using PAUP 4.0b10 (199). All the substitutions were weighted equally. All gaps and indel characters were treated as missing data. Replicates were set at 1 000 with TBR branch swapping and holding 10 trees with the MulTrees option on and Maxtrees set to 10 000. A strict consensus tree was generated from all equally parsimonious trees. Support for the nodes were calculated with 1 000 bootstrap replicates also using TBR branch swapping, but with the MulTrees setting turned off. Nodes with bootstrap values equal or higher than 75% were considered to be strongly supported, between 50-74% as moderately supported and below 50% to not be supported. The Consistency Index (CI) is an indication of the amount of homoplasy in the dataset. It is the fraction of the minimum amount of changes needed for the phylogenetic tree divided by the observed changes. The value is normally negatively correlated with the amount taxa in the tree. The Retention Index (RI) is also an indication of homoplasy, but also indicates how well the synapomorphies or shared characters explain the tree topology.

2.2.5 Maximum likelihood

The analysis for the maximum likelihood phylogenetic tree was conducted in RAxML v. 8.0.24 on the CIPRES Science Gateway website (121). The NEXUS file of the ITS region genetic data matrix that was generated in BioEdit was converted into a PHYLIP file using SequenceMatrix as RAxML requires a PHYLIP file as input. All the settings for the analysis were kept at default settings except for the following settings in the advances options (Configure Bootstrapping) section:

- “Print branch lengths (-k)” was switch on
- “Specify an Explicit Number of Bootstraps +” was switched off
- “Let RAxML halt bootstrapping automatically +” was switched on

The completed analysis produced the resulting phylogenetic tree in Newick format file. This Newick file was converted into a phylogenetic tree online in Tred (<http://www.reelab.net/tred>). From Tred the tree was exported as a PDF. The PDF was converted into an encapsulated postscript (.eps) file in Adobe Professional. The EPS file was opened in Microsoft PowerPoint where final editing was done.

2.2.6 Bayesian inference

The Bayesian inference analyses were also performed on the CIPRES Science Gateway website. For this analysis a minimum of two independent analyses were conducted in parallel. The model of evolution was assessed by using appropriate modelling software packages, e.g. PartitonFinder or mrmodeltest. The analyses were allowed to run over a user-specified number of iterations. If the analyses were allowed a sufficient amount of iterations, the two independent analyses converged on the same tree topology. During the analyses a predetermined user-defined percentage of the generated trees were stored as the analyses converged on a steady-state tree topology. The first ten percent of the stored trees from the independent analyses were discarded as the topologies at the beginning of the analyses were less likely to be accurate. The remaining stored trees from the two independent analyses were combined and a majority-rule consensus tree was generated. Node support of a Bayesian inference tree is indicated as posterior probabilities rather than bootstrap values. These values are indicative of the percentage of all the stored trees that contain the node in question. For a node to be considered supported it has to have a value equal or higher than 95%.

In the case of this ITS region sequence matrix, the best model for the matrix was GTR+G. We conducted two independent analyses and allowed for 10 000 000 generations while storing every

1000th tree. Upon completion the log-likelihood distributions were inspected to ascertain whether they had become stationary, and the values were plotted to see if they fluctuated. We also used the default temperature of 0.2 for heating the chains. Swapping between the chains and the acceptance values of these swaps were monitored and found to be in the acceptable range. The first 10%, i.e. the burnin, of the trees from both analyses was discarded and the remaining trees from both analyses were combined (18002 stored trees) in LogCombiner in the BEAST package. The file containing the combined trees from both analyses was executed in PAUP and a majority rule consensus tree was generated. The majority rule consensus tree was saved as a PICT file and was subsequently edited in Microsoft PowerPoint 2007.

2.3 Results

2.3.1 Molecular data

All the taxa contained within the matrix had usable sequence information. For *Bulnesia arborea*, however, a limited supply of DNA was available from Kew and repeated attempts to amplify the ITS region depleted the available stock. The only usable sequence that was obtained before the stock was depleted was from the internal primer, 2G. Thus within the matrix *Bulnesia arborea* only contains approximately between 50-66% of the sequence information. The taxa from Madagascar, *Zygophyllum madagascariensis* and *Zygophyllum madecassum* had a peculiar property that in the sequence chromatograms signal intensity dropped drastically around the 100 bp mark within the ITS 1 spacer. Even using several different primers combinations and sequencing conditions we were not able distinguish the peaks from the background signal. A probable reason these bases could not be sequenced was a long poly-G region upstream from these bases, which could have caused slippage of the Taq-polymerase on the template DNA strand. The regions containing these bases were removed and treated as missing data. The matrix overall contained 63 taxa with a total length of 830 bp.

The sequences of the analysed specimens of *Zygophyllum orbiculatum* and the single *Zygophyllum stapffii* specimen were almost identical, barring two ambiguous base calls obtained from one of the *Zygophyllum orbiculatum* samples (Craven 5101). A single ambiguous base call was also found in another *Zygophyllum orbiculatum* (HK 2757). A single point mutation was shared between the *Zygophyllum stapffii* (DUB 1222) and one of the *Zygophyllum orbiculatum* specimens (HK 2753).

2.3.1.1 Maximum parsimony (PAUP)

In Table 2-3 are the attributes of the genetic matrix that was used to construct the maximum parsimony phylogenetic tree using PAUP. The matrix consisted of 63 taxa and 830 bp per taxon. Indels were treated as missing data for this analysis.

Table 2-3: The matrix attributes of the ITS gene region as determined by PAUP.

Information of the parsimony analysis performed	
Total Characters	830
Constant Characters	372
Parsimony Uninformative Characters	108 (13.01%)
Parsimony Informative Characters	350 (42.17%)
Variability (Uninformative + Informative)	458 (55.18%)
Number of Trees	9904
Minimum Length	784
Maximum Length	4326
Values of Tree	
Tree Length	1438
CI	0.545
RI	0.815

What is immediately evident is the high percentage of variability of this region. The percentage variability is approximately 55%, which is far higher than that of the *trnLF* and *rbcL* regions from Bellstedt *et al.* (2008), which were approximately 35% and 20% respectively (17). In the parsimony analysis, a total of 9904 equally parsimonious trees were retrieved and these were combined to produce a strict consensus tree. A bootstrap tree was also calculated. The strict consensus tree with the accompanying bootstrap values plotted on the supported nodes is shown in Figure 2-3. Only nodes with a support of 50% and higher are indicated. The Consistency Index (CI) in the case of this analysis is low, at 0.545. The Retention Index (RI) is relatively high in this ITS region analysis, at 0.815.

All specimens of *Zygophyllum orbiculatum* and *Zygophyllum stapffii* are retrieved as a single well supported clade with no internal support.

The phylogenetic tree indicates support for the subfamily Zygophylloideae. As with previous studies each of the large groupings (indicated in different colours) was mostly well supported. However the relationships between the large groupings within Zygophylloideae, namely the Asian *Zygophyllum* clade, the *Fagonia* & *Melocarpum* clade, the *Zygophyllum orbiculatum/stapffii* clade, the monotypic *Augea* clade, the southern African and Australian *Zygophyllum* clade and the *Agrophyllum/Tetraena* clade remain unsupported.

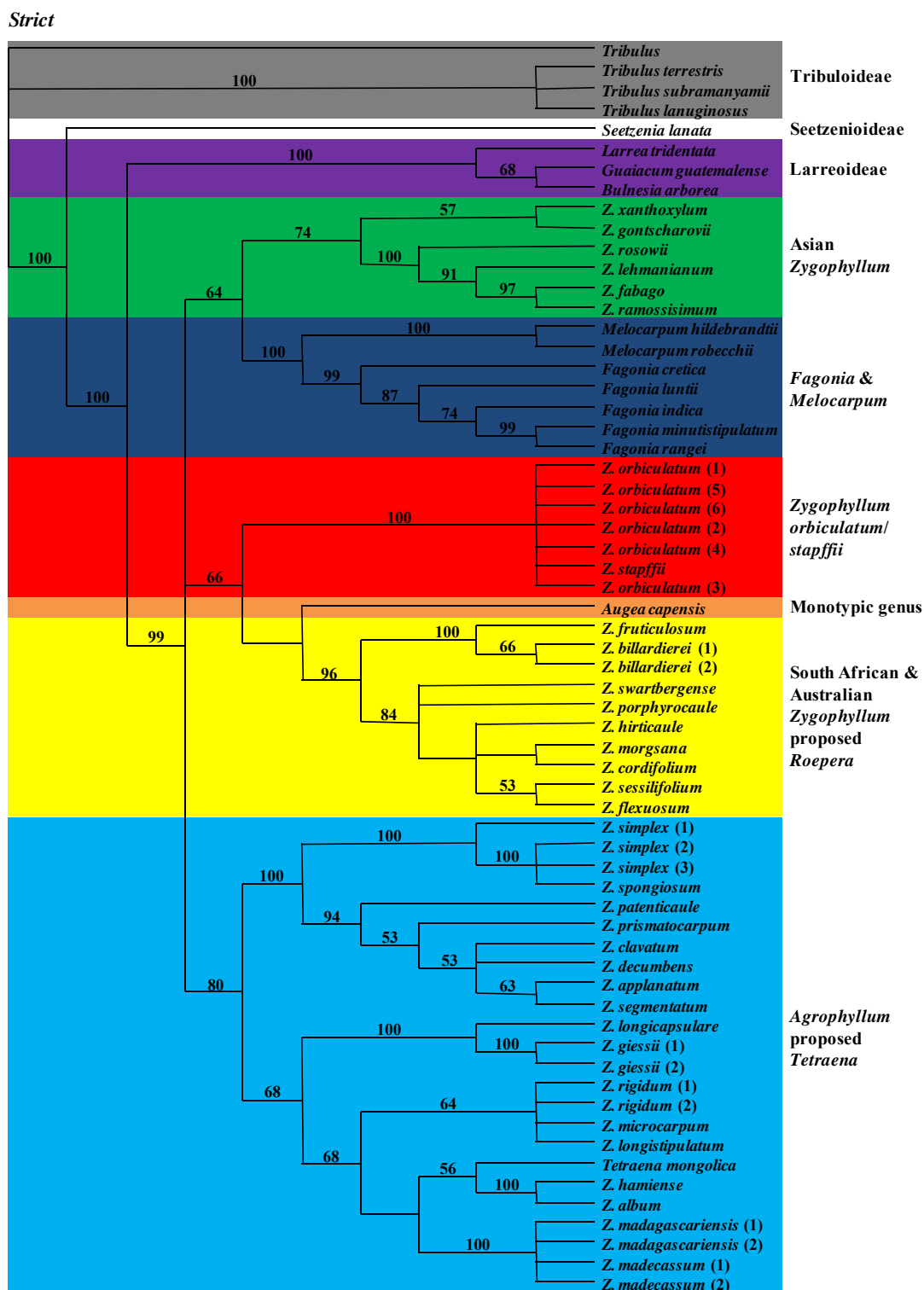


Figure 2-3: A strict consensus phylogenetic tree with bootstrap support values on the nodes. The strict consensus tree was drawn from 9904 equally parsimonious trees.

2.3.1.2 Maximum Likelihood (RAxML)

The tree produced by the maximum likelihood analysis, shown in Figure 2-4, retrieves all *Zygophyllum orbiculatum* and the single *Zygophyllum stapffii* specimen in a monophyletic clade well supported clade. Within this clade a strongly supported subclade of the two Namibian

Zygophyllum orbiculatum specimens (3 and 4) and the Namibian *Zygophyllum stapffii* specimen sister to the rest of the *Zygophyllum orbiculatum* specimens which were collected in Angola.

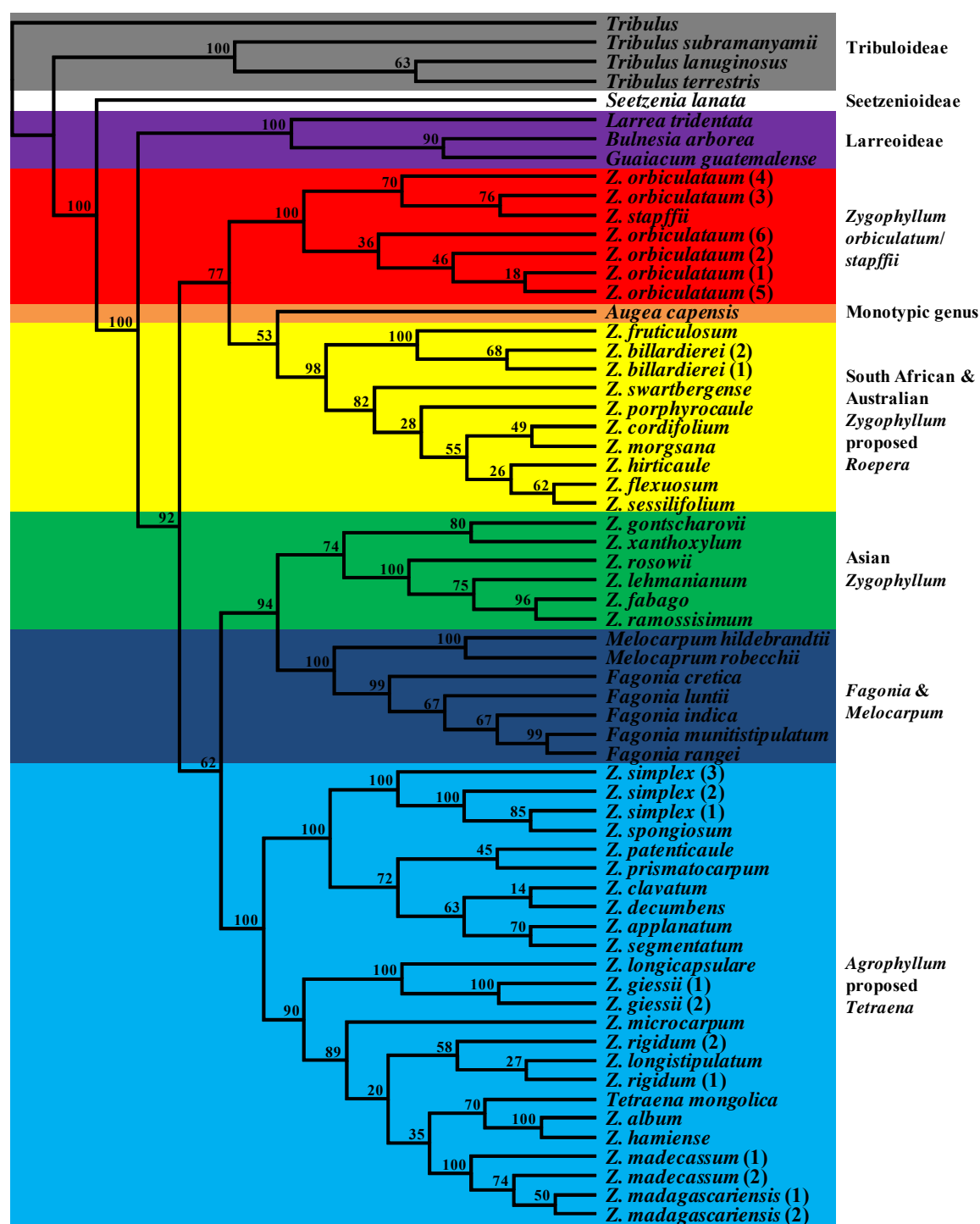


Figure 2-4: The most likely phylogenetic tree retrieved in a maximum likelihood analysis of the ITS region, indicating bootstrap support of the nodes. The main groupings are indicated in different colours. A node is considered supported if the value is equal or higher than 70%.

There was strong support for the subfamily Zygophylloideae and the large groupings (indicated in different colours) in the subfamily. There is strong support for a basal sister relationship between the groupings of southern African and Australian members of the subgenus

Zygophyllum, *Zygophyllum orbiculatum* and *Zygophyllum stapffii*, *Augea capensis* and subgenus *Agrophyllum*, *Fagonia* & *Melocarpum* and Asian *Zygophyllum*. The node separating *Zygophyllum orbiculatum* and *Zygophyllum stapffii* from *Augea capensis* and southern African and Australian members of the subgenus *Zygophyllum* is also supported. Good support for the sister relationship of Asian *Zygophyllum* and *Fagonia* & *Melocarpum* was also retrieved.

2.3.1.3 Bayesian Inference (MrBayes)

In this analysis all *Zygophyllum orbiculatum* specimens and the single *Zygophyllum stapffii* specimen appears as a single monophyletic clade with strong support.

As can be seen in the majority rule consensus tree, Figure 2-5, there is support for the subfamily Zygophylloideae, as well as for the large groupings (indicated in different colours) in the subfamily. Although the tree topology is identical to maximum likelihood analysis the only fully supported node within the subfamily Zygophylloideae is a sister relationship between Asian *Zygophyllum* and *Fagonia* & *Melocarpum*.

2.4 Discussion

In the context of this study, it is important that two of the *Zygophyllum orbiculatum* samples *Zygophyllum orbiculatum* (3) and *Zygophyllum orbiculatum* (4), by definition would have been classified as *Zygophyllum stapffii* based on their distributions. However, as they are unifoliolate they were classified as *Zygophyllum orbiculatum*. In the maximum likelihood analysis of the ITS region data these two specimens group with *Zygophyllum stapffii*, i.e. group together based on their geographic distribution in Namibia. All of the remaining *Zygophyllum orbiculatum* samples are retrieved basal to this clade without support in a strongly supported larger clade. In the parsimony and Bayesian Inference analyses the *Zygophyllum orbiculatum* samples and the *Zygophyllum stapffii* sample were retrieved in polytomies with a poorly supported subclade in the Bayesian analysis. Furthermore, the ITS sequences of *Zygophyllum orbiculatum* and *Zygophyllum stapffii* were almost identical, except for a few single base changes. From all of this evidence, it can therefore be concluded that *Zygophyllum orbiculatum* and *Zygophyllum stapffii* are conspecific. Previously, Bellstedt *et al.*, based on chloroplast *trnLF* and *rbcL* sequence data of two *Zygophyllum orbiculatum* and one *Zygophyllum stapffii* specimens, concluded that

Zygophyllum orbiculatum and *Zygophyllum stapffii* were conspecific (17). This conclusion is supported by the results obtained in this study. The identical morphology of the flowers, seed capsules, stems and even the individual leaves of the two species, with the exception that the leaf arrangement of *Zygophyllum orbiculatum* is unifoliolate and that of *Zygophyllum stapffii* is bifoliolate, gives further support to this conclusion. In terms of the taxonomy this means that the

earliest known name should be used for the taxon, i.e. *Zygophyllum orbiculatum* Welw. ex Oliver (140).

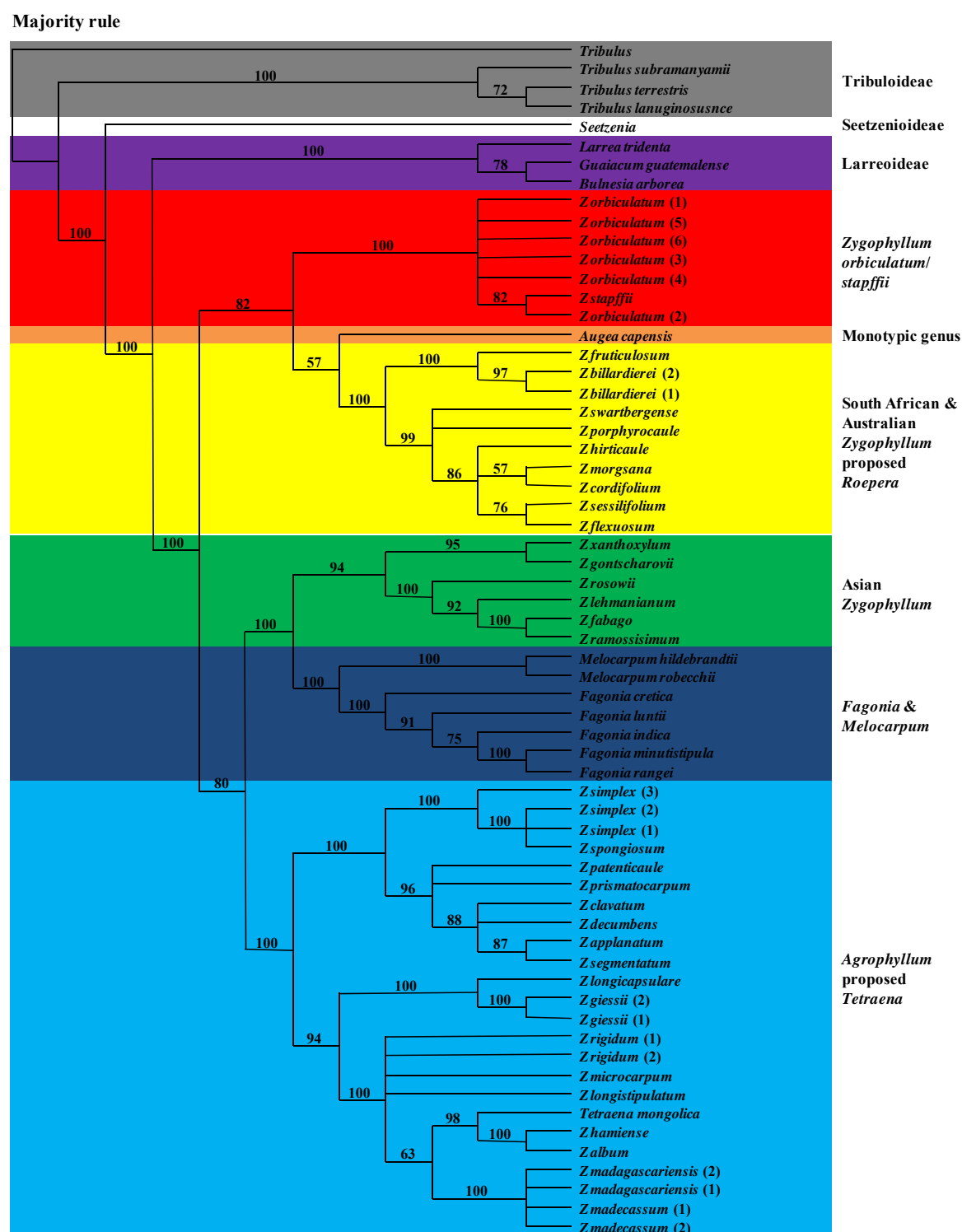


Figure 2-5: A majority rule consensus tree of the Bayesian Inference analyses of the ITS region. The numbers on the nodes are the posterior probability values. A node is considered supported when the corresponding value is equal or higher than 95%. The main groupings within Zygophylloideae are indicated in colour.

As was the case with the previous genetic studies on the subfamily Zygophylloideae using chloroplast markers (Beier *et al.*, 2003; Bellstedt *et al.* 2008), the ITS region also retrieved the

same large monophyletic groups (indicated with different colours in the phylogenetic trees), but was unable to resolve the phylogenetic relationships between these large groupings within the subfamily Zygophylloideae (13, 16).

In all of the analyses a highly supported monophyletic subfamily Zygophylloideae was retrieved. In none of the analyses could the relationships between the major groupings in the subfamily be resolved and a supported tree be obtained. None of the previous studies based on chloroplast sequences could resolve these relationships either (13, 17, 179, 180). An Incongruence Length Difference (ILD) test was performed on the ITS region phylogeny and the combined chloroplast *trnL*F and *rbcL* phylogeny (17) in order to determine whether such a combination was advisable. The results of the ILD test indicated that the data should not be combined as the chloroplast data were in conflict with the nuclear data. Consequently a phylogenetic analysis in which the chloroplast sequence data were combined with the ITS region sequence data was not performed.

There was resolution within each of the clades representing major groupings in the phylogenies generated from the ITS sequence data. This indicates high variability which is typical of the ITS marker. This marker has also been used in a study within the subfamily Zygophylloideae of the genus *Fagonia* (14). Within *Fagonia* this allowed a resolution of Old World and New World representatives of the genus. The ITS region has been used in a large number of plant groups for phylogenetic inference to the point where parts of the region have actually been proposed for DNA barcoding (42, 69, 99). This gene region is therefore very valuable in plant phylogenetics in spite of the fact that it could be problematic for phylogenetic analysis (see Table 1-5) (2, 9, 10, 46).

However, the evidence presented in this chapter indicates that additional data will be required to resolve the phylogenetic relationships within the subfamily Zygophylloideae (13, 17).

3 Next-Generation Sequencing and Combined Trees

3.1 Introduction

The first in-depth systematic study on the subfamily Zygophylloideae, by Beier *et al.* in 2003, using morphological characters, as well as molecular (*trnL* intron) data could not resolve the phylogenetic relationships within the subfamily. The study by Bellstedt *et al.* in 2008 utilizing both *trnLF* and *rbcL* data with a far wider taxon sampling also failed to resolve the relationships within the subfamily Zygophylloideae. The phylogenetic relationships within the subfamily Zygophylloideae using the ITS region as a molecular marker, in this study, also failed to resolve the phylogenetic relationships within the subfamily. In each of these studies though, the same large strongly supported monophyletic subgroupings appeared, i.e. the subgenus *Agrophyllum*/genus *Tetraena*; Asian subgenus *Zygophyllum*/*Zygophyllum sensu stricto*; genus *Fagonia* and *Melocarpum*; southern African and Australian members of subgenus *Zygophyllum*/*Roepera*; the species *Augea capensis* and *Zygophyllum stapffii/orbiculatum* yet the phylogenetic relationships of these large subgroupings relative to each other remained unresolved.

In order to address this problem the decision was taken to attempt a phylogenomic approach. The Central Analytical Facility at Stellenbosch University had recently, in 2011, acquired an ion semiconductor sequencer, and it was decided that it would be worth investigating this new technology to further this study. Even though the technology was very much in its infancy, and the amount of data that was theoretically achievable relative to the price of a sequencing analysis, was much cheaper which made a pilot study worth investigating. At the time we decided that whole genome sequencing was not ideal as the size of a typical *Zygophyllum* nuclear genome was unknown and too large. It was therefore decided to focus on the smaller chloroplast genome. The decision was taken to isolate and sequence the chloroplast genomes of a single species from each of the above-mentioned subgroupings identified within the subfamily Zygophylloideae (13, 17) and to use these sequences for phylogenetic inference. However, investigations have revealed that some members of *Zygophyllum* use C₄ photosynthesis whilst others use C₃ photosynthesis or CAM to fix CO₂. The use of the sequences of genes involved in photosynthesis for phylogenetic inference therefore needs to be treated with extreme care as these genes may be under selection pressure for these photosynthesis types.

3.2 Materials and Methods

3.2.1 Species that were investigated using the Next-Generation Sequencing approach.

Seven taxa, from the subgenus Zygophylloideae, and also an outgroup taxon from the closely related subfamily Tribuloideae were chosen for this study. Most of the chloroplast genome genes of *Bulnesia arborea* from the subfamily Larreoideae were available on Genbank as they had been sequenced as part of a larger study within the Angiosperms (210). The sequences from *Bulnesia arborea* were used in this analysis as Larreoideae is more closely related to the subfamily Zygophylloideae than the subfamily Tribuloideae.

The selected species (motivation for selection shown) from the monophyletic subgroups of the genus *Zygophyllum* were *Zygophyllum foetidum* (as representative of the southern African and Australian subgenus *Zygophyllum* named *Roepera* by Beier *et al.*, 2003; chosen because it a widespread species), *Zygophyllum fabago* (as representative of the Asian members of the subgenus *Zygophyllum*, retained by Beier *et al.*, 2003, type of the genus and of its section; fresh leaf material was supplied by our collaborator, Dr Gudrun Kadereit, University of Mainz, Germany), *Zygophyllum stapffii* (taxon with disputed placement in either subgenus *Agrophyllum* or subgenus *Zygophyllum* and conspecific with *Zygophyllum orbiculatum*) and lastly *Zygophyllum turbinatum* (as representative of the subgenus *Agrophyllum* named *Tetraena* by Beier *et al.*, 2003; chosen because it belongs to the large section *Bipartita* and was accessible). The taxa from the selected additional groups and outgroups are *Augea capensis* (currently a monotypic genus within subfamily Zygophylloideae), *Fagonia rangei* (as representative from the *Fagonia* & *Melocarpum* clade within subfamily Zygophylloideae), *Bulnesia arborea* (from the closest related subfamily to Zygophylloideae, namely Larreoideae) and lastly *Tribulus terrestris* (from the subfamily Tribuloideae) (13).

Below in Figure 4-1 are images of the above-mentioned taxa that were analysed in this study. The closely related outgroup species, *Bulnesia arborea*, is not shown as all pictures available of the plant on the internet were copyright protected. The reader is referred to (<http://www.freundfloweringtrees.com/bulnesia-arborea-verawood.html>).



Figure 3-1: Top left: (a), *Tribulus terrestris*, commonly known as the caltrop from the subfamily Tribuloideae (picture courtesy of Forest & Kim Starr). (b), *Fagonia rangei*. Unlike most members of the subfamily Zygophylloideae, *Fagonia* species have trifoliate leaves. (c), *Zygophyllum turbinatum*. (d), *Zygophyllum fabago* or the Syrian bean caper is the type for the genus *Zygophyllum* and was identified by Linnaeus in 1753 (102). (e), *Zygophyllum foetidum*. (f), *Augea capensis*. (g), *Zygophyllum stapffii* (pictures b-g by Dirk Bellstedt).

The list of plant species that were selected for this study is shown in Table 3-1. All species are from the family Zygophyllaceae, but only four are from the currently recognized genus *Zygophyllum*.

Table 3-1: The list of eight Zygophyllaceae species that were investigated in the whole-chloroplast genome sequencing analyses using the ion-semiconductor sequencer. Four of the species are currently considered to be of the genus *Zygophyllum*. The first taxon in the table is the outgroup species that was used as the reference genome. The two subsequent taxa in the table are from the two closest related subfamilies to subfamily Zygophylloideae, namely Tribuloideae (*Tribulus terrestris*) and Larreoideae (*Bulnesia arborea*), to which the rest of the taxa in the table belong.

Plant Species	Collection Locality
<i>Corynocarpus laevigata</i>	Outgroup from the order Cucurbitales and the reference genome (Genbank).
<i>Tribulus terrestris</i>	Stellenbosch Train station, Western Cape, South Africa. (P.D.W. van der Merwe, s.n.)
<i>Bulnesia arborea</i>	Chloroplast genes retrieved from Genbank.
<i>Fagonia rangei</i>	Close to embankment of the Orange River near Vioolsdrif border outpost between South Africa and Namibia. (Namibian side) (Bellstedt, 1314)
<i>Augea capensis</i>	Tankwa Karoo near Ceres, Western Cape, South Africa. (Bellstedt, 934)
<i>Zygophyllum stapffii</i>	Near Swakopmund, Namibia. (Bellstedt, 1222)
<i>Zygophyllum turbinatum</i>	Tankwa Karoo near Ceres, Western Cape, South Africa. (Bellstedt, 1386)
<i>Zygophyllum fabago</i>	Germany (Botanical Garden, Mainz University), Native to the Middle East.
<i>Zygophyllum foetidum</i>	Bridge across the Breede River between Robertson and McGregor Western Cape, South Africa (Robertson side) (Bellstedt 1387).

For the sake of clarity and not to overcomplicate the discussions of the results the following shortened names were used. These names conform to the taxonomy of Beier *et al.*, in 2003 (13). The name *Fagonia* was chosen for the large grouping representing the *Fagonia* & *Melocarpum* clade. The name *Tetraena* was chosen to represent the subgenus *Agrophyllum*. The name *Roepora* was chosen to represent the southern African and Australian members of the subgenus

Zygophyllum. The name *Augea* was chosen to represent the monotypic genus *Augea capensis*. The name *Zygophyllum stapffii* was chosen to represent the *Zygophyllum orbiculatum* and *Zygophyllum stapffii* conspecific species as the plant sample was taken from the southern most occurrence of the species in Namibia.

3.2.2 The chloroplast genome isolation and DNA purification procedures

In an effort to enrich the chloroplast DNA to be used in NGS a kit from Sigma specifically designed for the purpose of isolating intact chloroplasts from fresh plant leaf material was acquired. Chloroplast isolation using the kit is a very easy and straightforward procedure to follow and no overly expensive buffers/reagents or high-speed centrifugation steps are required as would be the case with a CsCl-gradient separation.

Chloroplast Isolation Buffer was added to plant material and this was macerated in a blender for a few seconds and filtered through a nylon mesh. The filtrate was briefly centrifuged at 200xg to sediment the chloroplasts in the filtrate. The supernatant was discarded and finally the pellet was resuspended in additional buffer and further purified in a 40%/80% Percoll[®] gradient using a swinging bucket centrifuge at 3 600xg. Intact chloroplasts were found in the interphase of the two concentrations of Percoll[®]. The intact chloroplasts were removed with a Pasteur pipette and the DNA was extracted in subsequent procedures.

Immediately after the chloroplasts were isolated the modified CTAB extraction protocol of Doyle & Doyle was used to isolate the DNA. The DNA was precipitated and resuspended after the second cleanup step. The purified DNA concentration and yield were determined on a NanoDrop 1000 Spectrophotometer system and the DNA was subsequently sent to the Stellenbosch University's Central Analytical Facility (CAF) for ion semi-conductor sequencing. At this facility the DNA sample was further purified and contaminating RNA was removed using RNA digestion. The run reports and FastQC reports for the different sample sequence analyses are given in addenda 6-2-6-7. Please note that through a technical glitch the run report and FastQC report for *Augea capensis* were not available at time of writing.

3.2.3 Contiguous sequence assembly and bioinformatic strategies

Sequence data from the ion semiconductor sequencing analyses were obtained in the form a large number of short reads. *De novo* assemblies of these reads from the ion semiconductor sequencing analyses were performed to generate contiguous sequences ("contigs") by either Newbler (<http://454.com/products/analysis-software/index.asp>) or Mira

3.2.5 Aligning contigs to the chloroplast reference genome in CodonCode Aligner.

Chloroplast genomes of most plants contain an inverted repeat region which has identical genetic information. For this reason only one of the inverted repeat regions were retained for the alignment of the contigs since some of the contigs would either align to one or the other of the IR regions which complicated the alignment process. Additionally the information of the individual genes of *Corynocarpus laevigata* were also downloaded and added to data that would be aligned to the reference genome since this would show exactly where each gene starts and ends on the reference genome. The *Corynocarpus laevigata* chloroplast genome contains 128 genes of which 30 are in the inverted repeat region, i.e. 15 on each of the two repeat regions. This meant that only 113 unique genes were required to account for all 128 genes on the reference chloroplast genome. All settings were kept at default, except the settings in the alignment section in the Preferences tab given below in sTable 3-2.

Table 3-2: The list of settings changed in the alignment section of the Preferences tab.

Preferences	Value
Algorithm	End to end alignments
Minimum percentage identity	50.0
Minimum overlap length	10
Minimum score	10
Maximum unaligned end overlap	90.0
Bandwidth (masimum gap size)	250.0
Word length	6
Match score	1
Mismatch penalty	-2
Additional first gap penalty	-3
Uncovered reference sequence	Leave as is

It was decided on that the only molecular data that would be used would be the regions where all taxa had information available, thus minimizing missing data which could lead to distorted tree topologies.

The only consistently sequenced region, other than chloroplast genetic information, for all the taxa was the nuclear ITS cassette. Since this region occurs in multiple copies, the odds of it being sequenced were quite high. A reference genome for this region was needed, preferably a complete cassette with the 18S rRNA, ITS1, 5.8S rRNA, ITS2 and 26S rRNA. BLAST searches of contigs consistently retrieved the *Humulus lupulus* (hops) ITS, 18S or 26S regions, which upon further investigation was revealed to form part of a full cassette of the region. This was used as the reference genome in CodonCode Aligner and all contigs of the applicable taxa were added, as well all the individual genes mentioned above of *Corynocarpus laevigata*, which were

available on Genbank. The contigs were filtered and aligned in CodonCode Aligner and the result was exported as a PHYLIP file. Within BioEdit the *Humulus lupulus* genome was removed from the matrix leaving only the reference genome and our sequenced taxa. The cassette containing these genes however was not available for *Bulnesia arborea* and several gaps were also identified in the the cassette of *Zygophyllum turbinatum* and in *Tribulus terrestris*. A sample of *Bulnesia arborea* leaf material was obtained from the Royal Botanic Garden Edinburgh (RBGE). Several primers were procured from Molecular Cell Biology Department of UCT (see Table 3-3) and the entire cassette was completely sequenced for this taxon and the gaps of *Zygophyllum turbinatum* and in *Tribulus terrestris* were also covered.

Table 3-3: The table indicates the primers that were used to amplify the 18S and 26S rDNA from the ITS cassette. The primers are indicated in the 5'-3' orientation (25, 94).

18S rRNA	
N-NS1	GTA GTC ATA TGC TTG TCT G
C-18H	GCC CTT CCG TCA ATT CCT TTA AGT TTC AGC
18Srev	CCT TCC TCT AAA CGA TAA GGT TC
26S rRNA	
641rev	TTG GTC CGT GTT TCA AGA CG
950rev	GCT ATC CTG AGG GAA ACT TC
1229rev	ACT TCC ATG ACC ACC GTC CT
1499rev	ACC CAT GTG CAA GTG CCG TT
1839rev	TTC ACC TTG GAG ACC TGA TG
2134rev	GGA CCA TCG CAA TGC TTT GT
2426rev	CCT ACA CCT CTC AAG TCA T
268rev	GCA TTC CCA AAC AAC CCG AC
2782rev	GGT AAC TTT TCT GAC ACC TC
3058rev	TTC GCG CCA CTG GCT TTT CA
3331rev	ATC TCA GTG GAT CGT GGC AG
N-nc26S1	CGA CCC CAG GTC AGG CG
N-nc26S2	GAG TCG GGT TGT TTG GGA
N-nc26S3	AGG GAA GCG GAT GGG GGC
N-nc26S4	TTG AAA CAC GGA CCA A
N-nc26S5	CGT GCA AAT CGT TCG TCT
N-nc26S6	TGG TAA GCA GAA CTG GCG
N-nc26S7	GAT GAG TAG GAG GGC GCG
N-nc26S8	ACG TTA GGA AGT CCG GAG
N-nc26S9	AAT GTA GGC AAG GGA AGT
N-nc26S10	TAA AAC AAA GCA TTG CGA
N-nc26S11	AAT CAG CGG GGA AAG AAG
N-nc26S12	GTC CTA AGA TGA GCT CAA
N-nc26S13	CCT ATC ATT GTG AAG CAG
N-nc26S14	TTA TGA CTG AAC GCC TCT
N-nc26S15	TGC CAC GAT CCA CTG AGA

3.2.6 Genetic markers used after the contig alignment to the reference genomes

It has been shown in literature that the chloroplast gene named *rbcL*, which is the large subunit of the ribulose-1,5-bisphosphate enzyme is under positive selection (34). Since it has also been shown that C₃, CAM and C₄ photosynthesis exists within this group (38, 112, 123, 131, 170, 220), it was decided to divide the sequenced protein-coding genes into two large groups. The first group was composed of all the genes involved directly or indirectly in photosynthesis, since if *rbcL* was under positive selection, the same might be true of the other photosynthetic genes. The second group of genes was composed of all of the genes not involved with the photosynthetic process. As mentioned, the nuclear encoded ITS cassette was also used in the subsequent phylogenetic analyses. These large ITS cassette DNA sequences were aligned with the shorter ITS regions from the previous study so that the alignment of the highly variable ITS1 and ITS2 was identical to that of the previous alignment matrix of the ITS region. The three groupings of genes that were used in the subsequent analyses are indicated below in Table 3-4, Table 3-5 and Table 3-6 respectively.

Table 3-4:-The table below indicates 21 genes that are directly (light reactions) or indirectly (dark reactions) involved in photosynthesis that were successfully sequenced and used.

Genes involved with the Photosynthetic process		Gene name (Abbreviated)	Gene Name
	1	<i>psb D</i> (protein-coding)	Photosystem II D2 protein
	2	<i>psb C</i> (protein-coding)	Photosystem II 44 kDa protein
	3	<i>psb A</i> (protein-coding)	Photosystem II 32 kDa protein
	4	<i>psb Z</i> (protein-coding)	Photosystem II reaction center Z protein
	5	<i>psa B</i> (protein-coding)	Photosystem I P700 apoprotein A2
	6	<i>psa A</i> (protein-coding)	Photosystem I P700 apoprotein A1
	7	<i>rbc L</i> (protein-coding)	Ribulose-1,5-bisphosphate carboxylase/oxygenase large subunit
	8	<i>psa J</i> (protein-coding)	Photosystem I reaction center subunit IX
	9	<i>psb B</i> (protein-coding)	Photosystem II 47 kDa protein
	10	<i>psb T</i> (protein-coding)	Photosystem II T protein
	11	<i>psb N</i> (protein-coding)	Photosystem II N protein
	12	<i>psb H</i> (protein-coding)	Photosystem II 10 kDa Phosphoprotein
	13	<i>cem A</i> (protein-coding)	Chloroplast envelope membrane protein (ycf10)
	14	<i>pet A</i> (protein-coding)	Cytochrome f
	15	<i>pet L</i> (protein-coding)	Cytochrome b6/f complex subunit VI
	16	<i>pet G</i> (protein-coding)	Cytochrome b6/f complex subunit V
	17	<i>pet B</i> (protein-coding)	Cytochrome b6
	18	<i>pet D</i> (protein-coding)	Cytochrome b6/f complex subunit IV
	19	<i>atp E</i> (protein-coding)	ATPase epsilon subunit
	20	<i>atp B</i> (protein-coding)	ATPase beta subunit
	21	<i>atp I</i> (protein-coding)	ATP synthase CF0 A subunit

Table 3-5:-The table indicates 20 chloroplast genes that are not involved in photosynthesis that were successfully sequenced. Fourteen of the genes in this list are coding genes, but for the tRNA genes only the introns were of importance as the genes themselves are very small and contain little or no informative characters. The other four genes are ribosomal RNA genes, which are not translated into amino acids, but are coding and functional as RNA.

Genes not involved with photosynthesis		Gene name (Abbreviated)	Gene Name
	1	<i>rps</i> 2 (protein-coding)	Ribosomal protein S2
	2	<i>rps</i> 14 (protein-coding)	Ribosomal protein S14
	3	<i>rps</i> 18 (protein-coding)	Ribosomal protein S18
	4	<i>rpl</i> 20 (protein-coding)	Ribosomal protein L20
	5	<i>rps</i> 12 (protein-coding)	Ribosomal protein S12
	6	<i>rps</i> 11 (protein-coding)	Ribosomal protein S11
	7	<i>rps</i> 8 (protein-coding)	Ribosomal protein S8
	8	<i>rpl</i> 14 (protein-coding)	Ribosomal protein L14
	9	<i>rpl</i> 16 (protein-coding)	Ribosomal protein L16
	10	<i>rpo</i> C2 (protein-coding)	RNA polymerase beta II subunit
	11	<i>rpo</i> C1 (protein-coding)	RNA polymerase beta I subunit
	12	<i>rpo</i> B (protein-coding)	RNA polymerase beta subunit
	13	<i>rpo</i> A (protein-coding)	RNA polymerase alpha subunit
	14	<i>mat</i> K (protein-coding)	maturase K
	15	<i>rrn</i> 16 (RNA-coding)	16S ribosomal RNA
	16	<i>rrn</i> 23 (RNA-coding)	23S ribosomal RNA
	17	<i>rrn</i> 4.5 (RNA-coding)	4.5S ribosomal RNA
	18	<i>rrn</i> 5 (RNA-coding)	5S ribosomal RNA
	19	<i>trn</i> A (intron)	tRNA (alanine)
	20	<i>trn</i> I (intron)	tRNA (isoleucine)

Table 3-6: The nuclear ITS cassette genes for all species that were also sequenced using the Ion Torrent semiconductor sequencing.

Nuclear Genetic		Gene name (Abbreviated)	Gene Name
	1	<i>rrn</i> 18 (RNA-coding)	18S ribosomal RNA
	2	ITS1 (intergenic)	Internal Transcribed Spacer 1
	3	<i>rrn</i> 5.8 (RNA-coding)	5.8S ribosomal RNA
	4	ITS2 (intergenic)	Internal Transcribed Spacer 2
	5	<i>rrn</i> 26 (RNA-coding)	26S ribosomal RNA

3.2.7 Phylogenetic inference methods used

The three gene groups that were generated were used in several phylogenetic analyses software packages, including PAUP b4.10, MrBayes 3.2.2 and RAxML 8.0.24 (73, 165, 193, 199). The process for each of the programs will be discussed below.

3.2.7.1 Preparation of the genetic information for phylogenetic analyses.

The alignment was exported from CodonCode Aligner as a PHYLIP file which was edited in BioEdit. Each of the genes that were identified for use in subsequent phylogenetic analyses was isolated within its own BioEdit alignment file. Subsequently all bases were checked by eye and aligned within their correct codon positions where applicable. For each of these isolated genes a NEXUS file was generated. These NEXUS files were used within SequenceMatrix v1.7.8 (204).

SequenceMatrix functions as a concatenation program and joins all selected files into a single alignment file. SequenceMatrix has a specific order in which it concatenates the NEXUS files based on the names (numbers first and then letters). Of all the options that were available, we chose to export these concatenated alignments as a PHYLIP file. Within the user interface of SequenceMatrix the order of the concatenation is clearly visible and could also be used to generate the partitioning scheme which was used in the model-based programs. The generated PHYLIP file was used for RAxML analyses, but for the PAUP and MrBayes analyses NEXUS files were generated. The generated PHYLIP files were opened with wordpad. The concatenated sequences were then copied out of the PHYLIP file and pasted into a pre-existing NEXUS file in which the new length for the dataset was noted. PAUP and MrBayes defines missing data and indels as dashes (“-“). RAxML distinguishes between these two. Indels are coded as (“-“) and missing data as (“?”). As indel-coding was not used for these analyses all question mark characters were replaced with dashes for PAUP and MrBayes.

3.2.7.2 The use of data partitioning in model-based phylogenetic programs

MrBayes and RAxML are known as model-based programs, meaning the data is analysed using several evolutionary models. The best-fitting model is used in the analyses, which leads to the best possible phylogenetic relationships given the data and available models. As genes are the functional entities being used to determine the phylogenetic relationships of the taxa being studied, it is logical to analyse each gene separately from the other to determine which evolutionary model fits best to each of the genes. Protein-coding genes can also be uniquely analysed by means of codon-partitioning. Each of the three base pair positions within the triplet codon which codes for amino acids can be analysed independently from the other two leading to improved resolution.

In this study 37 protein-coding genes were included, the rest being either RNA-coding, intergenic spacers or introns. This means there are 111 partitions for the protein-coding genes alone and an overall amount of 117 user-defined partitions. The program PartitionFinder is designed to work specifically with model-based phylogenetic programs like RAxML, MrBayes and BEAST. PartitionFinder is a Python-based (Python v2.7.8 Windows x86-64) application that runs in Microsoft Command Prompt. Data is added in PHYLIP format as well as a CFG file containing the user defined partition information. Within this CFG file the type of analysis is defined, i.e RAxML, MrBayes or BEAST. For each of these analysis there are three possible model selection settings BIC (Bayesian information criterion), AIC (Akaike information criterion) or AICc (Akaike information criterion corrected). Each criterion has a set amount of

evolutionary models it can implement on the dataset using the user-defined partitions as the scaffold for the evolutionary models.

The user-defined partitions are subsequently condensed into fewer partitions, as well as describing the best evolutionary model for these newly generated combined partitions. In the case of RAxML the resulting partition scheme can be directly used in analyses on the CIPRES Science Gateway website (121). For MrBayes the partitioning scheme needs to be incorporated into a NEXUS file, also containing the gene matrix, which also indicates substitution models for each of the PartitionFinder-generated partitioning schemes.

For each of the RAxML and MrBayes the model selection was set to BIC.

3.2.7.3 *MrBayes*

Once PartitionFinder suggested the appropriate models for the condensed set of partitions a NEXUS file was generated and the substitution models were defined. For each of the analyses performed in MrBayes the the settings in Table 3-7 were implemented.

Table 3-7: The setting below were used in all Bayesian inference analyses contained in this study.

Mcmc nruns	2
Nchains	4
Ngen	10 000 000
Temp	0.2
Smaplefreq	1 000
Printfreq	1 000
Savebrlens	Yes

3.2.7.4 *Conflict in nuclear and chloroplast gene information of taxa*

The apparent conflict in the chloroplast and nuclear data of taxa was further investigated using the non-photosynthetic gene regions and the ITS cassette gene region. Only the non-photosynthetic genes were used as it was assumed that the photosynthetic genes were possibly under selection pressure, which could lead to distorted relationships in phylogenetic inference.

The complete concatenated gene sequences of taxa that show conflict in their nuclear and chloroplast gene information were duplicated and the nuclear regions in these sequences were replaced with missing data leaving only the chloroplast data and *vice versa*, according to the method of Pirie *et al.* (154). Only a maximum likelihood analysis was performed for this investigation with codon partitions for protein-coding genes and gene region partitions where applicable.

There are two main hypotheses that could account for conflict between data of genes from different cellular compartments. The first is that a hybridization event between an ancestor of *Zygophyllum* group and an ancestor of the *Roepera* group hybridized. In subsequent offspring of this primary hybrid the nuclear genome recombined to retain the ancestral nuclear ITS cassette of the *Zygophyllum* group and the chloroplast genome of the ancestor of the *Roepera* group. Thus the *Zygophyllum* group would show evidence of a chloroplast capture event due to this hybridization (37, 58, 139, 196, 203). The second hypothesis is that rapid divergence of such groups of organisms leads to incomplete lineage sorting. Incomplete lineage sorting occurs when certain groupings of gene alleles combine in different combinations in rapidly diverging lineages from the most recent common ancestor.

Due to time constraints only an initial investigation into which of these two hypotheses is more likely was conducted. This investigation was conducted in Mesquite using its coalescence package. The most recently calculated divergence for the order Zygophyllales and Cucurbitales is approximately 100 Mya (Supplementary Table S1) (126). The initial input trees were the trees from the ITS cassette region and the non-photosynthetic chloroplast genes analyses from RAxML (maximum likelihood). Both trees were scaled so that they could be directly compared to each other. Only three initial population sizes were tested, namely 10 000, 100 000 and 1 000 000. A thousand trees were simulated and a generation time of 2 years was used. The reason why a generation time of 2 years was chosen was because e.g. *Zygophyllum simplex* is an annual whilst many other species can form seeds at the age of two to three years.

3.2.8 Tree editing

3.2.8.1 Tree editing for PAUP

The trees that were generated within in PAUP were exported as a PICT file. These PICT files were opened and edited in Microsoft PowerPoint 2007.

3.2.8.2 Tree editing for MrBayes

The two sets of trees files that were generated in the two independent runs within MrBayes were first imported into LogCombiner v1.8.0 which is part of the BEAST v1.8.0 software package (45). Within this program the first 10% of all the generated trees were removed and the two independent sets of tree files were combined into a single tree file (.t). This combined tree file was imported into Mesquite 3.0 (build644) (106). Within Mesquite a majority rule consensus tree was generated with proportional branch lengths. This tree was exported in Newick format (28). The Newick format file was imported into the phylogenetic tree drawing program named

Tred (<http://www.reelab.net/tred>). The tree was exported as a PDF and was subsequently imported into Adobe Acrobat Reader Professional v9.0. The tree was exported from this program as an encapsulated postscript file (.eps). This EPS file was opened in Microsoft PowerPoint 2007 in which the final editing was performed.

3.2.8.3 *Tree editing for RAxML*

The phylogenetic trees for the RAxML analyses were exported from the completed analyses on the CIPRES Science Gateway (121) in Newick file format. The same procedure was followed from this point as was used for the Bayesian trees.

3.3 Results

3.3.1 Phylogenetic analyses of the genes involved in photosynthesis

3.3.1.1 *Parsimony (PAUP)*

For the parsimony analysis of the genes involved in photosynthesis the matrix contained 17 435 bp of 21 protein-coding genes for nine taxa. A single most parsimonious tree was retrieved. The tree parameters are shown in Table 3-8 while the tree is shown in Figure 3-3(a).

Within this tree there is full support for the subfamily Zygophylloideae, and a basal position for the subgenus *Tetraena* within the subfamily Zygophylloideae. There is moderate support for the remainder of the ingroups within Zygophylloideae, but the only supported node is the sister relationship of *Augea* and *Zygophyllum stapffii*.

Table 3-8: The matrix information of the parsimony analysis of the genes involved in photosynthesis.

Information of the parsimony analysis performed	
Total Characters	17435
Constant Characters	14658
Parsimony Uninformative Characters	1648 (9.45%)
Parsimony Informative Characters	1129 (6.48%)
Variability (Uninformative + Informative)	2777 (15.93%)
Number of Trees	1
Minimum Length	3158
Maximum Length	4960
Values of Tree	
Tree Length	3924
CI	0.805
RI	0.575

3.3.1.2 *RAxML (Maximum Likelihood)*

The maximum likelihood phylogenetic tree is shown in Figure 3-3(b). The topology of the phylogenetic tree from the maximum likelihood analysis of the genes involved in photosynthesis is identical to the phylogenetic tree of the parsimony analysis. Again there is full support for the subfamily Zygothylloideae with the basal-most group being the *Tetraena* clade. There is also strong support for a sister relationships of *Augea* and *Zygothylum stapffii*. This group is embedded within the *Zygothylum* clade with strong support.

3.3.1.3 *MrBayes (Bayesian Inference)*

The majority-rule consensus phylogenetic tree of the Bayesian inference is shown in Figure 3-3(c). The topology of the phylogenetic tree is identical to the two previous analyses. There is strong support for the subfamily Zygothylloideae, with the *Tetraena* clade being the most basal group. The sister relationship of *Augea* and *Zygothylum stapffii* was fully supported. This group is embedded with the *Zygothylum* clade with strong support. The separation of this larger grouping from the *Fagonia* clade and the *Roepera* clade is fully supported. The sister relationship of the *Fagonia* clade and the *Roepera* clade is fully supported.

3.3.2 **Phylogenetic analyses of the genes not involved in photosynthesis**

3.3.2.1 *Parsimony (PAUP)*

The parsimony analysis for the chloroplast genes not involved in photosynthesis contained protein-coding, rRNA-coding, as well as intronic regions. The sequence matrix contained 22 346 bp for 9 taxa. The matrix information is shown in Table 3-9. A single most parsimonious tree that was retrieved is shown in Figure 3-3(d).

The topology for this phylogenetic tree is slightly different to that of the phylogenetic tree of the photosynthetic genes. There is full support the subfamily Zygothylloideae with the *Fagonia* clade at the basal position. There is full support for the sister relationship of *Augea* and *Zygothylum stapffii*. This group is embedded within the *Roepera* clade with 95% bootstrap support. This group is again embedded within the *Zygothylum* clade with 92% bootstrap support.

Table 3-9: The matrix information of the genes not involved in photosynthesis.

Information of the parsimony analysis performed	
Total Characters	22346
Constant Characters	18042
Parsimony Uninformative Characters	2702 (12.09%)
Parsimony Informative Characters	1602 (7.17%)
Variability (Uninformative + Informative)	4304 (19.26%)
Number of Trees	1
Minimum Length	5001
Maximum Length	7482
Values of Tree	
Tree Length	5945
CI	0.841
RI	0.620

3.3.2.2 *RAxML (Maximum Likelihood)*

In Figure 3-3(e) the single most likely tree of the maximum likelihood analysis of the chloroplast genes not involved in photosynthesis is shown. The tree topology is identical to the parsimony analysis for these genes. The sister relationship between *Augea* and *Zygophyllum* clade is fully supported. This grouping is embedded within the *Roepera* clade with 98% bootstrap support. This grouping is again embedded within the *Zygophyllum* clade albeit slightly less supported at 73%. This larger group's separation from the *Fagonia* clade, as well as from the *Tetraena* clade is fully supported. The sister relationship between the *Fagonia* clade and the *Tetraena* clade is slightly less supported at 69%.

3.3.2.3 *MrBayes (Bayesian Inference)*

The majority rule consensus tree of the chloroplast genes not involved in photosynthesis is shown in Figure 3-3(f). The topology of this tree is again identical to the trees obtained in the two previous phylogenetic analyses of these genes. There is complete support for the subfamily Zygophylloideae. The sister relationship between *Augea* and *Zygophyllum stapffii* is completely supported, This grouping is again embedded within the *Roepera* clade with complete support which in turn is again embedded within the *Zygophyllum* clade with complete support. The separation of this larger grouping from the *Fagonia* clade and the *Tetraena* clade is fully supported. The sister relationship of the *Fagonia* clade and the *Tetraena* clade is the only node not supported, as the posterior probability is only 82%.

3.3.3 Phylogenetic analyses of the genes of the nuclear ITS cassette

3.3.3.1 Parsimony (PAUP)

The parsimony analysis of the nuclear ITS cassette, containing three rRNA genes and 2 intergenic spacer sequences, had a total length of 5 930 bp for 9 taxa. The matrix parameters are given in Table 3-10.

Table 3-10: The matrix parameters of the nuclear ITS cassette.

Information of the parsimony analysis performed	
Total Characters	5930
Constant Characters	5044
Parsimony Uninformative Characters	523 (8.82%)
Parsimony Informative Characters	363 (6.12%)
Variability (Uninformative + Informative)	886 (14.94%)
Number of Trees	1
Minimum Length	1137
Maximum Length	1738
Values of Tree	
Tree Length	1486
CI	0.765
RI	0.419

A single most parsimonious tree was retrieved which is shown in Figure 3-3(g). The topology for the phylogenetic tree of this region is again different from the previous two gene region analyses. There is full support for the subfamily Zygophylloideae with the *Zygophyllum* clade being in the most basal position. The sister relationship of *Augea* and *Zygophyllum stapffii* is strongly supported. This group is embedded within the *Roepera* clade with strong support.

3.3.3.2 RAxML (Maximum Likelihood)

The most likely phylogenetic tree from the maximum likelyhood analysis of the nuclear ITS cassette is shown in Figure 3-3(h). The topology of this phylogenetic tree is identical to tree obtained from the parsimony analysis of this region. The entire subfamily Zygophylloideae is fully supported with the *Zygophyllum* clade at the most basal position. The sister relationship of *Augea* with *Zygophyllum stapffii* is completely supported. This group is embedded within the *Roepera* clade with strong support. This larger grouping separated from the *Fagonia* clade and the *Tetraena* clade is moderately supported at 70%. The sister relationship between the *Fagonia* clade with the *Tetraena* clade is weakly supported at 54%.

3.3.3.3 MrBayes (Bayesian Inference)

The majority-rule consensus tree of the nuclear ITS cassette is shown in Figure 3-3(i). The phylogenetic tree topology is identical to the two previous analyses of this region. The subfamily Zygophylloideae is completely supported with the *Zygophyllum* clade at the most basal position. The sister relationship of *Augea* with *Zygophyllum stapffii* is completely supported. This group is embedded within the *Roepera* clade with almost complete support. This larger grouping separation from the *Fagonia* clade and the *Tetraena* clade is completely supported. The sister relationship of the *Fagonia* clade with the *Tetraena* clade is completely supported.

3.3.4 Phylogenetic analyses of combined chloroplast genes

3.3.4.1 Parsimony (PAUP)

For this analysis all gene sequences of the chloroplast genomes were combined. The matrix contained a total number of 39 781 bp for nine taxa. The matrix parameters are indicated in Table 3-11.

Table 3-11: The matrix parameters of all the chloroplast gene sequences as calculated by the parsimony analysis.

Information of the parsimony analysis performed	
Total Characters	39781
Constant Characters	32700
Parsimony Uninformative Characters	4350 (10.93%)
Parsimony Informative Characters	2731 (6.87%)
Variability (Uninformative + Informative)	7081 (17.80%)
Number of Trees	1
Minimum Length	8159
Maximum Length	12442
Values of Tree	
Tree Length	9871
CI	0.827
RI	0.600

A single most parsimonious tree was retrieved from the parsimony analysis. The phylogenetic tree is shown in Figure 3-4(a). There is complete support for the subfamily Zygophylloideae with the *Tetraena* clade as the most basal group. The sister relationship of *Augea* with *Zygophyllum stapffii* is completely supported. This group is embedded within the *Roepera* clade with strong support. This group is embedded within the *Zygophyllum* clade with strong support. This larger group is embedded within the *Fagonia* clade, but is only weakly supported at 64%.

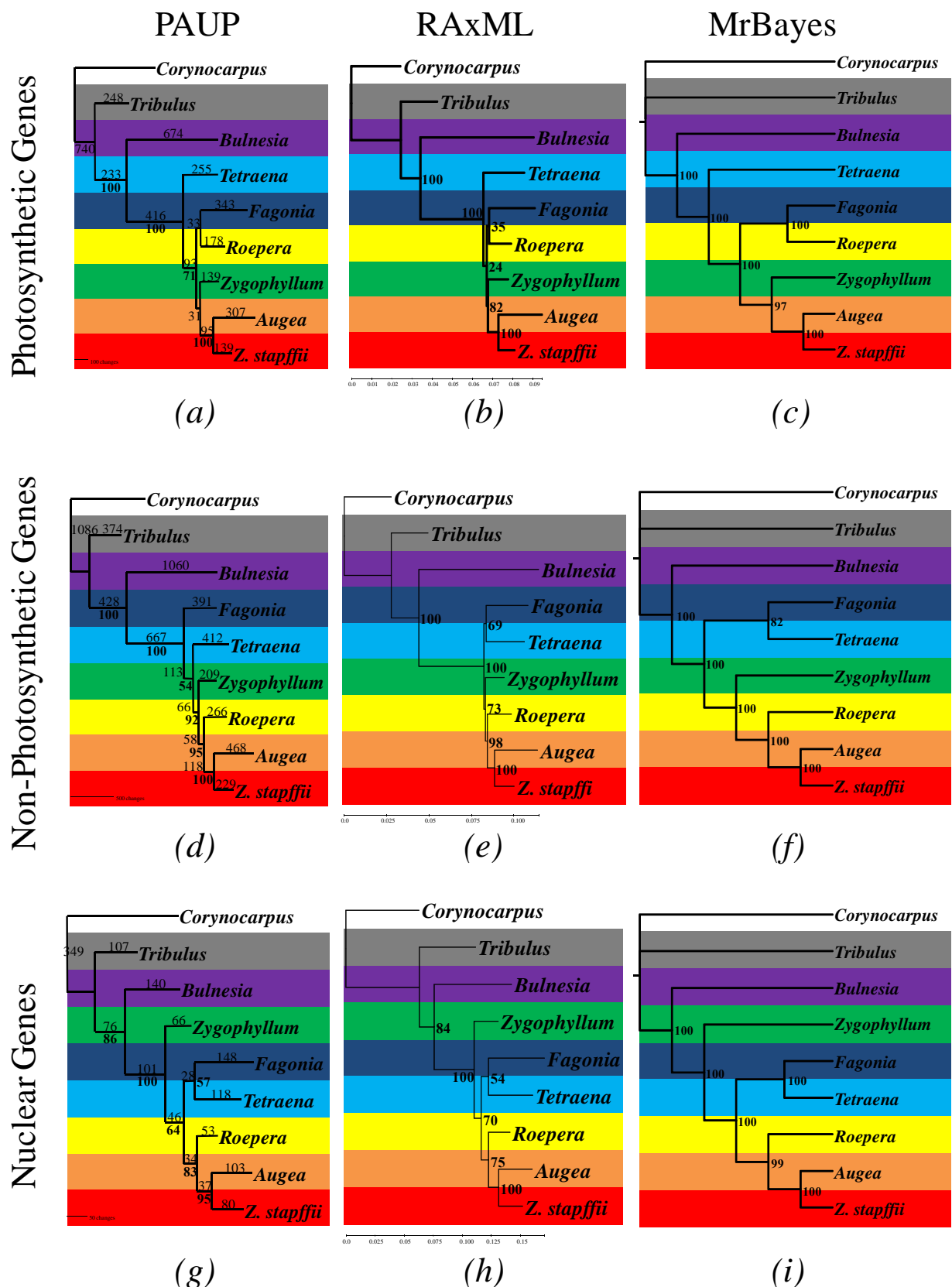


Figure 3-3: (a), The most parsimonious tree of the genes involved in photosynthesis with bootstrap support indicated below the nodes. (b), The most likely tree from the maximum likelihood analysis of the genes involved in photosynthesis with proportional branch lengths and bootstrap support for the nodes. (c), The majority-rule consensus tree calculated from the stored trees of two independent Bayesian inference analyses of the genes not involved in

photosynthesis, with posterior probabilities indicated. (d), The most parsimonious tree of the genes not involved in photosynthesis with bootstrap support indicated below the nodes. (e), The most likely tree from the maximum likelihood analysis of the genes not involved in photosynthesis with proportional branch lengths and bootstrap support for the nodes. (f), The majority-rule consensus tree calculated from the stored trees of two independent Bayesian inference analyses of the genes not involved in photosynthesis, with posterior probabilities indicated. (g), The most parsimonious tree of the nuclear ITS cassette genes and the spacer sequences, with bootstrap support indicated below the nodes. (h), The most likely tree from the maximum likelihood analysis of the nuclear ITS cassette genes and the spacer sequences, with proportional branch lengths and bootstrap support for the nodes. (i), The majority-rule consensus tree calculated from the stored trees of two independent Bayesian inference analyses of the nuclear ITS cassette genes and the spacer sequences, with posterior probabilities indicated.

3.3.4.2 *RAxML (Maximum Likelihood)*

The most likely phylogenetic tree from the maximum likelihood analysis is shown in Figure 3-4(b). The topology is slightly different from the parsimony analysis of the same sequence data matrix. There is complete support for the subfamily Zygothylloideae which is separated into two completely supported clades. The first clade consists of a group containing the *Fagonia* clade, as well as the *Tetraena* clade, which is strongly supported. The second group contains a completely supported sister relationship of *Augea* with *Zygothylum stapffii*. This group is embedded within the *Roepera* clade with almost complete support. This grouping is embedded within the *Zygothylum* clade with strong support.

3.3.4.3 *MrBayes (Bayesian Inference)*

The majority-rule consensus tree from the Bayesian inference analyses is shown in Figure 3-4(c). The topology is identical to tree topology of the parsimony analysis. The entire subfamily Zygothylloideae is completely supported, with the *Tetraena* clade as the most basal group. The sister relationship of *Augea* and *Zygothylum stapffii* is completely supported. This group is embedded within the *Roepera* clade with complete support. In turn, this larger grouping is embedded within the *Zygothylum* clade with complete support which is again embedded within the *Fagonia* clade, but this is unsupported, with a posterior probability value of 69%.

3.3.5 **Phylogenetic analyses of the combined chloroplast genes and the nuclear ITS cassette**

3.3.5.1 *Parsimony (PAUP)*

The gene sequence matrix of the combined chloroplast and nuclear gene regions used contained 45 711 bp. The parsimony analysis sequence matrix parameters are shown in Table 3-12.

Table 3-12: The gene sequence matrix parameters of the combined chloroplast and nuclear gene sequences analysis.

Information of the parsimony analysis performed	
Total Characters	45711
Constant Characters	37744
Parsimony Uninformative Characters	4873 (10.66%)
Parsimony Informative Characters	3094 (6.77%)
Variability (Uninformative + Informative)	7967 (17.43%)
Number of Trees	1
Minimum Length	9296
Maximum Length	14180
Values of Tree	
Tree Length	11372
CI	0.817
RI	0.575

A single most parsimonious phylogenetic tree was retrieved from the analysis in Figure 3-4(d). The entire subfamily Zygophylloideae was completely supported with the *Tetraena* clade as the most basal group. The sister relationship between *Augea* and *Zygophyllum stapffii* is completely supported. This group is embedded within the *Roepera* clade with strong support. This larger group is embedded within the *Zygophyllum* clade with strong support. This larger group is embedded within the *Fagonia* clade, but is unsupported.

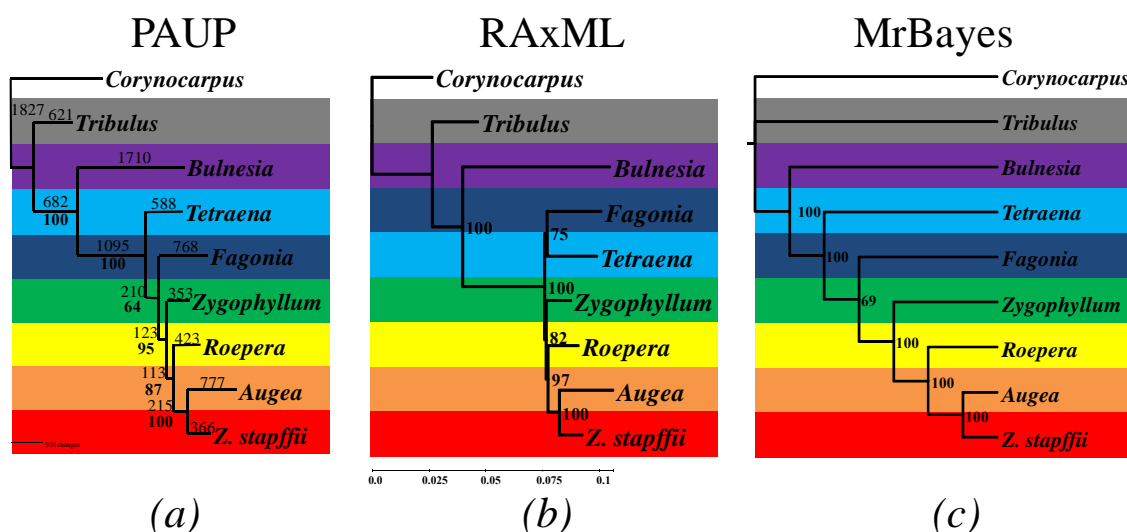
3.3.5.2 *RAxML (Maximum Likelihood)*

The single most likely phylogenetic tree of the combined chloroplast and nuclear gene regions is shown in Figure 3-4(e). The subfamily Zygophylloideae is completely supported. There is complete support for two groups within the subfamily within the subfamily Zygophylloideae. The first group contains the *Fagonia* clade and the *Tetraena* clade, which is strongly supported. The second group contains a completely supported sister relationship of *Augea* and *Zygophyllum stapffii*. This group is embedded within the *Roepera* clade with strong support which in turn is embedded within the *Zygophyllum* clade with strong support.

3.3.5.3 *MrBayes (Bayesian Inference)*

The majority-rule consensus tree of the Bayesian inference analyses are indicated in Figure 3-4(f). The subfamily Zygophylloideae is completely supported. There is complete support for two groupings within the subfamily Zygophylloideae. The first group contains the *Fagonia* clade and the *Tetraena* clade, which is completely supported. The second group contains a completely supported sister relationship of *Augea* and *Zygophyllum stapffii*. This group is embedded within

Photosynthetic and Non-Photosynthetic Genes



Photosynthetic, Non-Photosynthetic and Nuclear Genes

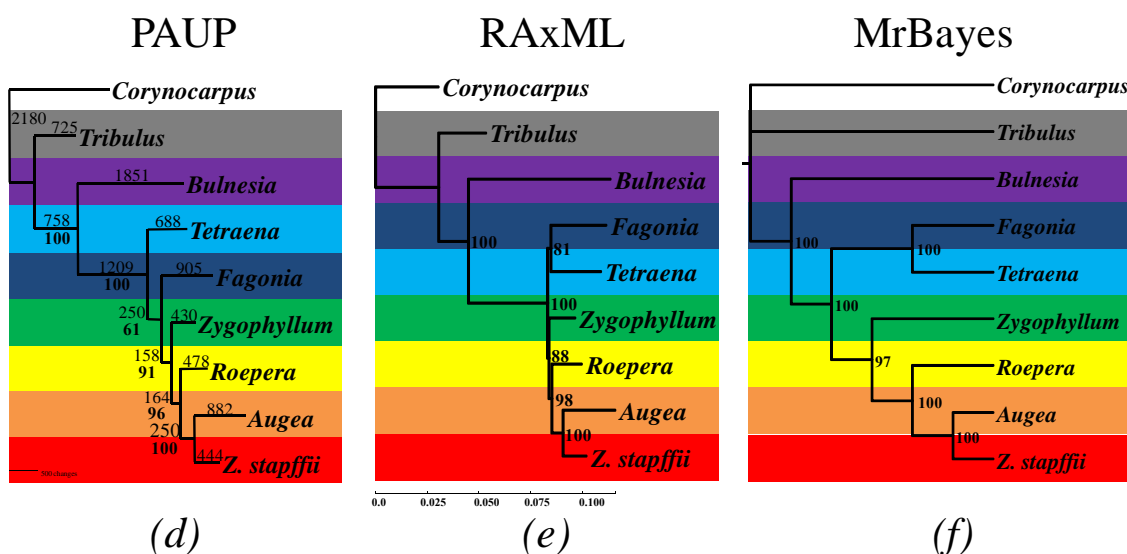


Figure 3-4: (a), The single most parsimonious phylogenetic tree of all the chloroplast gene regions, with bootstrap values indicated below the nodes. (b), The single most likely phylogenetic tree of the maximum likelihood analysis, of all gene regions, with proportional branch lengths and bootstrap support. (c), The majority-rule consensus tree of the combined stored trees from two independent Bayesian inference analyses, of all gene regions used, with posterior probabilities indicated. (d), The single most parsimonious phylogenetic tree of all the gene regions, with bootstrap values indicated below the nodes. (e), The single most likely phylogenetic tree of all the gene regions used, from the maximum likelihood analysis with proportional branch lengths and bootstrap support. (f), The majority-rule consensus tree of the combined stored trees from two independent Bayesian inference analyses, of all gene regions used, with posterior probabilities indicated.

the *Roepera* clade with complete support. This group is embedded within the *Zygophyllum* clade with very strong support.

3.3.6 Conflicting nuclear and chloroplast signal in Asian *Zygophyllum*

As can be seen from the phylogenetic trees of the chloroplast photosynthetic genes, the chloroplast non-photosynthetic genes and the nuclear genes region, all three groups of genes retrieved a different tree topology for the subfamily Zygophylloideae. Selection in the photosynthetic genes in order to conform to certain functional phenotypes in their respective photosynthetic mechanisms may have influenced the tree topology of the photosynthetic genes. Thus this tree topology may not reflect the true phylogenetic relationships between these taxa and consequently the decision was taken not to include these gene sequences in these analyses. In the phylogenetic analysis of the nuclear ITS cassette region the only difference in tree topology to that of the phylogenetic analysis of the non-photosynthetic gene tree topology was that the Asian *Zygophyllum* clade moved to the basal position in the subfamily Zygophylloideae (Figure 3-3 (h) compared to Figure 3-3 (e)). This was hypothesized as being either due to an ancient hybridization event in the evolution of the Asian *Zygophyllum* subfamily Zygophylloideae, or that rapid divergence within the subfamily resulted in incomplete lineage sorting.

An initial test was performed to assess which of these two hypotheses is more likely, but an in-depth study on all the variables could not be performed due to time constraints. The preliminary results of the initial analyses indicated that if the initial populations were between 10 000 and 100 000 that hybridization is the more likely hypothesis. If the population sizes were between 100 000 and 1 000 000 the more likely hypothesis is incomplete lineage sorting. As populations of common *Zygophyllum* species such as *Zygophyllum decumbens* and *Zygophyllum simplex* are larger than 1 000 000 (207), this gives support to the second hypothesis that incomplete lineage sorting gave rise to the conflicting signal from the nuclear as opposed to the chloroplast signal from the Asian *Zygophyllum*.

3.4 Discussion

The advent of NGS has considerably advanced the field of molecular systematics, but has highlighted some possible shortcomings in the analysis methods being performed on the sequence data. Massive amounts of data are generated at an unprecedented rate and can be used for very detailed phylogenetic analyses. This necessitates the need for phylogenetic analysis programs and the ability to handle the large amounts of data. In many other studies in which similar data were analysed by model-based methods in comparison to parsimony-based methods,

it has been concluded that the model-based phylogenetic inference methods perform better than parsimony-based methods (146, 211). This could be due to the fact that parsimony analyses search for the shortest tree to fit the data with no explicit model of evolution whilst model-based analyses make use of the best models to fit the data (146, 211).

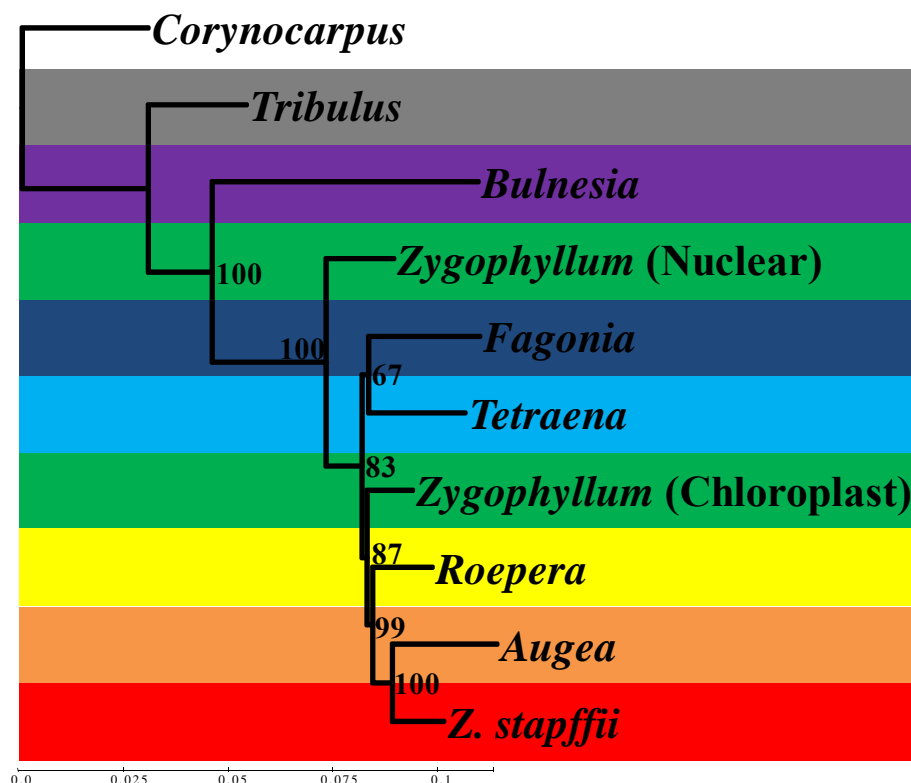


Figure 3-5: The maximum likelihood analysis phylogenetic tree retrieved after the *Zygophyllum fabago* nuclear and chloroplast gene sequences were separated from each other in order to test for conflict in the datasets.

If all of the analyses performed in this study are compared a common feature that becomes very clear is that the groups are separated by short internal branches on long terminal branches, which indicates that the groups within the subfamily Zygophylloideae have radiated from a common ancestor due to an ancient rapid radiation event. Parsimony analyses have been shown to be very sensitive to long branch attraction in such data sets (146, 211) and therefore appeared to have its shortcomings in this study. The same results as were obtained with the model-based analyses with the smaller data sub sets, were obtained with parsimony analysis, but when all the data were combined, parsimony did distort some of the nodes in the basal-most positions within the subfamily Zygophylloideae. For this reason, conclusions from the phylogenomic study should rather be made from the model-based phylogenetic inferences. Bellstedt *et al.* (2012) also showed ancient rapid radiation of members of the subfamily Zygophylloideae, in a dated phylogeny (16). Although the tree had a slightly different topology to the one found in this study, the radiation was estimated to coincide with the onset of aridification during the Miocene,

i.e. 20 to 18 Mya (217). It was postulated that during times of aridification species pre-adapted to these conditions could radiate rapidly into these newly formed climatological and ecological niches (15, 27).

The coalescence analysis in which the nuclear and the chloroplast gene data from *Zygophyllum fabago* were separated indicated that the Asian *Zygophyllum*, as a group, contains evidence of incomplete lineage sorting. Within the previous analyses of the subfamily Zygophylloideae utilizing the ITS region, all the taxa that are considered to be from the Asian *Zygophyllum* lineage grouped with *Zygophyllum fabago*, which indicates that *Zygophyllum fabago* is not a unique case in this clade regarding the difference in the nuclear and chloroplast gene signals in the phylogenetic analyses. This supports that the entire Asian *Zygophyllum* clade originated from a common ancestor and subsequently radiated into the Asian *Zygophyllum* group. This can also be interpreted to be evidence of an older rapid radiation in the subfamily.

A recent similar study attempted to elucidate the phylogenetic placement of the COM-clade (Celestrales-Oxalidales-Malpighiales) within the angiosperm phylogeny, using genes from the chloroplast, mitochondrial and nuclear genomes in maximum likelihood analyses. The study found that the 78-gene chloroplast gene matrix placed the COM-clade within the Fabidae, the 4-gene mitochondrial gene matrix placed the COM-clade in the Malvidae and the 5-gene nuclear gene matrix also placed the COM-clade in the Malvidae. The authors postulated that either ancient (121–108 Mya) incomplete lineage sorting or ancient hybridization was the possible cause, similar to what has been observed in this study (15, 198, 218).

Classical phylogenetic analysis cannot accommodate for recombination which is the result of hybridization or incomplete lineage sorting, it can only present a diverging phylogenetic hypothesis. To address this problem, Pirie *et al.*, in 2009 (154) have proposed the “taxon duplication” approach for the analysis of the relationships of a taxon showing recombination and this approach is gaining more and more acceptance in recent years (4, 74, 148). Although the taxon appears in two different positions in a phylogeny, it indicates the parentage of genes in the nuclear and chloroplast genome of that taxon. In the subfamily Zygophylloideae, the detection of incomplete lineage sorting at the molecular level supports what is observed in the morphology of characters such as seed dehiscence, seed mucilage and seed attachment of the major groups, i.e. different combinations of these characters in the different major groups in the subfamily as shown by Bellstedt *et al.*, in 2008 (17).

NGS technologies have been available for less than a decade and have made it possible to sequence entire chloroplast genomes of plants very rapidly and relatively cheaply. Since these technologies are still fairly new there are only a few hundred fully sequenced chloroplast

genomes available on Genbank and in most cases there are only single representative species for entire genera and even families (*Macademia integrifolia*, representing the genus *Macademia*, as well as the family *Proteaceae*). At this early stage of the NGS era studies are sacrificing the number of species being investigated for a drastic increase of phylogenetic markers.

One such study that has recently been published aimed to resolve the phylogenetic relationships of the commelinid clade in the Angiosperm phylogeny, in 2013 (11). This clade contains the orders Arecales, Commelinales, Poales, Zingiberales and an unplaced family Dasypogonaceae. This study utilized 83 plastid genes sequenced from representative species within the commelinids to resolve the basal relationships of this clade. This study showed that these lineages diverged long ago and rapidly (short internal nodes and long branches). This study could however not resolve the basal most node within the commelinid clade, but did show a sister relationship of Arecales with Dasypogonaceae, as well as a sister relationship of Poales with Zingiberales/Commelinales which were unresolved in previous studies in which the sequences of a limited number of markers were used for phylogenetic inference (11). Phylogenies based on complete or nearly complete chloroplast gene sets of selected single representative species have also recently been used to resolve the relationships in the entire Viridiplantae (168).

Previous studies on the subfamily Zygophylloideae consistently retrieved large strongly supported groupings i.e. the subgenus based on chloroplast and based on ITS region phylogenies *Agrophyllum*/genus *Tetraena*; Asian subgenus *Zygophyllum*/*Zygophyllum sensu stricto*; genus *Fagonia* and *Melocarpum*; southern African and Australian members of subgenus *Zygophyllum*/*Roepera*; the species *Augea capensis* and *Zygophyllum stapffii/orbiculatum*. The species within these groupings never moved from one grouping into another, which indicated that they are strongly associated with their respective clades. Thus choosing a specific single representative species from each of these subgroupings should not affect the overall topology as these species are strongly associated/embedded in their respective groupings. The phylogenomic study also utilizes genes from the exact same compartments as those from the previous studies (>40 chloroplast markers as opposed to two from Bellstedt *et al.*, 2008(17)/ entire ITS cassette as opposed to the ITS region only (Chapter 2)).

The phylogenetic relationships within the subfamily Zygophylloideae have been a point of contention for many years, as neither morphological nor single-gene molecular studies, previously, had enough information to resolve their phylogenetic relationships. This study has therefore brought much needed clarity to the phylogenetic history of the subfamily Zygophylloideae.

4 Conclusions and Future Perspectives

From the sequence evidence as well as phylogenetic analyses of the ITS region it can be concluded that *Zygophyllum orbiculatum* and *Zygophyllum stapffii* are conspecific. This is in agreement with the very similar morphology of the two species with the exception that *Zygophyllum orbiculatum* possesses unifoliolate leaves and the *Zygophyllum stapffii* possesses bifoliolate leaves. The taxonomic implications of this study are that the two species should be combined. The earliest name should have precedence regarding the naming of this taxon. This means *Zygophyllum stapffii* Schinz should be synonymized with *Zygophyllum orbiculatum* Welwitsch ex. Oliver.

As in previous studies, the ITS region was unable to resolve the basal relationships within the subfamily Zygophylloideae. It did however retrieve a monophyletic subfamily Zygophylloideae just as in previous studies (13, 17). From this it was concluded that more sequence information would be required in order to resolve the phylogenetic relationships within the subfamily Zygophylloideae.

A phylogenomic approach was used in an attempt to resolve the basal relationships within the subfamily in which next generation sequencing aimed at sequencing the chloroplast genome of representatives of the large groupings of the Zygophylloideae and outgroups was used. In this approach, sequence data of 21 photosynthetic protein-coding genes, 20 non-photosynthetic gene sequences (16 protein-coding sequences, 4 RNA-coding genes and 2 intron sequences) and the nuclear ITS cassette (3 RNA-coding genes and two intergenic spacers) was used to try to resolve the phylogenetic relationships of between the groupings within the subfamily Zygophylloideae. The topology of the tree obtained from the parsimony analysis showed different and less supported basal relationships, which was interpreted to possibly be the result of long branch attraction. In contrast, the two model-based phylogenetic inference methods, maximum likelihood and Bayesian inference, retrieved identical and strongly supported phylogenetic relationships within the subfamily Zygophylloideae. For these reasons further conclusions were only made from the model-based phylogenetic inferences. The short internal branches of the major groupings in the subfamily Zygophylloideae appear to indicate that the lineages diverged rapidly in an older radiation from a common ancestor. *Tetraena* was retrieved sister to the *Fagonia* and *Melocarpum* clade and not as member of a monophyletic genus *Zygophyllum*. *Augea capensis* and *Zygophyllum orbiculatum/stapffii* were retrieved as sister taxa which were embedded within *Roepera*. The ITS cassette analyses from the phylogenomic study, when compared to the combined chloroplast analyses, showed conflict in the position of the Asian

Zygophyllum. This was addressed by using a “taxon duplication” approach. Preliminary tests for incongruence indicated that this could be ascribed to incomplete lineage sorting rather than hybridization.

It is important to note that the incomplete lineage sorting detected in the phylogenomic analysis in *Zygophyllum fabago* is representative of the whole group of Asian *Zygophyllum* as this taxon and the other species from Asia appear in a single group in the ITS region phylogenetic analyses (as presented in Chapter 2).

These results have certain taxonomic implications for the subfamily. Beier *et al.* (2003) proposed the recircumscription of the subgenus *Agrophyllum* into a new genus *Tetraena* which is supported (13). This study showed that the Asian *Zygophyllum* have a unique combination of nuclear and chloroplast genes. This unique attribute, which is not prevalent in any of the other groupings within the subfamily Zygophylloideae, necessitates that this group should be classified as a unique genus with the name *Zygophyllum* as the type species, *Zygophyllum fabago* as described by Linnaeus in 1753, is found in this group (102). Although Beier *et al.*, (2003) (13) did not detect the incomplete lineage sorting within the Asian *Zygophyllum*, this study does support his conclusion that this group should be retained as genus *Zygophyllum*. In the phylogenomic analysis, *Augea* was found to be sister to the taxon *Zygophyllum orbiculatum/stapffii* and is embedded within the *Roepera* clade. However, *Zygophyllum orbiculatum/stapffii* is morphologically distinct from *Augea capensis* and both are distinct from *Roepera*. This necessitates that *Zygophyllum orbiculatum/stapffii* should be reclassified as a monotypic genus and be given a new and unique binomial name to distinguish it from both the genus *Roepera* and the genus *Augea*. This study further confirms that the taxonomic placements by Beier *et al.*, in 2003, of *Zygophyllum stapffi* in *Tetraena* and *Zygophyllum orbiculatum* in *Roepera* are not valid (13).

Future perspectives entail investigating all the protein-coding genes, especially the photosynthetic protein-coding genes, for selection using the Ka/Ks ratio analyses, as these genes might be under selection due to different photosynthetic phenotypes (C₃, C₄ and CAM as discussed before). A dated phylogenetic investigation should also be performed using BEAST in order to ascertain when the rapid divergence within the subfamily Zygophylloideae occurred.

5 References

1. **Al-Turki TA, Filfilan SA, Mehmood SF.** 2000. A cytological study of flowering plants from Saudi Arabia. *Willdenowia*. 30:339–358
2. **Álvarez I, Wendel JF.** 2003. Ribosomal ITS sequences and plant phylogenetic inference. *Mol. Phylogenet. Evol.* 29(3):417–434
3. **Ansorge WJ.** 2009. Next-generation DNA sequencing techniques. *N. Biotechnol.* 25(4):195–203
4. **Antonelli A, Humphreys AM, Lee WG, Linder HP.** 2011. Absence of mammals and the evolution of New Zealand grasses. *Proc. Biol. Sci.* 278:695–701
5. **Aoki S, Uehara K, Imafuku M, Hasebe M, Ito M.** 2004. Phylogeny and divergence of basal angiosperms inferred from *APETALA3*- and *PISTILLATA*-like *MADS*-box genes. *J. Plant Res.* 117(3):229–244
6. **Ascherson P, Beyer R, Friedel E, Jacobasch E, Kärnbach L, et al.** 1888. *Verhandlungen des Botanischen Vereins für die Provinz Brandenburg*. Berlin: R. Gaertner's verlagbuchhandlung (Herman Heyfelder)
7. **Atherton RA, McComish BJ, Shepherd LD, Berry LA, Albert NW, Lockhart PJ.** 2010. Whole genome sequencing of enriched chloroplast DNA using the Illumina GAI platform. *Plant Methods*. 6(22):
8. **Bailey CD, Doyle JJ.** 1999. Potential phylogenetic utility of the low-copy nuclear gene *pistillata* in dicotyledonous plants: comparison to nrDNA ITS and *trnL* intron in *Sphaerocardamum* and other Brassicaceae. *Mol. Phylogenet. Evol.* 13(1):20–30
9. **Baldwin BG.** 1992. Phylogenetic utility of the internal transcribed spacers of nuclear ribosomal DNA in plants: an example from the compositae. *Mol. Phylogenet. Evol.* 1(1):3–16
10. **Baldwin BG, Sanderson MJ, Porter JM, Wojciechowski MF, Campbell CS, et al.** 1995. The ITS region of nuclear ribosomal DNA: A valuable source of evidence on Angiosperm phylogeny. *Ann. Missouri Bot. Gard.* 82(2):247–77
11. **Barrett CF, Davis JI, Leebens-Mack J, Conran JG, Stevenson DW.** 2013. Plastid genomes and deep relationships among the commelinid monocot angiosperms. *Cladistics*. 29(1):65–87
12. **Barrett CF, Specht CD, Leebens-Mack J, Stevenson DW, Zomlefer WB, Davis JI.** 2014. Resolving ancient radiations: can complete plastid gene sets elucidate deep relationships among the tropical gingers (Zingiberales)? *Ann. Bot.* 113(1):119–133
13. **Beier B-A, Chase MW, Thulin M.** 2003. Phylogenetic relationships and taxonomy of subfamily Zygophylloideae (Zygophyllaceae) based on molecular and morphological data. *Plant Syst. Evol.* 240(1-4):11–39

14. **Beier B-A, Nylander JAA, Chase MW, Thulin M.** 2004. Phylogenetic relationships and biogeography of the desert plant genus *Fagonia* (Zygophyllaceae), inferred by parsimony and Bayesian model averaging. *Mol. Phylogenet. Evol.* 33(1):91–108
15. **Bell CD, Soltis DE, Soltis PS.** 2010. The age and diversification of the angiosperms re-revisited. *Am. J. Bot.* 97(8):1296–1303
16. **Bellstedt DU, Galley C, Pirie MD, Linder HP.** 2012. The Migration of the Palaeotropical Arid Flora: Zygophylloideae as an Example. *Syst. Bot.* 37(4):951–959
17. **Bellstedt DU, van Zyl L, Marais EM, Bytebier B, de Villiers CA, et al.** 2008. Phylogenetic relationships, character evolution and biogeography of southern African members of *Zygophyllum* (Zygophyllaceae) based on three plastid regions. *Mol. Phylogenet. Evol.* 47(3):932–949
18. **Bentham G, Hooker JD.** 1862. *Genera plantarum: ad exemplaria imprimis in Herbariis Kewensibus servata definita / auctoribus G. Bentham et J.D. Hooker.* Londini: A. Black,
19. **Bergthorsson U, Adams KL, Thomason B, Palmer JD.** 2003. Widespread horizontal transfer of mitochondrial genes in flowering plants. *Nature.* 424(6945):197–201
20. **Borgstro E, Lundin S, Lundeberg J.** 2011. Large Scale Library Generation for High Throughput Sequencing. *PLoS One.* 6(4):4–9
21. **Bousquet J, Strauss SH, Doerksen AH, Price RA.** 1992. Extensive variation in evolutionary rate of *rbcL* gene sequences among seed plants. *Proc. Natl. Acad. Sci. U. S. A.* 89:7844–7848
22. **Bousquet J, Strauss SH, Li P.** 1992. Complete Congruence between Morphological and *rbcl*-based Molecular Phylogenies in Birches and Related Species (Betulaceae). *Mol. Biol. Evol.* 9(6):1076–1088
23. **Bremer B, Bremer K, Chase MW, Reveal JL, Soltis DE, et al.** 2003. An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG II. *Bot. J. Linn. Soc.* 141(4):399–436
24. **Brown R.** 1814. General remarks, geographical and systematical on the botany of Terra Australis. In *A voyage to Terra Australis* 2, ed M Flinders
25. **Bult C, Kdillersjo M, Suh Y.** 1992. Amplification and Sequencing of 16/18S rDNA from Gel-Purified Total Plant DNA. *Plant Mol. Biol. Report.* 10(3):273–284
26. **Bytebier B, Antonelli A, Bellstedt DU, Linder HP.** 2011. Estimating the age of fire in the Cape flora of South Africa from an orchid phylogeny. *Proc. Biol. Sci.* 278(1703):188–195
27. **Cantino PD, Doyle JA, Graham SW, Judd WS, Olmstead RG, et al.** 2007. Towards a Phylogenetic Nomenclature of Tracheophyta. *Taxon.* 56(3):822
28. **Cardona G, Rosselló F, Valiente G.** 2008. Extended Newick: it is time for a standard representation of phylogenetic networks. *BMC Bioinformatics.* 9:532

29. **Carlquist S.** 2005. Wood anatomy of Krameriaceae with comparisons with Zygophyllaceae: phyletic, ecology and systematics. *Bot. J. Linn. Soc.* 149:257–270
30. **Chase MW, Fay MF, Reveal JL, Soltis DE, Soltis PS, et al.** 2009. An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG III. *Bot. J. Linn. Soc.* 161:105–121
31. **Chase MW, Reveal JL, Hortoria LHB, Biology P, Building M.** 2009. A phylogenetic classification of the land plants to accompany APG III. *Bot. J. Linn. Soc.* 161:122–127
32. **Chase MW, Soltis DE, Olmstead RG, Morgan D, Donald H, et al.** 1993. Phylogenetics of Seed Plants: An Analysis of Nucleotide Sequences from the Plastid Gene *rbcL*. *Ann. Missouri Bot. Gard.* 80(3):528–580
33. **Chippindale PT, Wiens JJ.** 2005. Re-evolution of the larval stage in the plethodontid salamander genus *Desmognathus*. *Herpetol. Rev.* 36(2):113–117
34. **Christin P-A, Salamin N, Muasya AM, Roalson EH, Russier F, Besnard G.** 2008. Evolutionary switch and genetic convergence on *rbcL* following the evolution of C₄ photosynthesis. *Mol. Biol. Evol.* 25(11):2361–2368
35. **Collin R, Miglietta MP.** 2008. Reversing opinions on Dollo’s Law. *Trends Ecol. Evol.* 23(September):602–9
36. **Corriveau JL, Coleman AW.** 1988. Rapid Screening Method to Detect Potential Biparental Inheritance of Plastid DNA and Results for Over 200 Angiosperm Species. *Am. J. Bot.* 75(10):1443–58
37. **Cristina Acosta M, Premoli AC.** 2010. Evidence of chloroplast capture in South American *Nothofagus* (subgenus *Nothofagus*, Nothofagaceae). *Mol. Phylogenet. Evol.* 54(1):235–242
38. **Crookston RK, Moss DN.** 1972. C-4 and C-3 Carboxylation Characteristics in the genus *Zygophyllum* (Zygophyllaceae). *Missouri Bot. Gard.* 59(3):465–470
39. **Davies K.** 2010. *It’s “Watson Meets Moore” as Ion Torrent Introduces Semiconductor Sequencing.* Bio-IT World. <http://www.bio-itworld.com/news/03/01/10/ion-torrent-semiconductor-sequencing.html>
40. **De Candolle AP.** 1824. *Prodromus systematis naturalis regni vegetabilis, sive, Enumeratio contracta ordinum generum specierumque plantarum huc usque cognitatum, juxta methodi naturalis, normas digesta /auctore Aug. Pyramo de Candolle.* Parisii: Sumptibus Sociorum Treuttel et Wurtz,
41. **Dollo L.** 1893. The Laws of Evolution. *Bull. la Société Belge Géologie*, pp. 164–166
42. **Dong W, Liu J, Yu J, Wang L, Zhou S.** 2012. Highly variable chloroplast markers for evaluating plant phylogeny at low taxonomic levels and for DNA barcoding. *PLoS One.* 7(4):1–9

43. **Doyle JA, Endress PK.** 2000. Morphological Phylogenetic Analysis of Basal Angiosperms: Comparison and Combination with Molecular Data. *Int. J. Plant Sci.* 161(S6):S121–S153
44. **Droege M, Hill B.** 2008. The Genome Sequencer FLX System - Longer reads, more applications, straight forward bioinformatics and more complete data sets. *J. Biotechnol.* 136:3–10
45. **Drummond AJ, Suchard MA, Xie D, Rambaut A.** 2012. Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Mol. Biol. Evol.* 29(8):1969–1973
46. **Dubcovsky J, Dvorák J.** 1995. Ribosomal RNA multigene loci: nomads of the Triticeae genomes. *Genetics.* 140(4):1367–1377
47. **Duvall MR, Robinson JW, Mattson JG, Moore A.** 2008. Phylogenetic analyses of two mitochondrial metabolic genes sampled in parallel from angiosperms find fundamental interlocus incongruence. *Am. J. Bot.* 95(7):871–884
48. **Dyer RA.** 1975. *The genera of Southern African Flowering Plants*. Department of Agricultural Technical Services, Pretoria. Volume 1 ed.
49. **Eid J, Fehr A, Gray J, Luong K, Lyle J, et al.** 2009. Real-time DNA sequencing from single polymerase molecules. *Science* (80-.). 323:133–138
50. **El Hadidi MN.** 1975. Zygophyllaceae in Africa. *Boissiera.* 24:317–323
51. **El Hadidi MN.** 1977. Tribulaceae as a distinct family. *Publ. from Cairo Univ. Herb.* 7. 8:103–108
52. **El Hadidi MN.** 1980. On the taxonomy of *Zygophyllum* section Bipartita. *Kew Bull.* 35(2):335–339
53. **El Hadidi MN.** 1985. Zygophyllaceae. In *Flora of Tropical East Africa*. A. A. Balkema
54. **Emshwiller E, Doyle JJ.** 1999. Chloroplast-expressed glutamine synthetase (ncpGS): potential utility for phylogenetic studies with an example from *Oxalis* (Oxalidaceae). *Mol. Phylogenet. Evol.* 12(3):310–319
55. **Endlicher SL.** 1841. Part 18: 1161. In *Genera Plantarum secundum ordines naturales disposita*. Fr. Beck, Vienna
56. **Engler A.** 1915. Zygophyllaceae. In *Die Pflazenwelt Afrikas. Die Vegetation der Erde.*, ed A Engler, CGO Drude, pp. 729–45
57. **Engler A.** 1931. Zygophyllaceae. In *Die Natürliche Pflanzenfamilien*. Vol. 19a: 144–84. Section 2. 457–58. Leipzig: Engelmann. 2nd ed.
58. **Fehrer J, Gemeinholzer B, Chrtek J, Bräutigam S.** 2007. Incongruent plastid and nuclear DNA phylogenies reveal ancient intergeneric hybridization in *Pilosella* hawkweeds (*Hieracium*, Cichorieae, Asteraceae). *Mol. Phylogenet. Evol.* 42:347–361

59. **Felsenstein J.** 1988. Phylogenies from molecular sequences: inference and reliability. *Annu. Rev. Genet.* 22:521–565
60. **Fortune PM, Schierenbeck KA, Ainouche AK, Jacquemin J, Wendel JF, Ainouche ML.** 2007. Evolutionary dynamics of Waxy and the origin of hexaploid *Spartina* species (Poaceae). *Mol. Phylogenet. Evol.* 43(3):1040–1055
61. **Fuchs J, Irestedt M, Fjeldså J, Couloux A, Pasquet E, Bowie RCK.** 2012. Molecular phylogeny of African bush-shrikes and allies: Tracing the biogeographic history of an explosive radiation of corvoid birds. *Mol. Phylogenet. Evol.* 64(1):93–105
62. **Galis F, Arntzen JW, Lande R.** 2010. Dollo’s law and the irreversibility of digit loss in *Bachia*. *Evolution (N. Y.)*. 64:2466–76
63. **Gaut BS, Muse S V., Clegg MT.** 1993. Relative rates of Nucleotide Substitution in the Chloroplast Genome. *Mol. Phylogenet. Evol.* 2(2):89–96
64. **Goldberg EE, Igić B.** 2008. On phylogenetic tests of irreversible evolution. *Evolution (N. Y.)*. 62:2727–2741
65. **Granot G, Grafi G.** 2014. Epigenetic information can reveal phylogenetic relationships within Zygothallales. *Plant Syst. Evol.*
66. **Grob GB., Gravendeel B, Eurlings MC.** 2004. Potential phylogenetic utility of the nuclear *FLORICAULA/LEAFY* second intron: comparison with three chloroplast DNA regions in *Amorphophallus* (Araceae). *Mol. Phylogenet. Evol.* 30(1):13–23
67. **Hansen AK, Escobar LK, Gilbert LE, Jansen RK.** 2007. Paternal, Maternal, And Biparental Inheritance Of The Chloroplast Genome In *Passiflora* (Passifloraceae): Implications For Phylogenetic Studies. *Am. J. Bot.* 94(1):42–46
68. **Haston E, Richardson JE, Stevens PF, Chase MW, Harris DJ, et al.** 2009. The Linear Angiosperm Phylogeny Group (LAPG) III: a linear sequence of the families in APG III. *Bot. J. Linn. Soc.* 161:128–131
69. **Hollingsworth PM.** 2011. Refining the DNA barcode for land plants. *Proc. Natl. Acad. Sci. U. S. A.* 108(49):19451–19452
70. **Huang S, Sirikhachornkit A, Faris JD, Su X, Gill BS, et al.** 2002. Phylogenetic analysis of the acetyl-CoA carboxylase and 3-phosphoglycerate kinase loci in wheat and other grasses. *Plant Mol. Biol.* 48(5-6):805–820
71. **Huang Y-Y, Matzke AJM, Matzke M.** 2013. Complete Sequence and Comparative Analysis of the Chloroplast Genome of Coconut Palm (*Cocos nucifera*). *PLoS One*. 8(8):1–12
72. **Hudson ME.** 2008. Sequencing breakthroughs for genomic ecology and evolutionary biology. *Mol. Ecol. Resour.* 8(1):3–17
73. **Huelsenbeck JP, Ronquist F.** 2001. MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics*. 17(8):754–755

74. **Humphreys AM, Antonelli A, Pirie MD, Linder HP.** 2011. Ecology and evolution of the diaspore “burial syndrome.” *Evolution (N. Y)*. 65:1163–1180
75. **Hunziker JH, Comas C.** 2002. *Larrea* interspecific hybrids revisited (Zygophyllaceae). *Darwiniana*. 40(1-4):33–38
76. Illumina Sequencing Technology. 2010
77. Illumina Sequencing Technology - Highest data accuracy, simple workflow, and a broad range of applications. 2010. Illumina, Inc
78. *Ion PGM™ Sequencer Specifications*. 2014.
<http://www.lifetechnologies.com/za/en/home/life-science/sequencing/next-generation-sequencing/ion-torrent-next-generation-sequencing-workflow/ion-torrent-next-generation-sequencing-run-sequence/ion-pgm-system-for-next-generation-sequencing/ion-pgm-system-specifications.html>
79. **Ismael AMA.** 1983. Some factors controlling the water economy of *Zygophyllum qatarense* growing in Qatar. *J. Arid Environ*. 6:239–46
80. **Jansen RK, Michaels HJ, Wallace RS, Kim K-J, Keeley SC, et al.** 1992. Chloroplast DNA Variation in the Asteraceae: Phylogenetic and Evolutionary Implications. In *Molecular Systematics of Plants*, ed PS Soltis, DE Soltis, JJ Doyle, pp. 252–79. Boston, MA: Springer US
81. **Johnson J.** 2012. *Ion Torrent Hits 400bp Read Length Mark...Why we're excited*. EdgeBio. <http://www.edgebio.com/ion-torrent-hits-400bp-read-length-markwhy-were-excited>
82. **Johnson LA, Soltis DE, Botany SS, Mar NJ.** 1994. MatK DNA Sequences and Phylogenetic Reconstruction in Saxifragaceae s. str.. . 19(1):143–156
83. **Jorgensen RA, Cluster PD.** 1988. Modes and tempos in the evolution of nuclear ribosomal DNA: New characters for evolutionary studies and new markers for genetic and population studies. *Ann. Missouri Bot. Gard*. 75(4):1238–1247
84. **Judd WS.** 2008. *Plant Systematics: A Phylogenetic Approach*. W. H. Freeman. 3rd, illustr ed.
85. **Kadereit G, Freitag H.** 2011. Molecular phylogeny of Camphorosmeae (Camphorosmoideae, Chenopodiaceae): Implications for biogeography , evolution of C₄ - photosynthesis and taxonomy. . 60(February):51–78
86. **Kare Bremer, Chase MW, Stevens PF, Anderberg AA, Backlund A, et al.** 1998. An ordinal classification for the families of flowering plants. *Ann. Missouri Bot. Gard*. 85(4):531–553
87. **Karow J.** 2010. *Ion Torrent Systems Presents \$50,000 Electronic Sequencer at AGBT*. In Sequence. <http://www.genomeweb.com/sequencing/ion-torrent-systems-presents-50000-electronic-sequencer-agbt>

88. **Kim JS, Kim J-H.** 2013. Comparative Genome Analysis and Phylogenetic Relationship of Order Liliales Insight from the Complete Plastid Genome Sequences of Two Lilies (*Lilium longiflorum* and *Alstroemeria aurea*). *PLoS One*. 8(6):e68180
89. **Kircher M, Kelso J.** 2010. High-throughput DNA sequencing-concepts and limitations. *BioEssays*. 32(6):524–536
90. **Knoop V.** 2004. The mitochondrial DNA of land plants: peculiarities in phylogenetic perspective. *Curr. Genet.* 46(3):123–139
91. **Kohlsdorf T, Wagner GP.** 2006. Evidence for the reversibility of digit loss: a phylogenetic study of limb evolution in *Bachia* (Gymnophthalmidae: Squamata). *Evolution*. 60(9):1896–1912
92. **Korlach J, Bjornson KP, Chaudhuri BP, Cicero RL, Flusberg BA, et al.** 2010. Real-time DNA sequencing from single polymerase molecules. In *Methods in enzymology*, ed NG Walter. 472(10):431–55. Elsevier Inc. Volume 472 ed.
93. **Kubitzki K.** 2007. *The families and genera of vascular plants*. Springer. Volume 9 ed.
94. **Kuzoff RK, Sweere JA, Soltis DE, Soltis PS, Zimmer EA.** 1998. The phylogenetic potential of entire 26S rDNA sequences in plants. *Mol. Biol. Evol.* 15(3):251–263
95. **Laport RG, Minckley RL, Ramsey J.** 2012. Phylogeny and Cytoecography of the North American Creosote Bush (*Larrea tridentata*, Zygophyllaceae). *Syst. Bot.* 37(1):153–164
96. **Lemey P, Salemi M, Vandamme A-M.** 2009. *The Phylogenetic Handbook A Practical Approach to Phylogenetic Analysis and Hypothesis Testing*. New York, NY: Cambridge University Press. Second Ed.
97. **Levene MJ, Korlach J, Turner SW, Foquet M, Craighead HG, Webb WW.** 2003. Zero-mode waveguides for single-molecule analysis at high concentrations. *Science*. 299(5607):682–86
98. **Lewis CE, Doyle JJ.** 2001. Phylogenetic utility of the nuclear gene malate synthase in the palm family (Arecaceae). *Mol. Phylogenet. Evol.* 19(3):409–420
99. **Li D-Z, Gao L-M, Li H-T, Wang H, Ge X-J, et al.** 2011. Comparative analysis of a large dataset indicates that internal transcribed spacer (ITS) should be incorporated into the core barcode for seed plants. *Proc. Natl. Acad. Sci. U. S. A.* 108(49):19641–19646
100. **Linder CR, Goertzen LR, Heuvel B V, Francisco-Ortega J, Jansen RK.** 2000. The complete external transcribed spacer of 18S-26S rDNA: Amplification and phylogenetic utility at low taxonomic levels in Asteraceae and closely allied families. *Mol. Phylogenet. Evol.* 14(2):285–303
101. **Lindley J.** 1853. *The vegetable kingdom; or, the structure, classification, and uses of plants, illustrated upon the natural system*. London: Bradbury and Evans. 3rd ed. ed.
102. **Linnaeus C.** 1753. *Caroli Linnaei ... Species plantarum: exhibentes plantas rite cognitatas, ad genera relatas, cum differentiis specificis, nominibus trivialibus, synonymis selectis, locis natalibus, secundum systema sexuale digestas...* Holmiae: Impensis Laurentii Salvii,

103. **Liu L, Li Y, Li S, Hu N, He Y, et al.** 2012. Comparison of next-generation sequencing systems. *J. Biomed. Biotechnol.* 2012:251364
104. **Löve Á.** 1979. IOPB Chromosome Number Reports LXIV. *Taxon.* 28(4):391–408
105. **Ma H, Yanofsky MF, Meyerowitz EM.** 1990. Molecular cloning and characterization of GPA1, a G protein alpha subunit gene from *Arabidopsis thaliana*. *Proc. Natl. Acad. Sci. U. S. A.* 87(10):3821–25
106. **Maddison WP, Maddison DR.** 2014. Mesquite: a modular system for evolutionary analysis. <http://mesquiteproject.org>
107. **Maia VH, Gitzendanner MA, Soltis PS, Wong GK-S, Soltis DE.** 2014. Angiosperm Phylogeny Based on 18S/26S rDNA Sequence Data: Constructing a Large Data Set Using Next-Generation Sequence Data. *Int. J. Plant Sci.* 175(6):613–50
108. **Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, et al.** 2005. Genome Sequencing in Open Microfabricated High Density Picoliter Reactors - Supplementary Methods: Library Preparation. *Nature.* 437(7057):1–34
109. **Marshall CR, Raff EC, Raff R a.** 1994. Dollo's law and the death and resurrection of genes. *Proc. Natl. Acad. Sci. U. S. A.* 91(December):12283–12287
110. **Mason-Gamer RJ, Weil CF, Kellogg EA.** 1998. Granule-bound starch synthase: structure, function, and phylogenetic utility. *Mol. Biol. Evol.* 15(12):1658–1673
111. **Mathews S, Spangler RE, Mason-gamer RJ, Kellogg EA.** 2002. Phylogeny of Andropogoneae inferred from Phytochrome B, BGSSI and ndhF. *Int. J. Plant Sci.* 163(3):441–50
112. **Matimati I, Musil CF, Raitt L, February EC.** 2012. Diurnal stem diameter variations show CAM and C₃ photosynthetic modes and CAM–C₃ switches in arid South African succulent shrubs. *Agric. For. Meteorol.* 161:72–79
113. **Maxam a. M, Gilbert W.** 1977. A new method for sequencing DNA. *Proc. Natl. Acad. Sci.* 74(2):560–564
114. **McCauley R a, Cortés-Palomec AC, Oyama K.** 2008. Isolation, characterization, and cross-amplification of polymorphic microsatellite loci in *Guaiacum coulteri* (Zygophyllaceae). *Mol. Ecol. Notes.* 8(3):671–674
115. **McCormack JE, Hird SM, Zellmer AJ, Carstens BC, Brumfield RT.** 2013. Applications of next-generation sequencing to phylogeography and phylogenetics. *Mol. Phylogenet. Evol.* 66(2):526–538
116. **Metzker ML.** 2008. Sequencing technologies — the next generation. *Nat. Rev. Genet.*
117. **Metzker ML.** 2010. Sequencing technologies - the next generation. *Nat. Rev. Genet.* 11(1):31–46

118. **Meyer M, Kircher M.** 2010. Illumina sequencing library preparation for highly multiplexed target capture and sequencing. *Cold Spring Harb. Protoc.* 2010(6):pdb.prot5448
119. **Michaels HJ, Scott KM, Olmstead RG, Szaro T, Robert K, et al.** 1993. Interfamilial Relationships of the Asteraceae: Insights from *rbcL* Sequence variation. *Ann. Missouri Bot. Gard.* 80:742–751
120. **Michener CD, Corliss JO, Cowan RS, Raven PH, Sabrosky CW, et al.** 1970. *Systematics In Support of Biological Research*. Washington, D.C.: Division of Biology and Agriculture
121. **Miller MA, Pfeiffer W, Schwartz T.** 2010. Creating the CIPRES Science Gateway for Inference of Large Phylogenetic Trees.
122. **Mishra MK, Suresh N, Bhat AM, Suryaprakash N, Kumar SS, et al.** 2011. Genetic molecular analysis of *Coffea arabica* (Rubiaceae) hybrids using SRAP markers. *Rev. Biol. Trop.* 59(June):607–617
123. **Mooney HA, Troughton JH, Berry JA.** 1977. Carbon Isotope Ratio Measurements of Succulent Plants in Southern Africa*. *Oecologia.* 30:295–305
124. **Moore MJ, Bell CD, Soltis PS, Soltis DE.** 2007. Using plastid genome-scale data to resolve enigmatic relationships among basal angiosperms. *Proc. Natl. Acad. Sci. U. S. A.* 104(49):19363–19368
125. **Moore MJ, Dhingra A, Soltis PS, Shaw R, Farmerie WG, et al.** 2006. Rapid and accurate pyrosequencing of angiosperm plastid genomes. *BMC Plant Biol.* 6:17
126. **Moore MJ, Soltis PS, Bell CD, Burleigh JG, Soltis DE.** 2010. Phylogenetic analysis of 83 plastid genes further resolves the early diversification of eudicots. *Proc. Natl. Acad. Sci. U. S. A.* 107(10):4623–4628
127. **Morey M, Fernández-Marmiesse A, Castiñeiras D, Fraga JM, Couce ML, Cocho JA.** 2013. A glimpse into past, present, and future DNA sequencing. *Mol. Genet. Metab.* 110:3–24
128. **Morgan DR, Soltis DE.** 1993. Phylogenetic Relationships among members of Saxifragaceae *sensu lato* based on *rbcL* Sequence data. *Ann. Missouri Bot. Gard.* 80:631–660
129. **Morozova O, Marra MA.** 2008. Applications of next-generation sequencing technologies in functional genomics. *Genomics.* 92(5):255–264
130. **Mosher JJ, Bernberg EL, Shevchenko O, Kan J, Kaplan L a.** 2013. Efficacy of a 3rd generation high-throughput sequencing platform for analyses of 16S rRNA genes from environmental samples. *J. Microbiol. Methods.* 95(2):175–181
131. **Muhaidat R, Sage RF, Dengler NG.** 2007. Diversity of Kranz anatomy and biochemistry in C₄ eudicots. *Am. J. Bot.* 94(3):362–381

132. **Muller H.** 1939. Reversibility in Evolution Considered From the Standpoint of Genetics. *Biol. Rev.* 14(December 1937):261–280
133. **Myllykangas S, Buenrostro J, Ji HP.** 2012. Overview of Sequencing Technology Platforms. In *Bioinformatics for High Throughput Sequencing.*, ed N Rodríguez-Ezpeleta, M Hackenberg, AM Aransay, pp. 11–26. Springer
134. **New England BioLabs® Inc.** 2014. NEBNext® for Ion Torrent™ - Library Preparation Kits
135. *New Products: PacBio's RS II; Cufflinks.* 2013. In Sequence. <http://www.genomeweb.com/sequencing/new-products-pacbios-rs-ii-cufflinks>
136. **Nyrén P.** 2007. The history of pyrosequencing. *Methods Mol. Biol.* 373(3):1–14
137. **Oh S-H, Potter D.** 2003. Phylogenetic utility of the second intron of *LEAFY* in *Neillia* and *Stephanandra* (Rosaceae) and implications for the origin of *Stephanandra*. *Mol. Phylogenet. Evol.* 29(2):203–215
138. **Ohyama K, Fukuzawa H, Kohchi T, Shirai H, Sano T, et al.** 1986. Chloroplast gene organization deduced from complete sequence of liverwort *Marchantia polymorpha* chloroplast DNA. *Nature.* 322(6079):572–574
139. **Okuyama Y, Fujii N, Wakabayashi M, Kawakita A, Ito M, et al.** 2005. Nonuniform concerted evolution and chloroplast capture: Heterogeneity of observed introgression patterns in three molecular data partition phylogenies of Asian *Mitella* (Saxifragaceae). *Mol. Biol. Evol.* 22(2):285–296
140. **Oliver D. . . assisted by other botanists.** 1868. *Flora of tropical Africa. Volume I.* London: Crown Agents for Overseas Governments and Administrations
141. *Overview of SOLiD™ Sequencing Chemistry.* <http://www.appliedbiosystems.com/absite/us/en/home/applications-technologies/solid-next-generation-sequencing/next-generation-systems/solid-sequencing-chemistry.html>
142. *PacBio Launches PacBio RS II Sequencer.* Next Gen Seek. <http://nextgenseek.com/2013/04/pacbio-launches-pacbio-rs-ii-sequencer/>
143. *PacBio Reveals Beta System Specs for RS; Says Commercial Release is on Track for First Half of 2011.* 2010. In Sequence. <http://www.genomeweb.com/sequencing/pacbio-reveals-beta-system-specs-rs-says-commercial-release-track-first-half-201>
144. *PacBio Ships First Two Commercial Systems; Order Backlog Grows to 44.* 2011. In Sequence. <http://www.genomeweb.com/print/968018>
145. *Pacific Biosciences: SMRT Sequencing Advantage.* 2014. SMRT Technology. <http://www.pacificbiosciences.com/products/smrt-technology/smrt-sequencing-advantage/>
146. **Palmer JD, Soltis DE, Chase MW.** 2004. The Plant Tree of Life: An Overview and some Points of View. *Am. J. Bot.* 91(10):1437–1445

147. **Patterson C, Williams DM, Humpries CJ.** 1993. Congruence between Molecular and Morphological Phylogenies. *Annu. Rev. Ecol. Syst.* 24:153–188
148. **Pelser PB, Kennedy AH, Tepe EJ, Shidler JB, Nordenstam B, et al.** 2010. Patterns and causes of incongruence between plastid and nuclear Senecioneae (Asteraceae) phylogenies. *Am. J. Bot.* 97(5):856–873
149. **Perkel J.** 2011. *Making contact with sequencing's fourth generation.* BioTechniques. <http://www.biotechniques.com/BiotechniquesJournal/2011/February/Making-Contact-with-Sequencings-Fourth-Generation/biotechniques-308942.html?service=print>
150. **Petrosino JF, Highlander S, Luna RA, Gibbs RA, Versalovic J.** 2009. Metagenomic pyrosequencing and microbial identification. *Clin. Chem.* 55(5):856–866
151. **Pettersson E, Lundeberg J, Ahmadian A.** 2009. Generations of sequencing technologies. *Genomics.* 93(2):105–111
152. **Philippe H, Blanchette M.** 2007. Introduction - Overview of the First Phylogenomics Conference. *BMC Evol. Biol.* 7(Suppl 1):S1–S16
153. **Pickrell WO, Rees MI, Chung S-K.** 2012. Next generation sequencing methodologies - an overview. In *Advances in protein chemistry and structural biology.* 89:1–26. Elsevier Inc. Volume 89 ed.
154. **Pirie MD, Humphreys AM, Barker NP, Linder HP.** 2009. Reticulation, data combination, and inferring evolutionary history: An example from Danthonioideae (Poaceae). *Syst. Biol.* 58(6):612–628
155. **Poggio L, Burghardt AD, Hunziker JH.** 1989. Nuclear DNA variation in diploid and polyploid taxa of *Larrea* (Zygophyllaceae). *Heredity (Edinb).* 63:321–328
156. **Poggio L, Hunziker JH, Wulff AF.** 1992. Cariotipo y Contenido de adn Nuclear de *Pintoa chilensis* y *Sisyndite sparteae* (Zygophyllaceae). *Darwiniana.* 31(1-4):11–15
157. **Poggio L, Wulff AF, Hunziker JH.** 1986. Chromosome Size, Nuclear Volume and DNA content in *Bulnesia* (Zygophyllaceae). *Darwiniana.* 27(1-4):25–38
158. **Popp M, Oxelman B.** 2001. Inferring the history of the polyploid *Silene aegaea* (Caryophyllaceae) using plastid and homoeologous nuclear DNA sequences. *Mol. Phylogenet. Evol.* 20(3):474–481
159. **Qiu Y, Lee J, Bernasconi-Quadroni F, Soltis DE, Soltis PS, et al.** 1999. The earliest angiosperms: evidence from mitochondrial , plastid and nuclear genomes. *Nature.* 402(November):404–407
160. **Quail M a, Smith M, Coupland P, Otto TD, Harris SR, et al.** 2012. A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics.* 13(1):341
161. **Roberts RJ, Carneiro MO, Schatz MC.** 2013. The advantages of SMRT sequencing. *Genome Biol.* 14(6):405

162. *Roche Shutting Down 454 Sequencing Business*. 2013. GenomeWeb.
<http://www.genomeweb.com/sequencing/roche-shutting-down-454-sequencing-business>
163. **Ronaghi M**. 1998. DNA SEQUENCING: A Sequencing Method Based on Real-Time Pyrophosphate. *Science* (80-.). 281(5375):363–65
164. **Ronaghi M, Karamohamed S, Pettersson B, Uhle M**. 1996. Real-Time DNA Sequencing Using Detection of Pyrophosphate Release. *Anal. Biochem.* 242:84–89
165. **Ronquist F, Huelsenbeck JP**. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics*. 19(12):1572–1574
166. **Rothberg JM, Hinz W, Rearick TM, Schultz J, Mileski W, et al**. 2011. An integrated semiconductor device enabling non-optical genome sequencing. *Nature*. 475(7356):348–352
167. **Rouse GW, Wilson NG, Worsaae K, Vrijenhoek RC**. 2015. Report A Dwarf Male Reversal in Bone-Eating Worms. *Curr. Biol.* 25(2):236–41
168. **Ruhfel BR, Gitzendanner MA, Soltis PS, Soltis DE, Burleigh JG**. 2014. From algae to angiosperms-inferring the phylogeny of green plants (Viridiplantae) from 360 plastid genomes. *BMC Evol. Biol.* 14(1):23
169. **Rusk N, Kiermer V**. 2008. Primer: Sequencing-the next generation. *Nat. Methods*. 5(1):15
170. **Sage RF, Christin P-A, Edwards EJ**. 2011. The C₄ plant lineages of planet Earth. *J. Exp. Bot.* 62(9):3155–3169
171. **Sanger F, Nicklen S**. 1977. DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. U. S. A.* 74(12):5463–5467
172. **Savolainen V, Chase MW, Hoot SB, Morton CM, Soltis DE, et al**. 2000. Phylogenetics of flowering plants based on combined analysis of plastid *atpB* and *rbcL* gene sequences. *Syst. Biol.* 49(2):306–362
173. **Savolainen V, Fay MF, Albachi DC, Backlund A, Van der Bank M, et al**. 2000. Phylogeny of the eudicots: a nearly complete familial analysis based on *rbcL* gene sequences. *Kew Bull.* 55:257–309
174. **Schoch CL, Seifert KA, Huhndorf S, Robert V, Spouge JL, et al**. 2012. Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi. *Proc. Natl. Acad. Sci. U. S. A.* 109(16):6241–6246
175. **Schreiber A**. 1963. Die Gattung *Zygophyllum* L. in Südwestafrika. *Sonderdruck aus den Mitteilungen der Bot. Staatssammlung München*. 5:49–114
176. **Scotland RW, Olmstead RG, Bennett JR**. 2003. Phylogeny Reconstruction: The Role of Morphology. *Syst. Biol.* 52(4):539–548

177. **Shaw J, Lickey EB, Beck JT, Farmer SB, Liu W, et al.** 2005. The Tortoise and the Hare II: Relative Utility of 21 Non-coding Chloroplast DNA Sequences for Phylogenetic Analysis. *Am. J. Bot.* 92(1):142–166
178. **Shaw J, Lickey EB, Schilling EE, Small RL.** 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(3):275–288
179. **Sheahan MC, Chase MW.** 1996. A phylogenetic analysis of Zygophyllaceae R. Br. based on morphological, anatomical and *rbcL* DNA sequence data. *Bot. J. Linn. Soc.* 122:279–300
180. **Sheahan MC, Chase MW.** 2000. Phylogenetic Relationships within Zygophyllaceae Based on DNA Sequences of Three Plastid Regions , with Special Emphasis on Zygophylloideae Phylogenetic Relationships within Zygophyllaceae Based on DNA Sequences of three Plastid Regions , with Special Emp. *Syst. Bot.* 25(2):371–384
181. **Shinozaki K, Ohme M, Tanaka M, Wakasugi T, Hayashida N, et al.** 1986. The complete nucleotide sequence of the tobacco chloroplast genome: its gene organization and expression. *EMBO J.* 5(9):2043–2049
182. **Silberfeld T, Leigh JW, Verbruggen H, Cruaud C, de Reviers B, Rousseau F.** 2010. A multi-locus time-calibrated phylogeny of the brown algae (Heterokonta, Ochrophyta, Phaeophyceae): Investigating the evolutionary nature of the “brown algal crown radiation.” *Mol. Phylogenet. Evol.* 56(2):659–674
183. Single Molecule Real Time (SMRT™) DNA Sequencing. 2009
184. *Six Years After Acquisition, Roche Quietly Shuttters 454.* 2013. Bio-IT World. <http://www.bio-itworld.com/2013/10/16/six-years-after-acquisition-roche-quietly-shuttters-454.html>
185. **Small RL, Cronn RC, Wendel JF.** 2004. L.A.S. JOHNSON REVIEW No. 2 - Use of nuclear genes for phylogeny reconstruction in plants. *Aust. Syst. Bot.* 17:145–170
186. **Small RL, Ryburn JA, Cronn RC, Seelanan T, Wendel JF.** 1998. The tortoise and the hare: choosing between noncoding plastome and nuclear Adh sequences for phylogeny reconstruction in a recently diverged plant group. *Am. J. Bot.* 85(9):1301–1315
187. *SMRT Cells: Consumables.* 2013. <http://www.pacificbiosciences.com/products/consumables/SMRT-cells/>
188. **Soltis DE, Smith SA, Cellinese N, Wurdack KJ, Tank DC, et al.** 2011. Angiosperm phylogeny: 17 genes, 640 taxa. *Am. J. Bot.* 98(4):704–730
189. **Soltis DE, Soltis PS, Chase MW, Mort ME, Albach DC, et al.** 2000. Angiosperm phylogeny inferred from 18S rDNA, *rbcL*, and *atpB* sequences. *Bot. J. Linn. Soc.* 133(4):381–461
190. **Soltis DE, Soltis PS, Clegg MT, Durbin M.** 1990. *RbcL* sequence divergence and phylogenetic relationships in Saxifragaceae sensu lato. *Proc. Natl. Acad. Sci. United States.* 87(June):4640–4644

191. **Soltis PS, Soltis DE, Chase MW.** 1999. Angiosperm phylogeny inferred from multiple genes as a tool for comparative biology. *Nature*. 402(6760):402–404
192. **Sonder OW.** 1860. Zygophyllaceae. In *Flora Capensis*, ed W. Harvey, OW Sonder. A.S. Robertson. Volume 1 ed.
193. **Stamatakis A.** 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*. 30(9):1312–1313
194. **Stanley EL, Bauer AM, Jackman TR, Branch WR, Mouton PLFN.** 2011. Between a rock and a hard polytomy: Rapid radiation in the rupicolous girdled lizards (Squamata: Cordylidae). *Mol. Phylogenet. Evol.* 58(1):53–70
195. **Steele K, Vilgalys R.** 1994. Phylogenetic analyses of Polemoniaceae using nucleotide sequences of the plastid gene *matK*. *Syst. Bot.* 19(1):126–142
196. **Stegemann S, Keuthe M, Greiner S, Bock R.** 2012. Horizontal transfer of chloroplast genomes between plant species. *Proc. Natl. Acad. Sci.* 109(7):2434–2438
197. **Straub SCK, Moore MJ, Soltis PS, Soltis DE, Liston A, Livshultz T.** 2014. Phylogenetic signal detection from an ancient rapid radiation: Effects of noise reduction, long-branch attraction, and model selection in crown clade Apocynaceae. *Mol. Phylogenet. Evol.*
198. **Sun M, Soltis DE, Soltis PS, Zhu X, Burleigh JG, Chen Z.** 2014. Deep phylogenetic incongruence in the angiosperm clade Rosidae. *Mol. Phylogenet. Evol.*
199. **Swofford DL.** 2002. PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods). *Sinauer Assoc. Sunderland, Massachusetts*
200. **Taberlet P, Gielly L, Pautou G, Bouvet J.** 1991. Universal primers for amplification of three non-coding regions of chloroplast DNA. *Plant Mol. Biol.* 17(5):1105–9
201. **Takhtajan A.** 1997. *Diversity and Classification of Flowering Plants*. Columbia University Press
202. **Takhtajan A.** 2009. *Flowering Plants*. Springer. Second Ed.
203. **Tsitrona A, Kirkpatrick M, Levin DA.** 2003. A model for chloroplast capture. *Evolution*. 57(8):1776–1782
204. **Vaidya G, Lohman DJ, Meier R.** 2011. SequenceMatrix: concatenation software for the fast assembly of multi-gene datasets with character set and codon information. *Cladistics*. 27(2):171–180
205. **Valcárcel V, Fiz-Palacios O, Wen J.** 2014. The origin of the early differentiation of Ivies (*Hedera* L.) and the radiation of the Asian Palmate group (*Araliaceae*). *Mol. Phylogenet. Evol.* 70:492–503
206. **Van Huyssteen DC.** 1937. *Morphologisch-systematische Studien über die Gattung Zygophyllum mit besonderer Berücksichtigung der afrikanischen Arten*. Fakultät der Friedrich Wilhelms-Universität zu Berlin.

207. **Van Zyl L.** 2000. *A Systematic Revision of Zygophyllum (Zygophyllaceae) in the Southern African Region.* University of Stellenbosch
208. **Voelkerding K V, Dames SA, Durtschi JD.** 2009. Next-generation sequencing: from basic research to diagnostics. *Clin. Chem.* 55(4):641–658
209. **Wagner GP, Booth G, Bagheri-chaichian H.** 1994. Phylogenies without Fossils. *Evolution.* 48(3):329–347
210. **Wang H, Moore MJ, Soltis PS, Bell CD, Brockington SF, et al.** 2009. Rosid radiation and the rapid rise of angiosperm-dominated forests. *Proc. Natl. Acad. Sci. U. S. A.* 106(10):3853–3858
211. **Whelan S, Liò P, Goldman N.** 2001. Molecular Phylogenetics: State-of-the- art Methods for looking into the Past. *Trends Genet.* 17(5):262–272
212. **Whiting MF, Bradler S, Maxwell T.** 2003. Ups and downs of evolution Insects that lost – then re-evolved – the ability to fly. *Nature.* 421(January):264–267
213. **Wilson MA, Gaut B, Clegg MT.** 1990. Chloroplast DNA Evolves Slowly in the Palm Family (Arecaceae). *Mol. Biol. Evol.* 7(4):303–314
214. **Y. Sun, D. Z. Skinner, G. H. Liang SHH.** 1994. Phylogenetic analysis of *Sorghum* and related taxa using internal transcribed spacers of nuclear ribosomal DNA. *Theor. Appl. Genet.* 89:26–32
215. **Yang TW, Yang YA, Xiong Z.** 2000. Paternal Inheritance of chloroplast DNA in interspecific hybrids in the genus *Larrea* (Zygophyllaceae). *Am. J. Bot.* 87(10):1452–1458
216. **Yuan Y-W, Liu C, Marx HE, Olmstead RG.** 2009. The pentatricopeptide repeat (PPR) gene family, a tremendous resource for plant phylogenetic studies. *New Phytol.* 182(1):272–283
217. **Zachos J, Pagani M, Sloan L, Thomas E, Billups K.** 2001. Trends, rhythms, and aberrations in global climate 65 Ma to present. *Science (80-).* 292(5517):686–693
218. **Zhenxiang X, Liu L, Rest JS, Davis CC.** 2014. Coalescent versus Concatenation Methods and the Placement of *Amborella* as Sister to Water Lilies. *Syst. Biol.* 63(6):919–932
219. **Zhong B, Liu L, Yan Z, Penny D.** 2013. Origin of land plants using the multispecies coalescent model. *Trends Plant Sci.* 18(9):492–495
220. **Ziegler H, Batanouny KH, Sankhla N, Vyas OP, Stichler W.** 1981. The Photosynthetic Pathway Types of some Desert Plants from India, Saudi Arabia, Egypt, and Iraq. *Oecologia.* 48:93–99

6 Appendices

6-1: Below is a list of the 128 genes found in the chloroplast genome of *Corynocarpus laevigata*.

<i>Corynocarpus laevigata</i> Complete Chloroplast Genome Gene List					
Gene Number	Symbol	Description	Start Position on the Genomic Accession	End Position on the Genomic Accession	Orientation
Large Single Copy Region (LSC)					
1	<i>trnH</i>	tRNA	14	88	minus
2	<i>psbA</i>	PSII 32 kDa protein	478	1539	minus
3	<i>trnK</i>	tRNA	1848	4450	minus
4	<i>matK</i>	maturase K	2170	3684	minus
5	<i>rps16</i>	ribosomal protein S16	5746	6869	minus
6	<i>trnQ</i>	tRNA	8024	8095	minus
7	<i>psbK</i>	PSII K protein	8471	8656	plus
8	<i>psbI</i>	PSII I protein	9065	9175	plus
9	<i>trnS</i>	tRNA	9297	9384	minus
10	<i>trnG</i>	tRNA	10162	10914	plus
11	<i>trnR</i>	tRNA	11225	11296	plus
12	<i>atpA</i>	ATPase alpha subunit	11941	13464	minus
13	<i>atpF</i>	ATPase subunit I	13545	14882	minus
14	<i>atpH</i>	ATPase subunit III	15346	15591	minus
15	<i>atpI</i>	ATP synthase CF ₀ A subunit	16680	17423	minus
16	<i>rps2</i>	ribosomal protein S2	17638	18348	minus
17	<i>rpoC2</i>	RNA polymerase beta II subunit	18609	22751	minus
18	<i>rpoC1</i>	RNA polymerase beta I subunit	22917	25726	minus
19	<i>rpoB</i>	RNA polymerase beta subunit	25732	28944	minus
20	<i>trnC</i>	tRNA	30240	30310	plus
21	<i>petN</i>	cytochrome b6/f complex subunit VIII	31217	31306	plus
22	<i>psbM</i>	PSII M protein	31940	32044	minus
23	<i>trnD</i>	tRNA	32616	32689	minus
24	<i>trnY</i>	tRNA	33147	33230	minus
25	<i>trnE</i>	tRNA	33290	33362	minus
26	<i>trnT</i>	tRNA	34379	34450	plus
27	<i>psbD</i>	PSII D2 protein	35966	37027	plus
28	<i>psbC</i>	PSII 44 kDa protein	36975	38396	plus
29	<i>trnS</i>	tRNA	38645	38737	minus
30	<i>psbZ</i>	photosystem II reaction center Z protein	39054	39242	plus
31	<i>trnG</i>	tRNA	39767	39837	plus
32	<i>trnM</i>	tRNA	40021	40094	minus
33	<i>rps14</i>	ribosomal protein S14	40259	40561	minus
34	<i>psaB</i>	PSI P700 apoprotein A2	40688	42892	minus
35	<i>psaA</i>	PSI P700 apoprotein A1	42918	45170	minus
36	<i>ycf3</i>	Ycf3 protein	45950	47914	minus

37	<i>trnS</i>	tRNA	48662	48740	plus
38	<i>rps4</i>	ribosomal protein S4	49050	49655	minus
39	<i>trnT</i>	tRNA	50032	50104	minus
40	<i>trnL</i>	tRNA	51380	52008	plus
41	<i>trnF</i>	tRNA	52389	52461	plus
42	<i>ndhJ</i>	NADH dehydrogenase 19 kDa subunit	53255	53731	minus
43	<i>ndhK</i>	NADH dehydrogenase 32 kDa subunit	53836	54516	minus
44	<i>ndhC</i>	NADH dehydrogenase subunit 3	54566	54928	minus
45	<i>trnV</i>	tRNA	55693	56365	minus
46	<i>trnM</i>	tRNA	56545	56617	plus
47	<i>atpE</i>	ATPase epsilon subunit	56796	57197	minus
48	<i>atpB</i>	ATPase beta subunit	57194	58690	minus
49	<i>rbcL</i>	ribulose-1,5-bisphosphate carboxylase/oxygenase large subunit	59493	60920	plus
50	<i>accD</i>	acetyl-CoA carboxylase beta subunit	61492	62982	plus
51	<i>psaI</i>	PSI reaction center subunit VIII	63764	63877	plus
52	<i>ycf4</i>	Ycf4 protein	64317	64871	plus
53	<i>cemA</i>	potential heme-binding protein	65348	66037	plus
54	<i>petA</i>	cytochrome f	66256	67218	plus
55	<i>psbJ</i>	PSII reaction center subunit X	68060	68182	minus
56	<i>psbL</i>	PSII reaction center subunit XII	68325	68441	minus
57	<i>psbF</i>	PSII reaction center subunit VI	68464	68583	minus
58	<i>psbE</i>	PSII reaction center subunit V	68593	68844	minus
59	<i>petL</i>	cytochrome b6/f complex subunit VI	70159	70254	plus
60	<i>petG</i>	cytochrome b6/f complex subunit V	70453	70566	plus
61	<i>trnW</i>	tRNA	70684	70757	minus
62	<i>trnP</i>	tRNA	70916	70989	minus
63	<i>psaJ</i>	PSI reaction center subunit IX	71398	71532	plus
64	<i>rpl33</i>	ribosomal protein L33	72069	72269	plus
65	<i>rps18</i>	ribosomal protein S18	72439	72744	plus
66	<i>rpl20</i>	ribosomal protein L20	73060	73413	minus
67	<i>rps12</i>	ribosomal protein S12	74305	74192	plus
68	<i>rps12</i>	ribosomal protein S12	74305	74192	plus
69	<i>clpP</i>	ATP-dependent protease	74488	76535	minus
70	<i>psbB</i>	photosystem II 47 kDa protein	77004	78530	plus
71	<i>psbT</i>	PSII T protein	78709	78810	plus
72	<i>psbN</i>	PSII N protein	78876	79007	minus
73	<i>psbH</i>	PSII 10 kDa phosphoprotein	79112	79333	plus
74	<i>petB</i>	cytochrome b6	79448	80888	plus
75	<i>petD</i>	cytochrome b6/f complex subunit IV	81081	82292	plus
76	<i>rpoA</i>	RNA polymerase alpha subunit	82469	83464	minus
77	<i>rps11</i>	ribosomal protein S11	83530	83952	minus
78	<i>rpl36</i>	ribosomal protein L36	84060	84173	minus
79	<i>rps8</i>	ribosomal protein S8	84621	85025	minus
80	<i>rpl14</i>	ribosomal protein L14	85176	85544	minus
81	<i>rpl16</i>	ribosomal protein L16	85704	87258	minus

82	<i>rps3</i>	ribosomal protein S3	87404	88060	minus
83	<i>rpl22</i>	ribosomal protein L22	88045	88545	minus
84	<i>rps19</i>	ribosomal protein S19	88630	88908	minus
Inverted Repeat Region A (IR _A)					
85	<i>rpl2</i>	ribosomal protein L2	88978	90488	minus
86	<i>rpl23</i>	ribosomal protein L23	90507	90788	minus
87	<i>trnI</i>	tRNA	90950	91023	minus
88	<i>ycf2</i>	Ycf2 protein	91112	98062	plus
89	<i>trnL</i>	tRNA	98650	98730	minus
90	<i>ndhB</i>	NADH dehydrogenase subunit 2	99233	101451	minus
91	<i>rps7</i>	ribosomal protein S7	101787	102254	minus
92	<i>trnV</i>	tRNA	104676	104747	plus
93	<i>rrn16</i>	16S ribosomal RNA	104974	106464	plus
94	<i>trnI</i>	tRNA	106761	107809	plus
95	<i>trnA</i>	tRNA	107874	108631	plus
96	<i>rrn23</i>	23S ribosomal RNA	108787	111595	plus
97	<i>rrn4.5</i>	4.5S ribosomal RNA	111694	111796	plus
98	<i>rrn5</i>	5S ribosomal RNA	112056	112176	plus
99	<i>trnR</i>	tRNA	112447	112520	plus
100	<i>trnN</i>	tRNA	113128	113199	minus
Small Single Copy Region (SSC)					
101	<i>ndhF</i>	NADH dehydrogenase subunit 5	114572	116809	minus
102	<i>rpl32</i>	ribosomal protein L32	117574	117732	plus
103	<i>trnL</i>	tRNA	119205	119284	plus
104	<i>ccsA</i>	cytochrome c biogenesis protein	119413	120396	plus
105	<i>ndhD</i>	NADH dehydrogenase subunit 4	120561	122063	minus
106	<i>psaC</i>	PSI 9 kDa protein	122190	122435	minus
107	<i>ndhE</i>	NADH dehydrogenase subunit 4L	122657	122959	minus
108	<i>ndhG</i>	NADH dehydrogenase subunit 6	123188	123718	minus
109	<i>ndhI</i>	NADH dehydrogenase 18 kDa subunit	124086	124595	minus
110	<i>ndhA</i>	NADH dehydrogenase subunit 1	124677	126963	minus
111	<i>ndhH</i>	NADH dehydrogenase 49 kDa subunit	126965	128146	minus
112	<i>rps15</i>	ribosomal protein S15	128260	128535	minus
113	<i>ycf1</i>	hypothetical protein	128983	134541	minus
Inverted Repeat Region B (IR _B)					
114	<i>trnN</i>	tRNA	134866	134937	plus
115	<i>trnR</i>	tRNA	135545	135618	minus
116	<i>rrn5</i>	5S ribosomal RNA	135889	136009	minus
117	<i>rrn4.5</i>	4.5S ribosomal RNA	136269	136371	minus
118	<i>rrn23</i>	23S ribosomal RNA	136470	139278	minus
119	<i>trnA</i>	tRNA	139434	140191	minus
120	<i>trnI</i>	tRNA	140256	141304	minus
121	<i>rrn16</i>	16S ribosomal RNA	141601	143091	minus
122	<i>trnV</i>	tRNA	143318	143389	minus
123	<i>rps7</i>	ribosomal protein S7	145811	146278	plus
124	<i>ndhB</i>	NADH dehydrogenase subunit 2	146614	148832	plus

125	<i>trnL</i>	tRNA	149335	149415	plus
126	<i>ycf2</i>	Ycf2 protein	150003	156953	minus
127	<i>rpl23</i>	ribosomal protein L23	157277	157558	plus
128	<i>rpl2</i>	ribosomal protein L2	157577	159087	plus

Transfer RNA genes
Gene products involved in the Photosynthetic Process
Maturase genes
RNA polymerase subunit genes
Hypothetical genes or genes with unknown functions
Carboxylase subunit genes
Protease genes
Ribosomal RNA genes

6-2: The Ion Torrent run report and FastQC report of *Zygophyllum fabago*

CAF - Stellenbosch - Torrent Browser

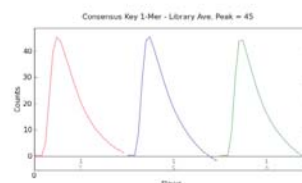
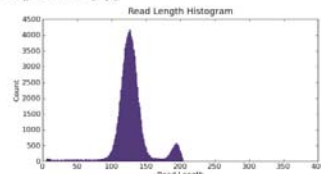
http://146.232.41.138/output/Home/SMA_12-v2.0_097/Defau...

Report for SMA_12-v2.0

Library Summary

Based on Predicted Per-Base Quality Scores - Independent of Alignment

Total Number of Bases [Mbp]	16.55
• Number of Q17 Bases [Mbp]	12.76
• Number of Q20 Bases [Mbp]	11.41
Total Number of Reads	129,360
Mean Length [bp]	128
Longest Read [bp]	204



Reference Genome Information

Genome Name	Corynocarpus laevigata CP
Genome Size	159,202 bases
Genome Version	31_01_2011-G8
Index Version	tmap-f2

Based on Full Library Alignment to Provided Reference

	AQ17	AQ20	Perfect
Total Number of Bases [Mbp]	0.10	0.07	0.06
Mean Length [bp]	54	43	39
Longest Alignment [bp]	143	143	138
Mean Coverage Depth	0.70x	0.40x	0.40x
Percentage of Library Covered	79%	78%	78%

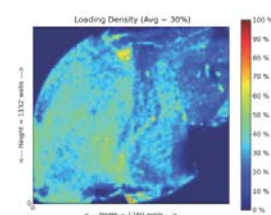
Read Alignment Distribution

Read Length [bp]	Reads	Unmapped	Excluded	Clipped	Perfect	1 mismatch	≥2 mismatches
50	87,365	78,059	2	0	304	620	8,380
100	84,689	75,603	2	2,233	59	109	6,683
150	7,766	7,278	0	467	0	0	21

Test Fragment Report

Ion Sphere™ Particle (ISP) Identification Summary

	Count	Percentage
Total Addressable Wells	1,262,520	
• Wells with ISPs	372,121	29%
• Live ISPs	309,216	83%
• Test Fragment ISPs	18,282	6%
• Library ISPs	290,934	94%
Library ISPs / Percent Enrichment		
• Filtered: Polyclonal	290,934	82%
• Filtered: Primer dimer	83,882	29%
• Filtered: Low quality	45	<1%
• Final Library Reads	77,647	27%
	129,360	44%



Report Information

Analysis Info

Run Name	R_2011_07_16_16_46_21_user_SMA-12
Run Date	2011-07-16 16:46:21
Analysis Name	SMA_12-v2.0
Analysis Date	2012-01-26
Analysis Cycles	65
Analysis Flows	260












R_2011_07_16_16_46_21_user_SMA-12_Auto_SMA-12_14.fastq Fa... file:///C:/Documents and Settings/14411911/My Documents/Ion Torrent ...

FastQC Report

Summary

Fri 20 Feb 2015

R_2011_07_16_16_46_21_user_SMA-12_Auto_SMA-12_14.fastq

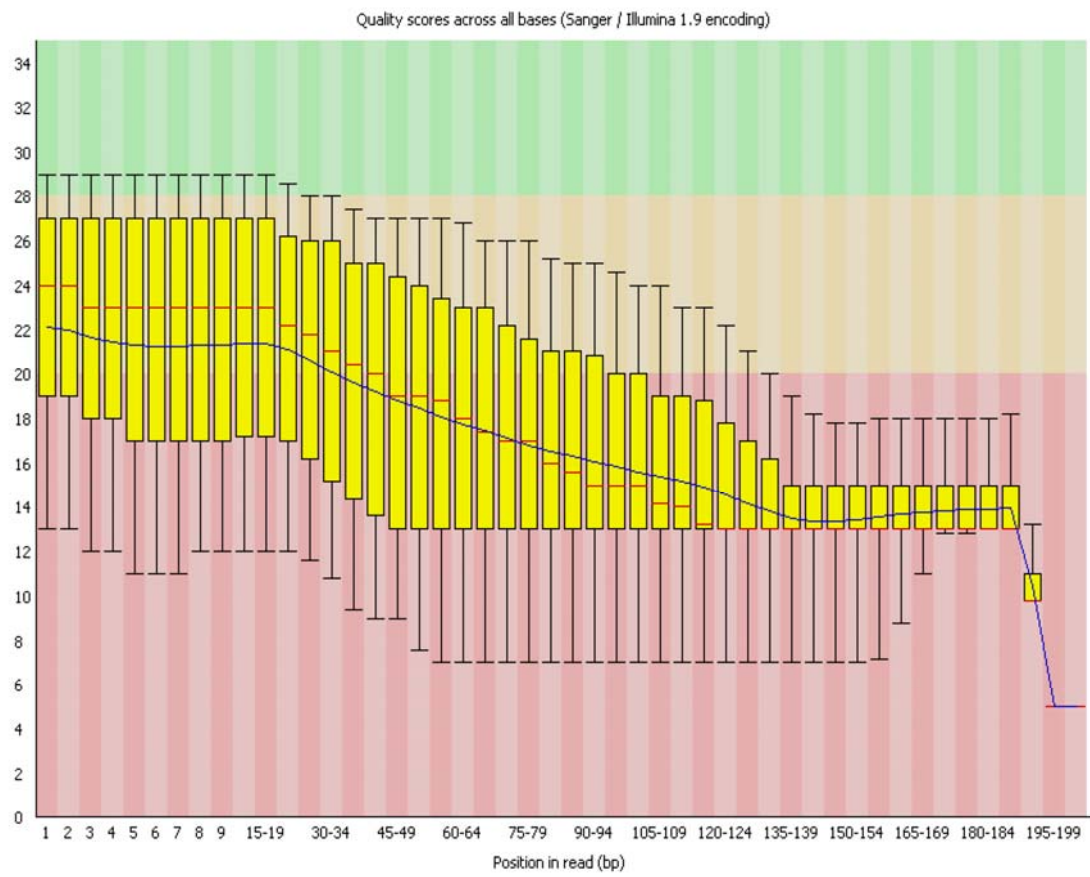
-  [Basic Statistics](#)
-  [Per base sequence quality](#)
-  [Per sequence quality scores](#)
-  [Per base sequence content](#)
-  [Per sequence GC content](#)
-  [Per base N content](#)
-  [Sequence Length Distribution](#)
-  [Sequence Duplication Levels](#)
-  [Overrepresented sequences](#)
-  [Adapter Content](#)
-  [Kmer Content](#)

Basic Statistics

Measure	Value
Filename	R_2011_07_16_16_46_21_user_SMA-12_Auto_SMA-12_14.fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	123388
Sequences flagged as poor quality	0
Sequence length	5-203
%GC	38

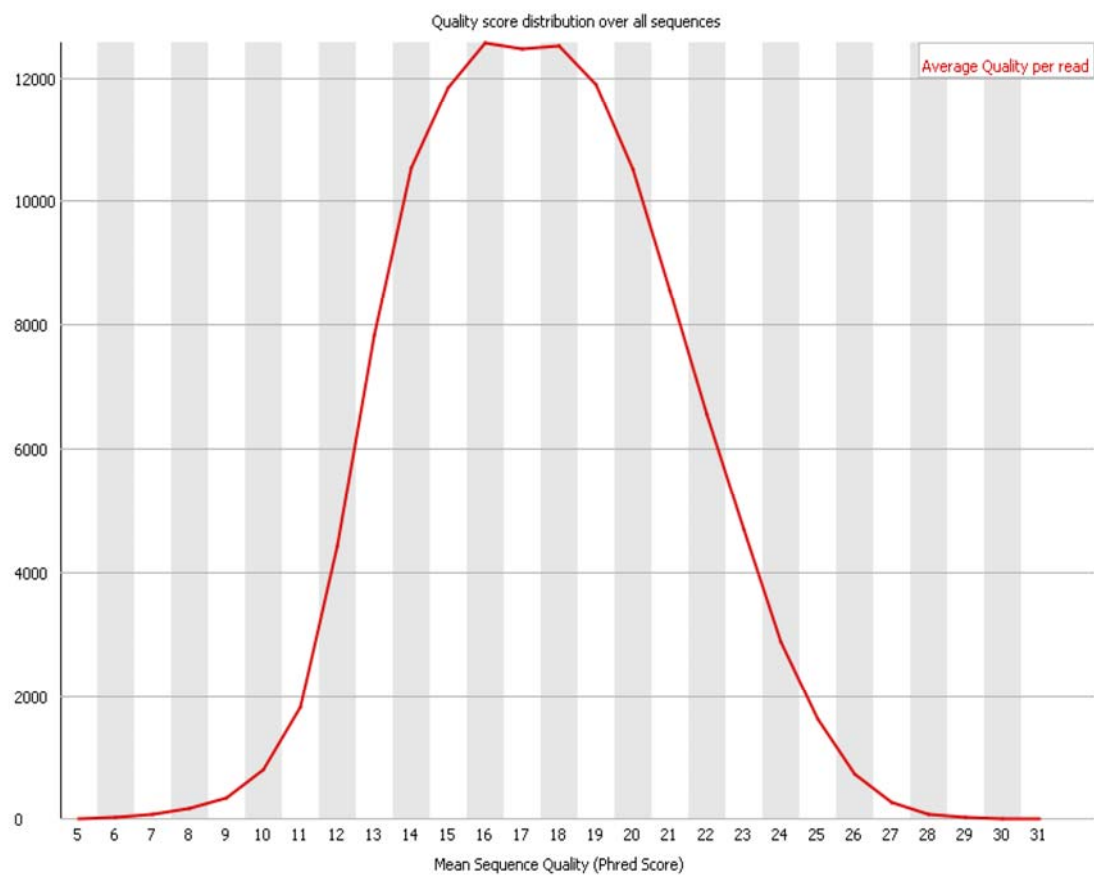
Per base sequence quality

R_2011_07_16_16_46_21_user_SMA-12_Auto_SMA-12_14.fastq Fa... file:///C:/Documents and Settings/14411911/My Documents/Ion Torrent ...



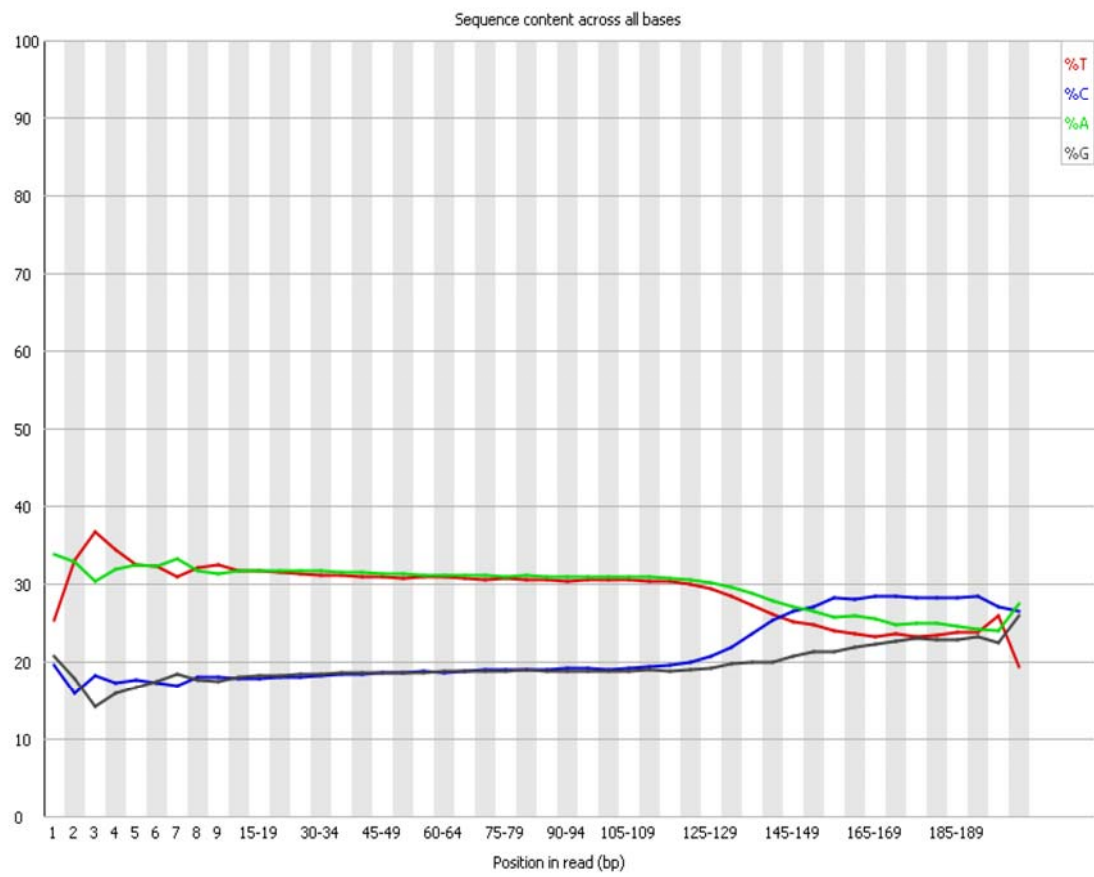
 **Per sequence quality scores**

R_2011_07_16_16_46_21_user_SMA-12_Auto_SMA-12_14.fastq Fa... file:///C:/Documents and Settings/14411911/My Documents/Ion Torrent ...



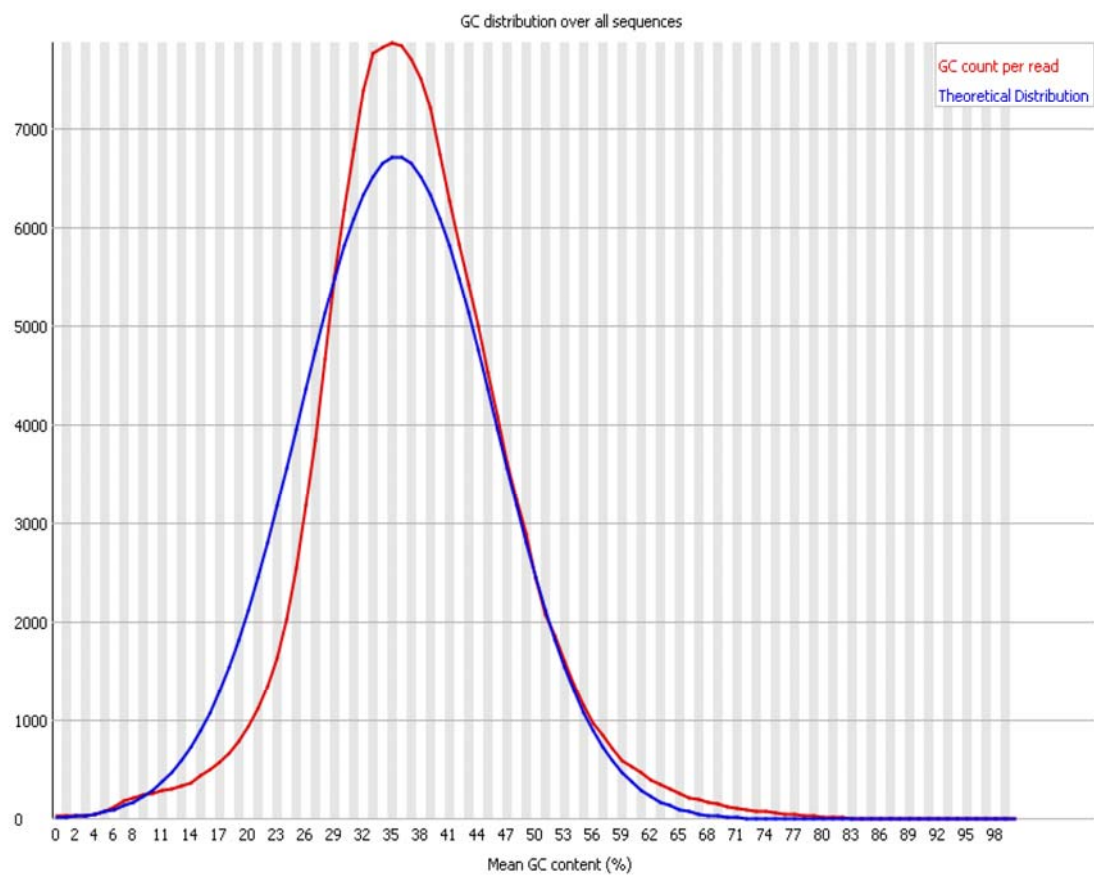
✖ Per base sequence content

R_2011_07_16_16_46_21_user_SMA-12_Auto_SMA-12_14.fastq Fa... file:///C:/Documents and Settings/14411911/My Documents/Ion Torrent ...



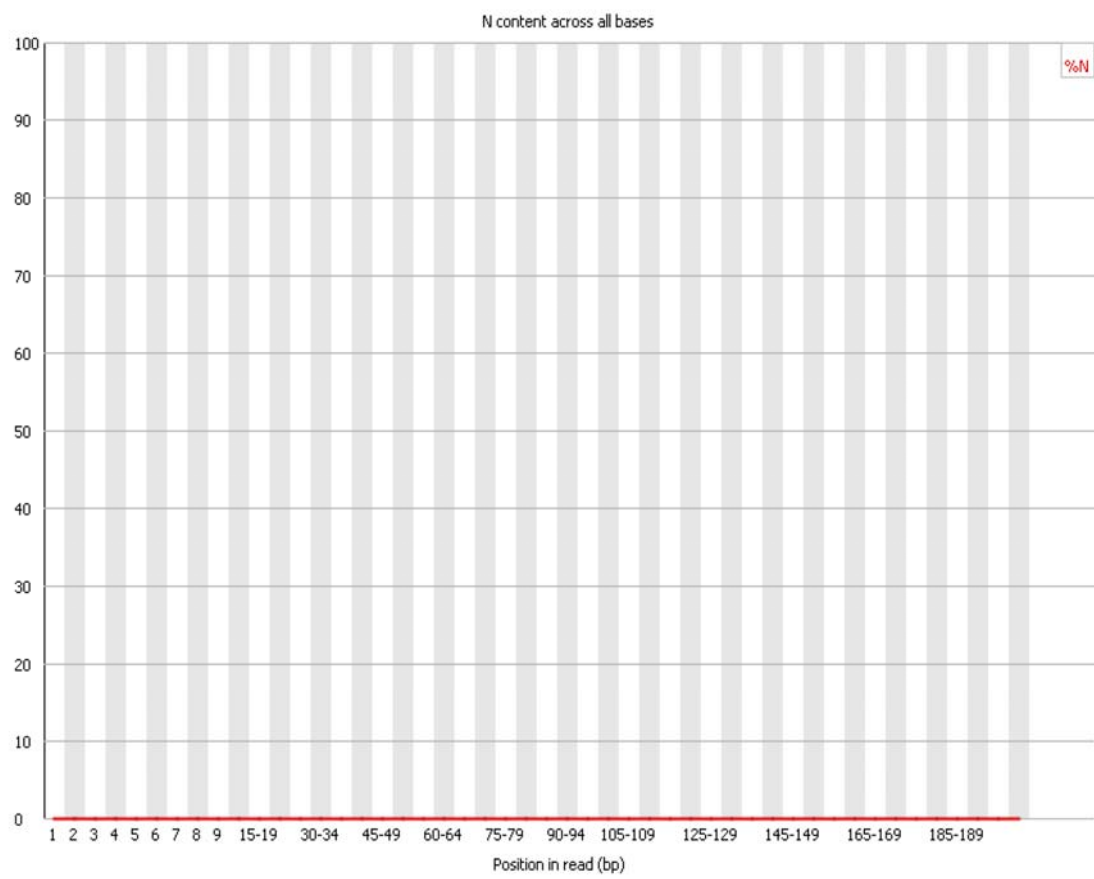
! Per sequence GC content

R_2011_07_16_16_46_21_user_SMA-12_Auto_SMA-12_14.fastq Fa... file:///C:/Documents and Settings/14411911/My Documents/Ion Torrent ...



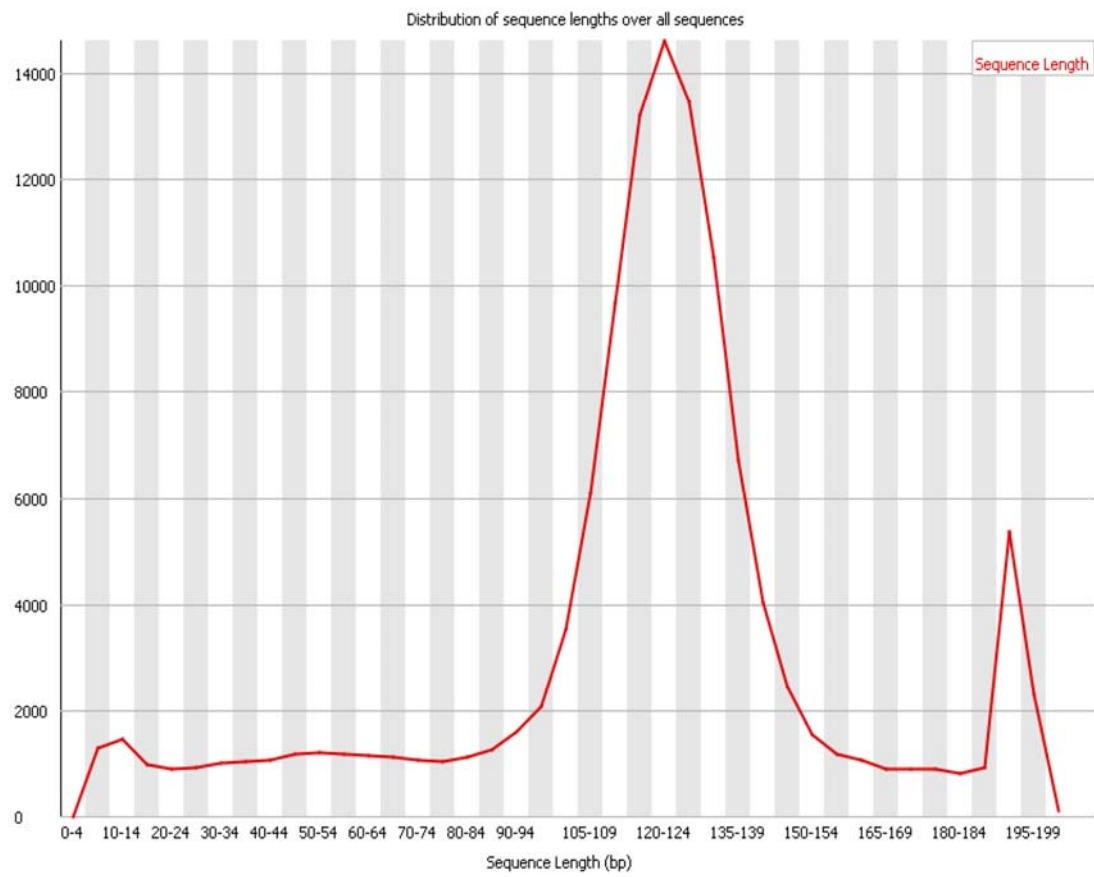
Per base N content

R_2011_07_16_16_46_21_user_SMA-12_Auto_SMA-12_14.fastq Fa... file:///C:/Documents and Settings/14411911/My Documents/Ion Torrent ...



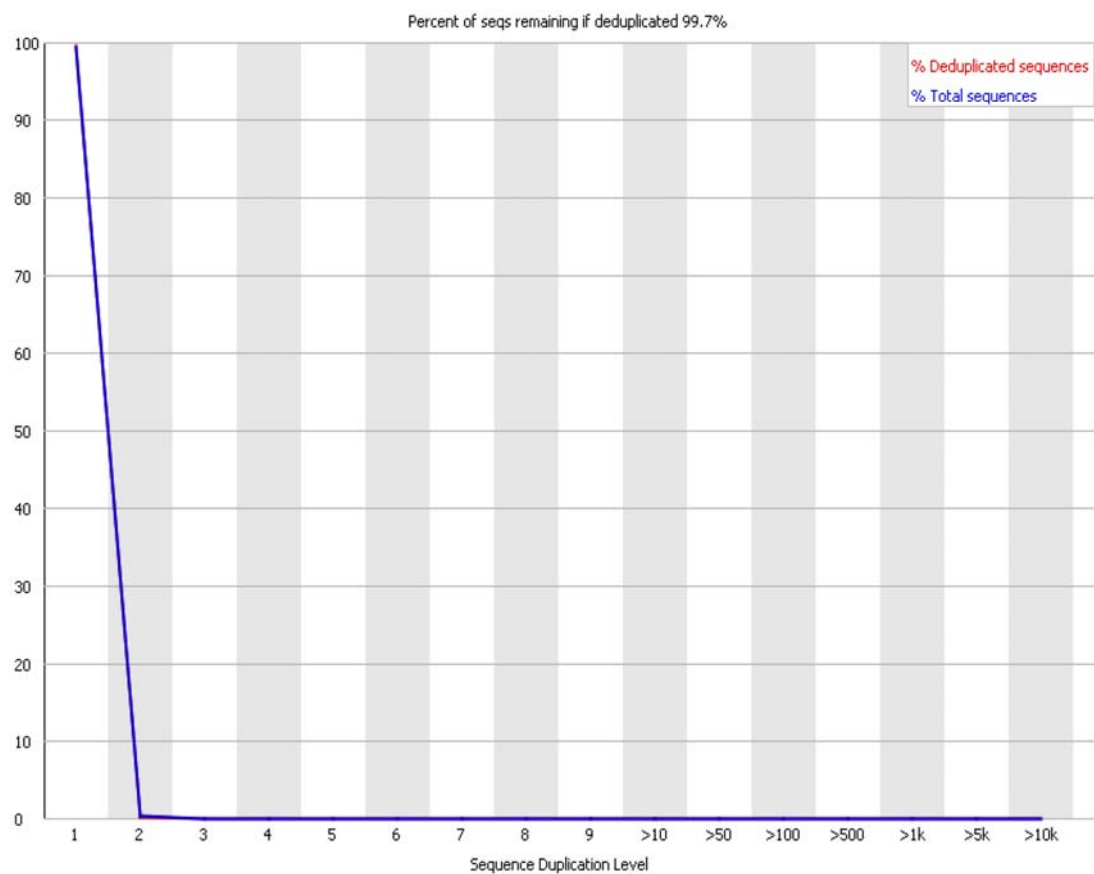
🚫 Sequence Length Distribution

R_2011_07_16_16_46_21_user_SMA-12_Auto_SMA-12_14.fastq Fa... file:///C:/Documents and Settings/14411911/My Documents/Ion Torrent ...



✔ Sequence Duplication Levels

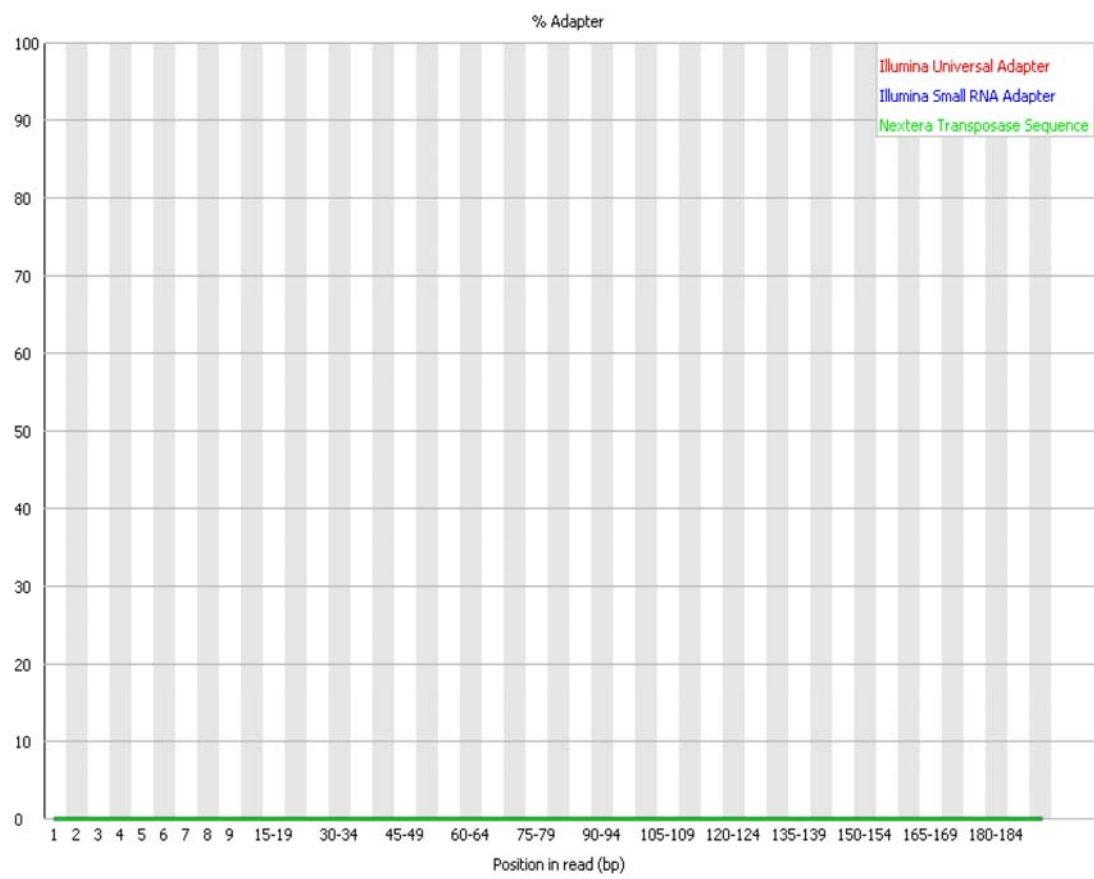
R_2011_07_16_16_46_21_user_SMA-12_Auto_SMA-12_14.fastq Fa... file:///C:/Documents and Settings/14411911/My Documents/Ion Torrent ...



 **Overrepresented sequences**
No overrepresented sequences

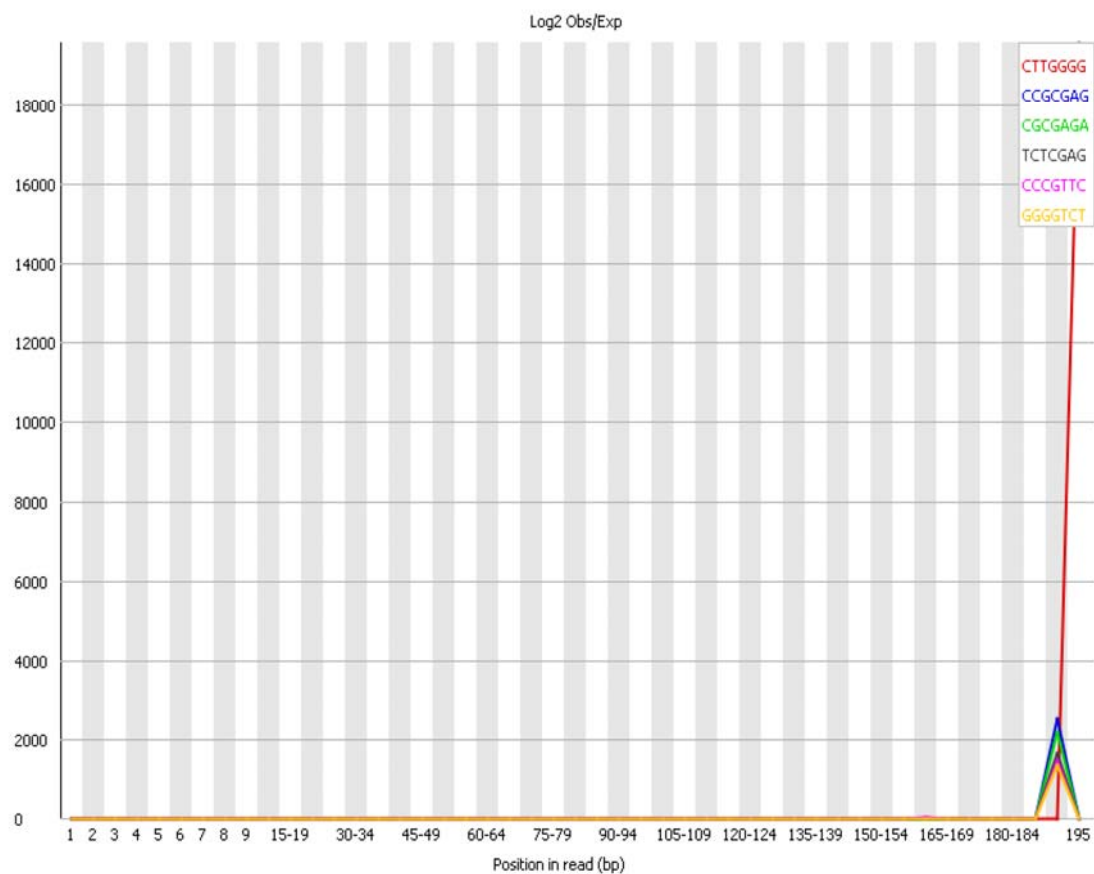
 **Adapter Content**

R_2011_07_16_16_46_21_user_SMA-12_Auto_SMA-12_14.fastq Fa... file:///C:/Documents and Settings/14411911/My Documents/Ion Torrent ...



 **Kmer Content**

R_2011_07_16_16_46_21_user_SMA-12_Auto_SMA-12_14.fastq Fa... file:///C:/Documents and Settings/14411911/My Documents/Ion Torrent ...



Sequence	Count	PValue	Obs/Exp Max	Max Obs/Exp Position
CTTGGGG	70	0.0	19546.143	195
CCGCGAG	30	0.0012048531	2533.7593	190-194
CGCGAGA	20	0.0016273911	2206.8225	190-194
TCTCGAG	45	0.0027106022	1689.173	190-194
CCCGTTC	30	0.0036611126	1471.215	190-194
GGGGTCT	55	0.004048856	1382.0505	190-194
TCCCGTT	60	0.0048182853	1266.8796	190-194
GGGTCTG	35	0.004982829	1261.0415	190-194
CTCGAGC	35	0.004982829	1261.0415	190-194
CATCCGG	65	0.005654573	1169.4272	190-194
GCGAGAT	35	0.0075322813	1028.7444	190-194
CCTTAAG	80	0.008564505	950.1598	190-194
CTGCCCC	20	2.4355314E-4	445.67755	180-184
CCGTCCC	30	4.5330584E-4	361.9656	185-189
TAACGCC	25	7.353991E-4	308.33353	185-189

R_2011_07_16_16_46_21_user_SMA-12_Auto_SMA-12_14.fastq Fa... file:///C:/Documents and Settings/14411911/My Documents/Ion Torrent ...

Sequence	Count	PValue	Obs/Exp Max	Max Obs/Exp Position
AAACGGG	55	7.785276E-4	301.5383	175-179
AGGCATT	85	0.0014297593	245.753	185-189
ATTCCGA	310	0.0018548601	245.20251	190-194
GCCCCCG	20	0.0017315368	231.9034	180-184
CCTGCCC	40	0.0019419442	222.83878	180-184

Produced by [FastQC](#) (version 0.11.2)

6-3: The Ion Torrent run report and FastQC report of *Zygothymus foetidus*.

CAF - Stellenbosch - Torrent Browser

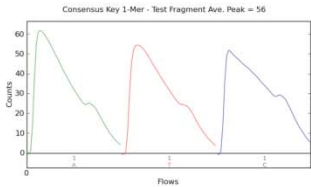
http://146.232.41.138/output/Home/Auto_SMA-14chl_16_0...

Report for Auto_SMA-14chl_16

Library Summary
Test Fragment Report
Test Fragment Summary

Test Fragment
TF_D

Percent (50AQ17 / Num)
7%

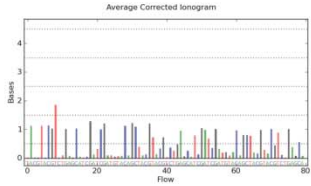
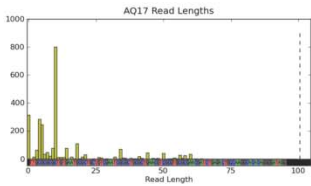


Test Fragment - TF_D

Quality Metrics

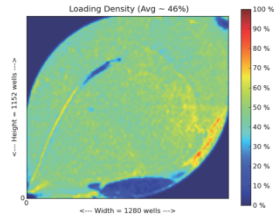
TF Name	TF_D
TF Seq	TTGCGCGCGCTGTGAATGCGCTGCTGCGAATCGCGCTGCGCTGAACGTC GCGTGCAGCAACGATCTGAGACTGCCAAGGCACACAGGGATAGG
Num	2,669
Avg Q17 read length	14
50AQ17	188

Graphs



Ion Sphere™ Particle (ISP) Identification Summary

Total Addressable Wells	Count	Percentage
• Wells with ISPs	1,262,519	
• Live ISPs	588,702	47%
• Test Fragment ISPs	52,998	9%
• Library ISPs	5,231	10%
	47,767	90%
Library ISPs / Percent Enrichment	Count	Percentage
• Filtered: Too short	47,767	8%
• Filtered: Keypass failure	12,794	27%
• Filtered: Mixed / Polyclonal	23,657	50%
• Filtered: Low Signal	5,650	12%
• Filtered: Poor Signal Profile	1,400	3%
• Filtered: 3' Adapter trim	3,464	7%
• Filtered: 3' Quality trim	0	<1%
• Final Library Reads	0	<1%
	650	1%



Report Information

Analysis Info

Run Name	R_2011_08_11_17_30_02_user_SMA-14chl
Run Date	2011-08-11 17:30:02
Analysis Name	Auto_SMA-14chl_16
Analysis Date	2011-10-18

CAF - Stellenbosch - Torrent Browser

http://146.232.41.138/output/Home/Auto_SMA-14chl_16_0...

Analysis Cycles	65
Project	chloroplastseq
Sample	zyg
Library	Library:none
PGM	SmartBlue
Chip Check	Passed
Chip Type	"314R"
Notes	314 chip for 2nd chloroplast sample
Flow Order	TACGTACGCTGAGCATCGATCGATGTACAGC
Library Key	TCAG

Software Version

Host	Stellenboschpgm1
Analysis	1.52-13
Alignment	1.47-6
Dbreports	1.106-18
Tmap	0.0.28-1
Docs	1.28-3
Tsconfig	1.5-10
Referencelibrary	1.6-2

File Links

[Library Sequence \(SFF\)](#)
[Library Sequence \(FASTQ\)](#)
[Full Library Alignments \(BAM\)](#)
[Library Alignments \(BAM Index\)](#)
[Test Fragments \(SFF\)](#)
[PDF of this Report](#)
[Customer Support Archive](#)

Plugin Summary

Select Plugins To Run	Refresh Plugin Status
-----------------------	-----------------------

[Request Support](#) | [Help/Documentation](#) | [Terms of Use](#)
 Copyright © 2012 Life Technologies Corporation
 This product should be used for research use only












R_2011_08_11_21_09_29_user_SMA-15_Auto_SMA-15_17.fastq Fa... file:///C:/Documents and Settings/14411911/My Documents/Ion Torrent ...

FastQC Report

Summary

Fri 20 Feb 2015

R_2011_08_11_21_09_29_user_SMA-15_Auto_SMA-15_17.fastq

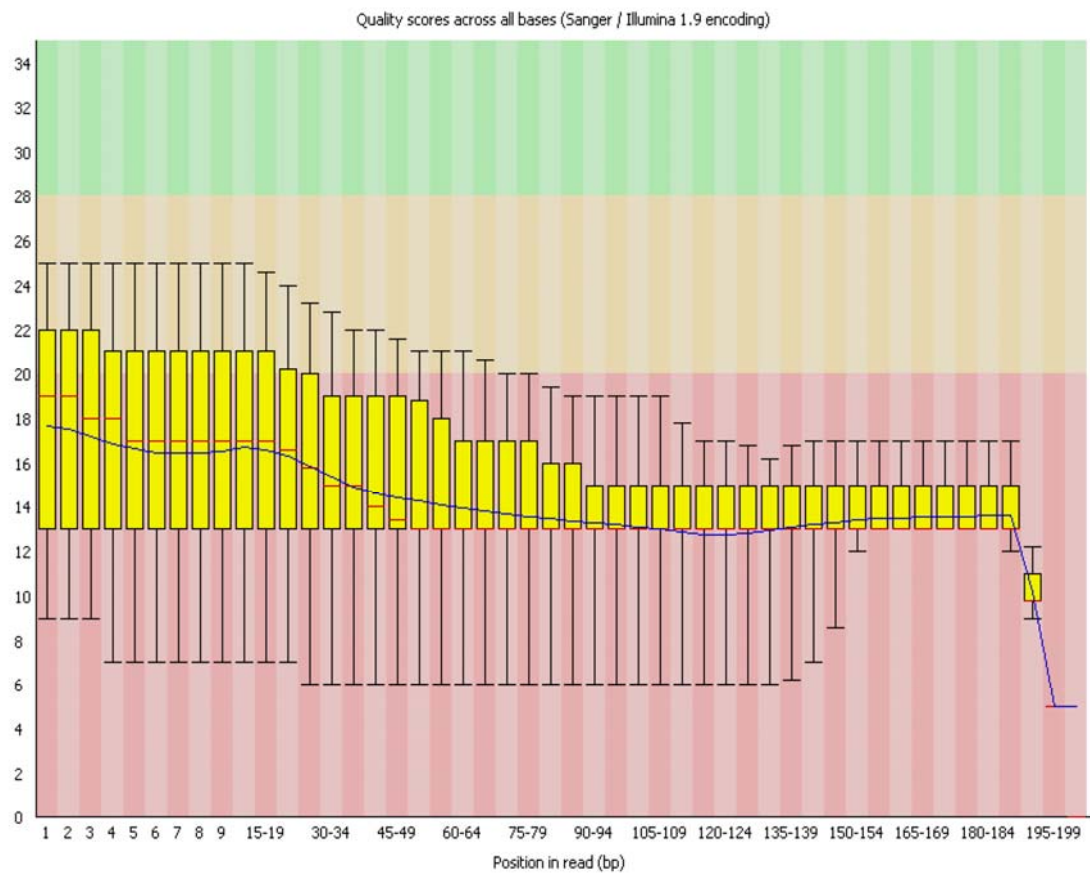
-  [Basic Statistics](#)
-  [Per base sequence quality](#)
-  [Per sequence quality scores](#)
-  [Per base sequence content](#)
-  [Per sequence GC content](#)
-  [Per base N content](#)
-  [Sequence Length Distribution](#)
-  [Sequence Duplication Levels](#)
-  [Overrepresented sequences](#)
-  [Adapter Content](#)
-  [Kmer Content](#)

Basic Statistics

Measure	Value
Filename	R_2011_08_11_21_09_29_user_SMA-15_Auto_SMA-15_17.fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	353413
Sequences flagged as poor quality	0
Sequence length	5-203
%GC	41

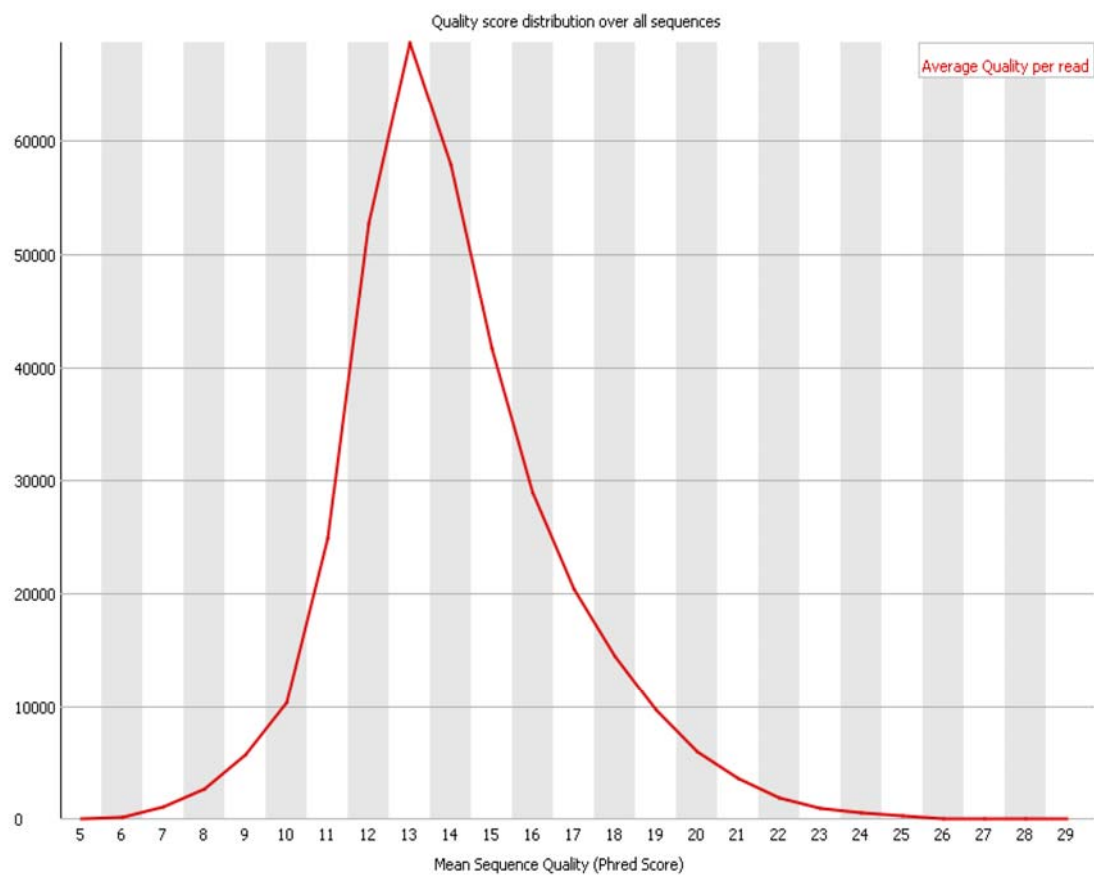
Per base sequence quality

R_2011_08_11_21_09_29_user_SMA-15_Auto_SMA-15_17.fastq Fa... file:///C:/Documents and Settings/14411911/My Documents/Ion Torrent ...



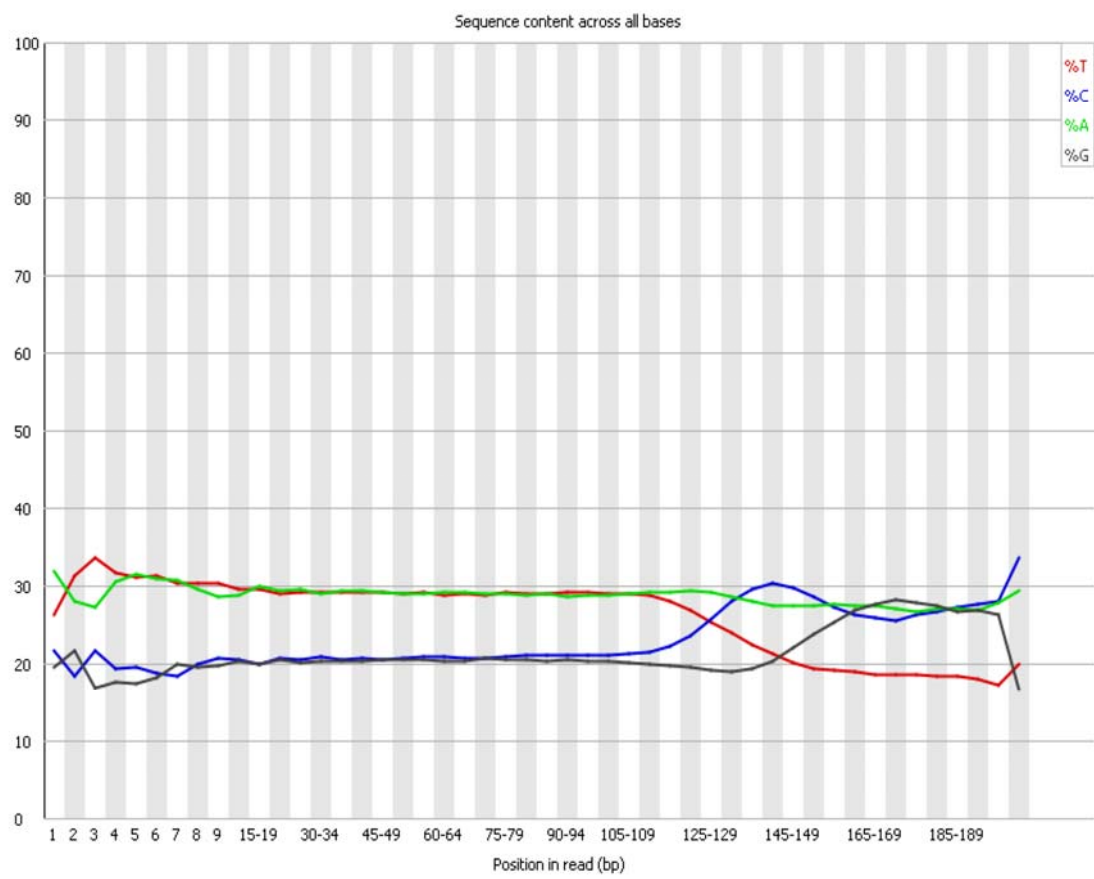
 **Per sequence quality scores**

R_2011_08_11_21_09_29_user_SMA-15_Auto_SMA-15_17.fastq Fa... file:///C:/Documents and Settings/14411911/My Documents/Ion Torrent ...



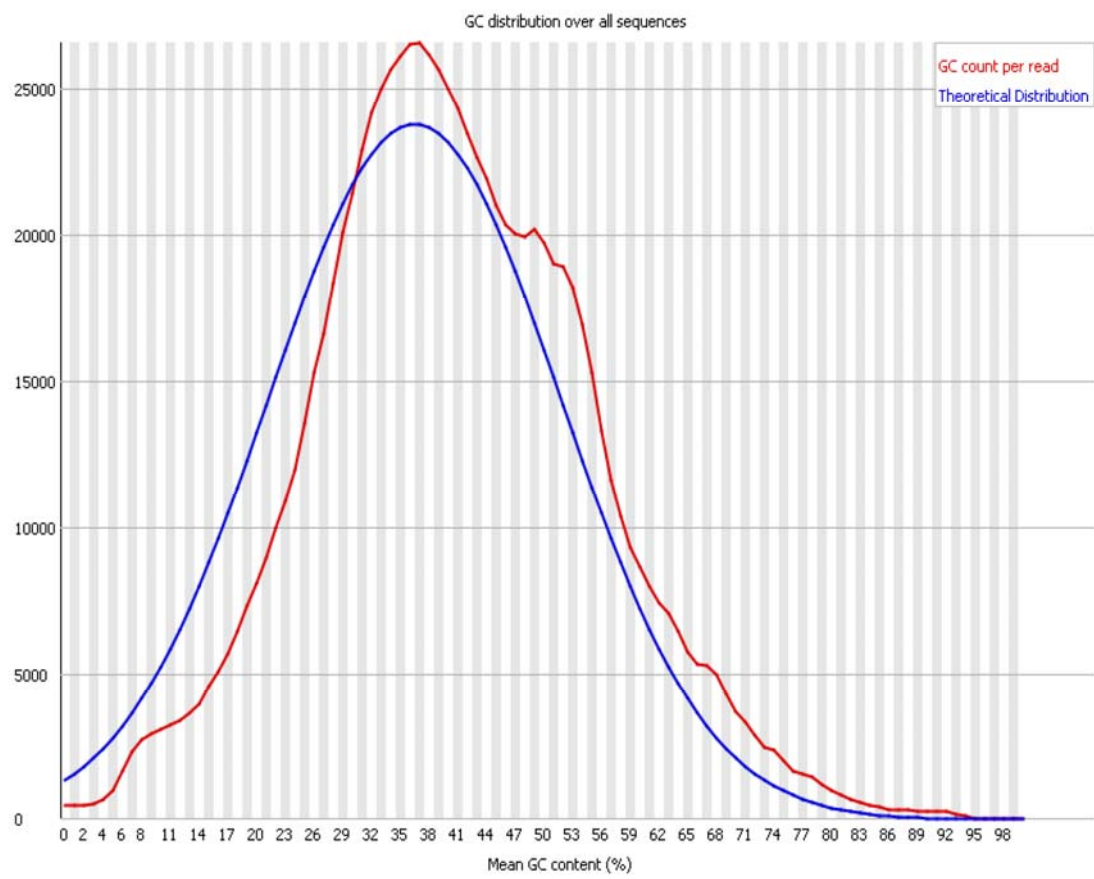
! Per base sequence content

R_2011_08_11_21_09_29_user_SMA-15_Auto_SMA-15_17.fastq Fa... file:///C:/Documents and Settings/14411911/My Documents/Ion Torrent ...



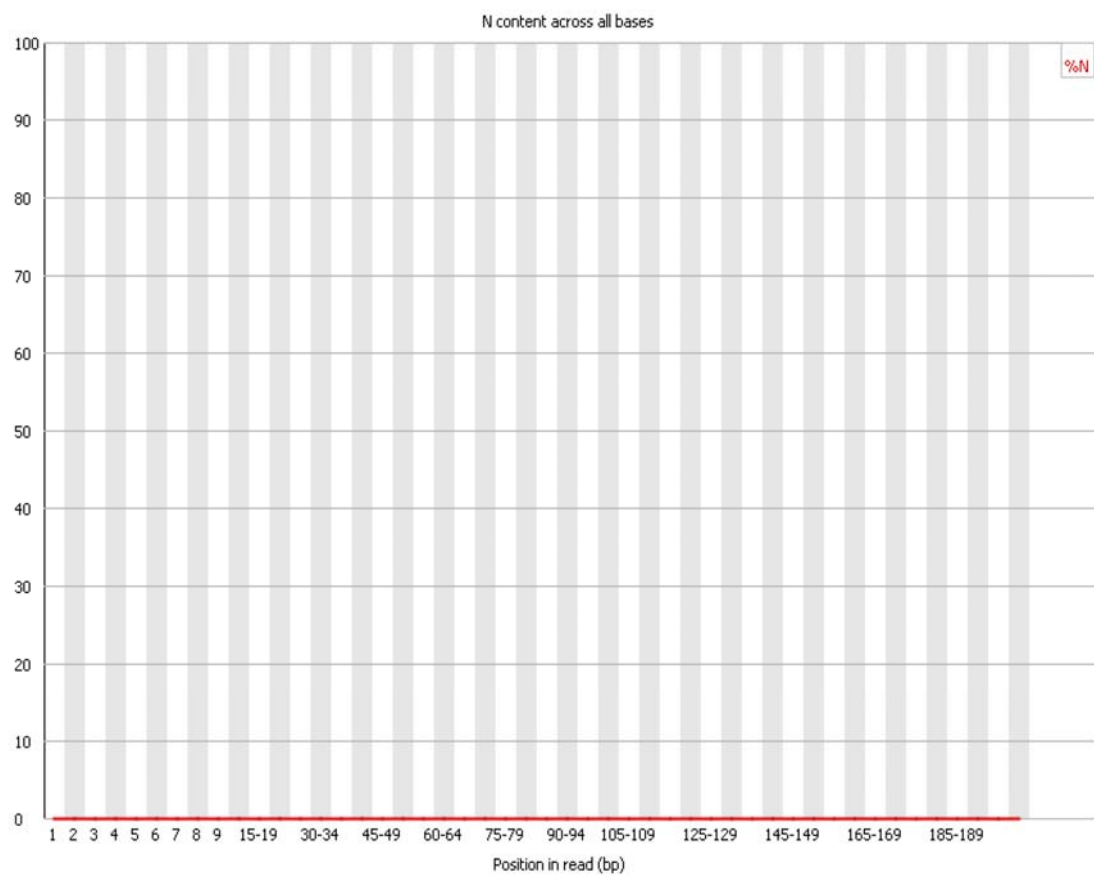
! Per sequence GC content

R_2011_08_11_21_09_29_user_SMA-15_Auto_SMA-15_17.fastq Fa... file:///C:/Documents and Settings/14411911/My Documents/Ion Torrent ...



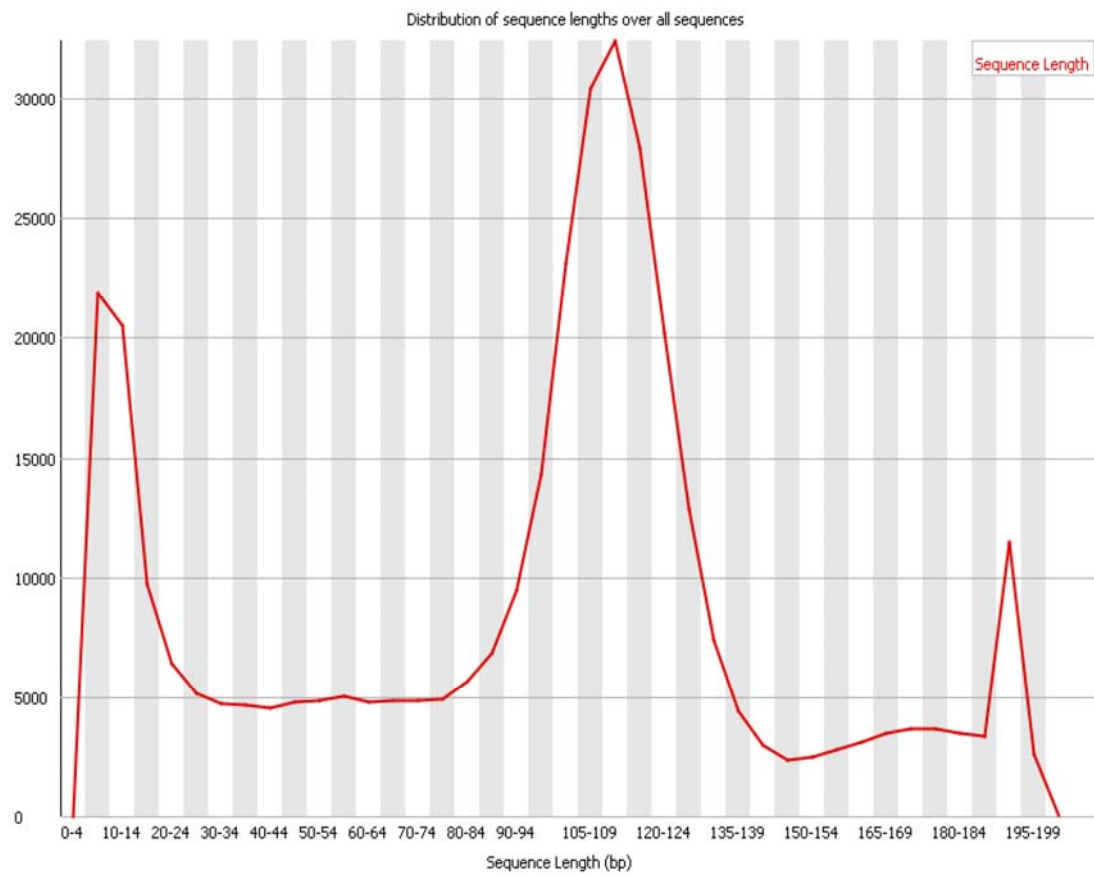
✓ Per base N content

R_2011_08_11_21_09_29_user_SMA-15_Auto_SMA-15_17.fastq Fa... file:///C:/Documents and Settings/14411911/My Documents/Ion Torrent ...



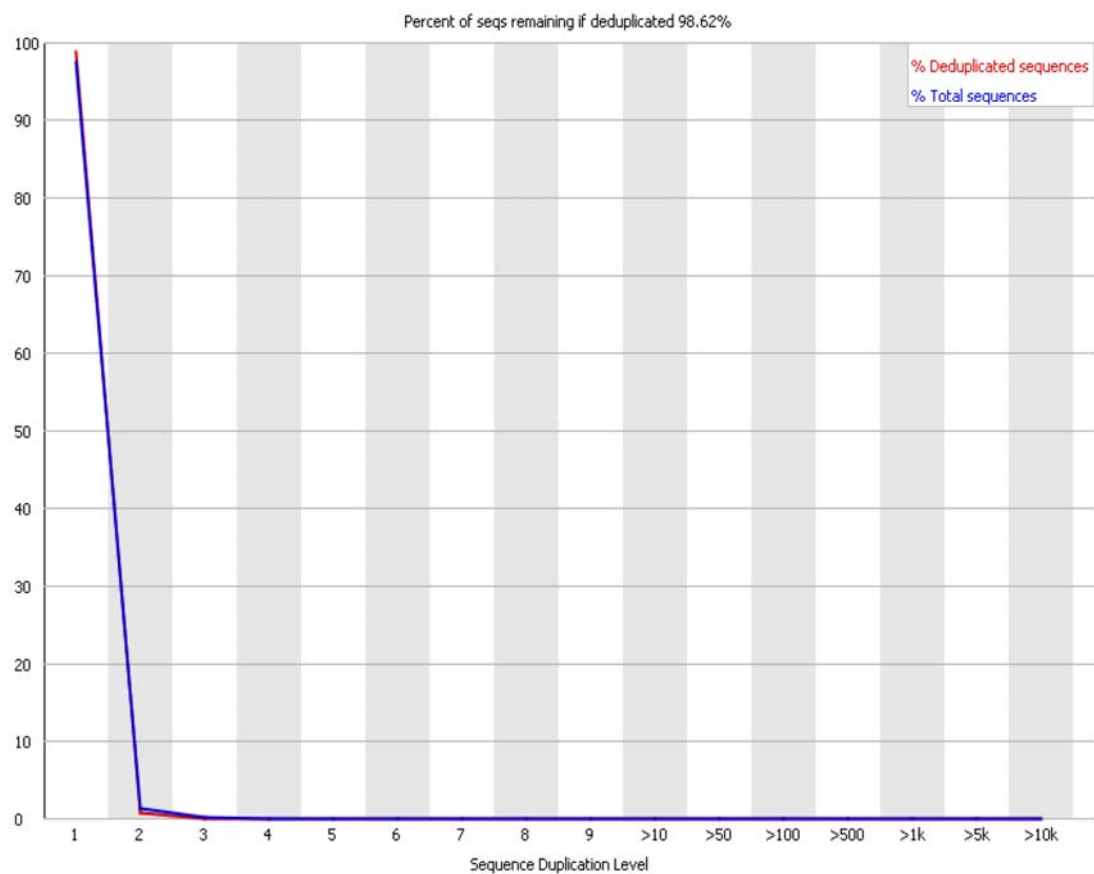
! Sequence Length Distribution

R_2011_08_11_21_09_29_user_SMA-15_Auto_SMA-15_17.fastq Fa... file:///C:/Documents and Settings/14411911/My Documents/Ion Torrent ...



✔ Sequence Duplication Levels

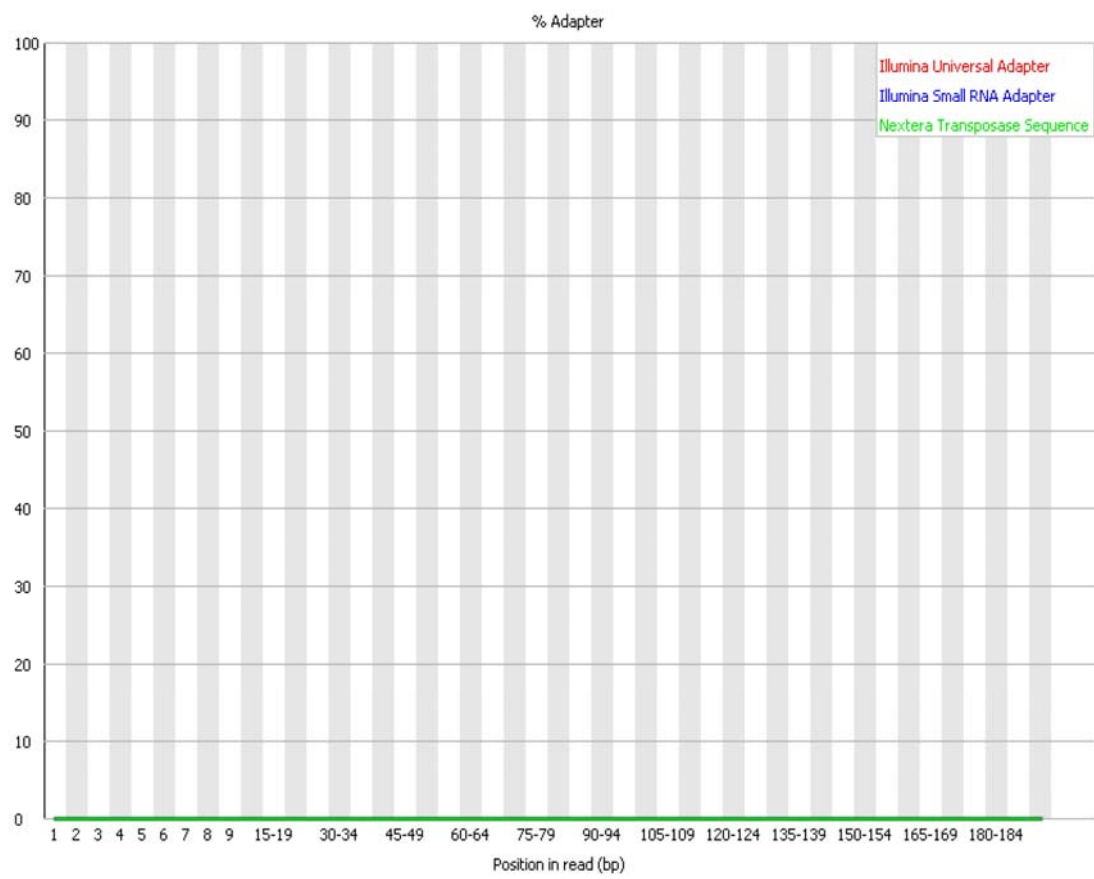
R_2011_08_11_21_09_29_user_SMA-15_Auto_SMA-15_17.fastq Fa... file:///C:/Documents and Settings/14411911/My Documents/Ion Torrent ...



 **Overrepresented sequences**
No overrepresented sequences

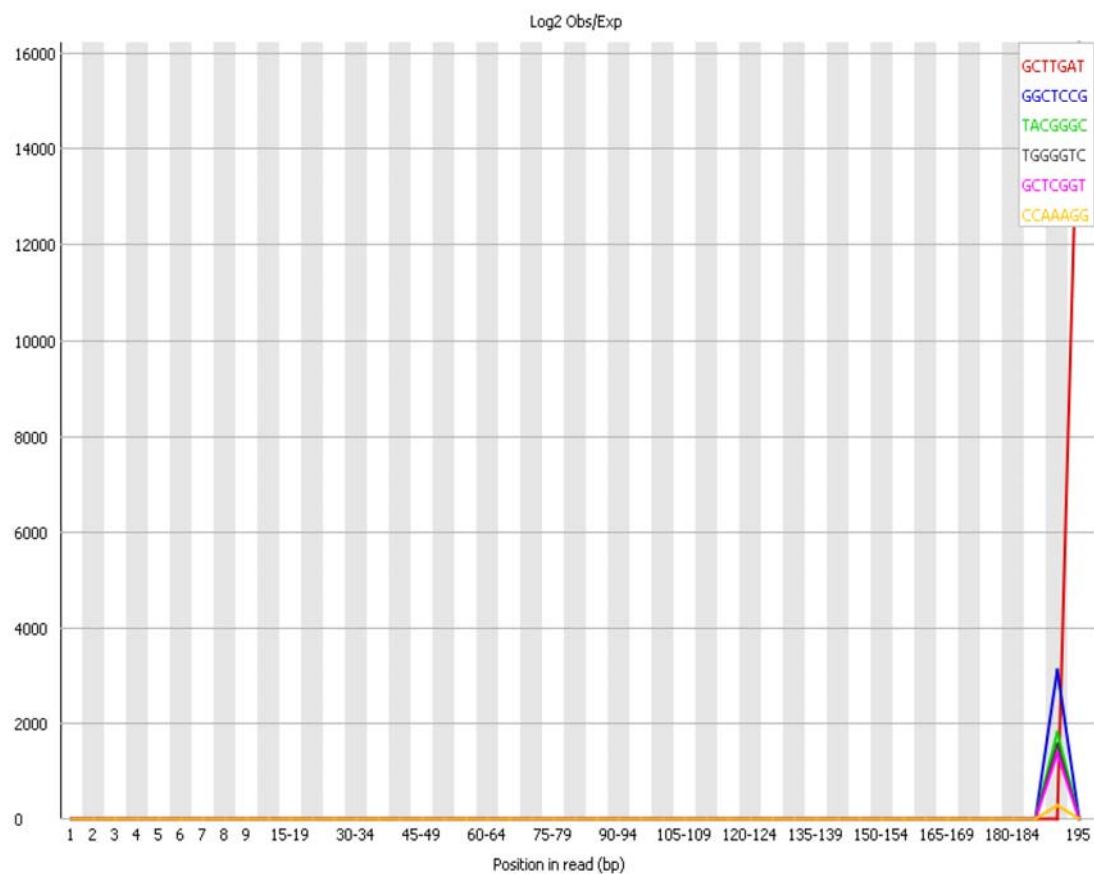
 **Adapter Content**

R_2011_08_11_21_09_29_user_SMA-15_Auto_SMA-15_17.fastq Fa... file:///C:/Documents and Settings/14411911/My Documents/Ion Torrent ...



 **Kmer Content**

R_2011_08_11_21_09_29_user_SMA-15_Auto_SMA-15_17.fastq Fa... file:///C:/Documents and Settings/14411911/My Documents/Ion Torrent ...



Sequence	Count	PValue	Obs/Exp Max	Max Obs/Exp Position
GCTTGAT	190	0.0	16203.184	195
GGCTCCG	35	8.003028E-4	3141.4336	190-194
TACGGGC	60	0.0023515793	1832.5029	190-194
TGGGGTC	70	0.0032005804	1570.7168	190-194
GCTCGGT	80	0.0041801143	1374.3772	190-194
CCAAAGG	355	0.0013348286	309.7188	190-194
AGCATTG	365	0.0027946716	301.23337	190-194
GAAACGG	130	0.0028209172	196.5276	185-189
ACGGAGC	85	0.008463361	136.41765	165-169
AAAGGGC	465	0.0023800544	112.21451	190-194
CATTCCG	190	1.9404513E-5	92.722084	185-189
GAAGGTA	325	7.974502E-4	78.611046	185-189
ATTCCGT	180	0.0024960465	67.87048	185-189
CCTGCCC	150	0.0017828249	65.78216	180-184
AAGGGCC	740	2.3646862E-11	65.004326	190-194

R_2011_08_11_21_09_29_user_SMA-15_Auto_SMA-15_17.fastq Fa... file:///C:/Documents and Settings/14411911/My Documents/Ion Torrent ...

Sequence	Count	PValue	Obs/Exp Max	Max Obs/Exp Position
GTCCCCC	95	1.450717E-4	61.90328	170-174
CATTCGC	85	0.0036657415	61.666656	175-179
CGGAATG	215	0.0012118455	59.41532	185-189
ACCCGTA	135	0.0049449867	57.154087	160-164
GGTACCC	120	0.009419403	57.138176	170-174

Produced by [FastQC](#) (version 0.11.2)

6-4: The Ion Torrent run report and FastQC report of *Zygophyllum turbinatum*.

CAF - Stellenbosch - Torrent Browser

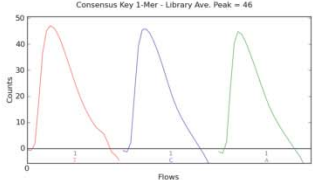
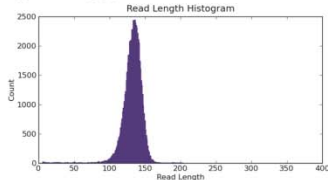
http://146.232.41.138/output/Home/Zygophyllum_096/Defau...

Report for *Zygophyllum*

Library Summary

Based on Predicted Per-Base Quality Scores - Independent of Alignment

Total Number of Bases [Mbp]	8.58
• Number of Q17 Bases [Mbp]	7.51
• Number of Q20 Bases [Mbp]	7.07
Total Number of Reads	65,462
Mean Length [bp]	131
Longest Read [bp]	203



Reference Genome Information

There was an alignment error. For details see the [Report Log](#)
Unable to process alignment for genome, because the **none** reference library was not found.

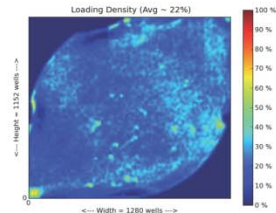
Read Alignment Distribution

alignTable.txt not found

Test Fragment Report

Ion Sphere™ Particle (ISP) Identification Summary

Total Addressable Wells	Count	Percentage
• Wells with ISPs	1,262,747	
• Live ISPs	280,759	22%
• Test Fragment ISPs	197,994	71%
• Library ISPs	15,927	8%
	182,067	92%
Library ISPs / Percent Enrichment	Count	Percentage
• Filtered: Polyclonal	182,067	69%
• Filtered: Primer dimer	50,004	27%
• Filtered: Low quality	13	<1%
• Final Library Reads	66,588	37%
	65,462	36%



Report Information

Analysis Info

Run Name	R_2011_11_19_16_02_37_user_SMA-26-Zygophyllum_turbinatum
Run Date	2011-11-19 16:02:37
Analysis Name	Zygophyllum
Analysis Date	2012-01-26
Analysis Cycles	8
Analysis Flows	260
Project	Dirk_Bellstedt_chloroplast_zygophyllum
Sample	chloroplast_zygophyllum
Library	none
PGM	SmartBlue
Chip Check	Passed
Chip Type	314R
Chip Data	single
Notes	
Barcode Set	
Flow Order	TACGTACGTCTGAGCATCGATCGATGTACAGC
Library Key	TCAG

Software Version

Torrent Suite	2.0
---------------	-----

CAF - Stellenbosch - Torrent Browser

http://146.232.41.138/output/Home/Zygophyllum_096/Defau...

host	stellenboschpgm1
ion-alignment	2.0.3-1
ion-analysis	2.0.7-1
ion-dbreports	2.0.15-1
ion-docs	2.0.5-1
ion-gpu	1.2-1
ion-onetouchupdater	2.0.1-1
ion-pgmupdates	2.0.4
ion-plugins	2.0.15-1
ion-publishers	2.0.8-1
ion-referencelibrary	1.6-2
ion-rsmts	2.0.4-1
ion-sampledats	1.2-0
ion-torrentR	2.0.3-1
ion-tsconfig	2.0.13-1
ion-tsups	1.0-1
ion-usbmount	0.0.19.1ion1
tmap	0.2.3-1

File Links

[Library Sequence \(SFF\)](#)
[Library Sequence \(FASTQ\)](#)
[Full Library Alignments \(BAM\)](#)
[Full Library Alignments \(BAI\)](#)
[Test Fragments \(SFF\)](#)
[PDF of this Report](#)
[Customer Support Archive](#)

Plugin Summary

Select Plugins To Run	Refresh Plugin Status	
-----------------------	-----------------------	--

Alignment Completed

Re-Alignment to Genome: *Corynocarpus laevigata* CP

[Download Output Files](#)

Based on Re-Alignment to Provided Reference

	AQ17	AQ20	Perfect
Total Number of Bases [Mbp]	0.07	0.05	0.04
Mean Length [bp]	56	45	40
Longest Alignment [bp]	157	157	157
Mean Coverage Depth	0.50x	0.30x	0.30x
Percentage of Library Covered	63%	62%	62%

Re-Alignment Read Distribution

Read Length [bp]	Reads	Unmapped	Excluded	Clipped	Perfect	1 mismatch	≥2 mismatches
50	45,149	39,189	3	0	241	390	5,326
100	44,383	38,488	1	988	38	81	4,787
150	1,902	1,638	0	179	2	0	83

Assembler

[Assembler.html](#)

Unknown

[Request Support](#) | [Help/Documentation](#) | [Terms of Use](#)
 Copyright © 2012 Life Technologies Corporation
 This product should be used for research use only












R_2011_11_19_16_02_37_user_SMA-26-Zygophyllum_turbinatum_Zy... file:///C:/Documents and Settings/14411911/My Documents/Ion Torrent ...

FastQC Report

Fri 20 Feb 2015

R_2011_11_19_16_02_37_user_SMA-26-Zygophyllum_turbinatum_Zygophyllum.fastq

Summary

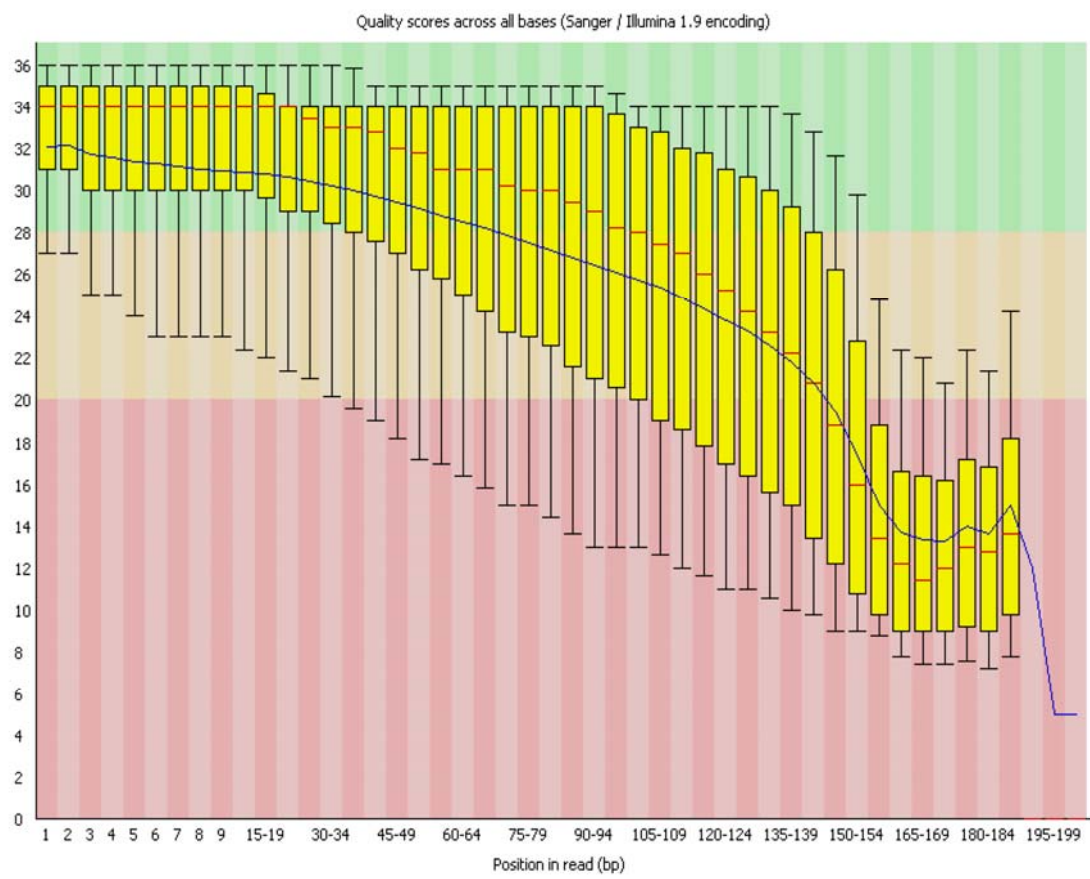
-  [Basic Statistics](#)
-  [Per base sequence quality](#)
-  [Per sequence quality scores](#)
-  [Per base sequence content](#)
-  [Per sequence GC content](#)
-  [Per base N content](#)
-  [Sequence Length Distribution](#)
-  [Sequence Duplication Levels](#)
-  [Overrepresented sequences](#)
-  [Adapter Content](#)
-  [Kmer Content](#)

Basic Statistics

Measure	Value
Filename	R_2011_11_19_16_02_37_user_SMA-26-Zygophyllum_turbinatum_Zygophyllum.fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	65462
Sequences flagged as poor quality	0
Sequence length	5-203
%GC	35

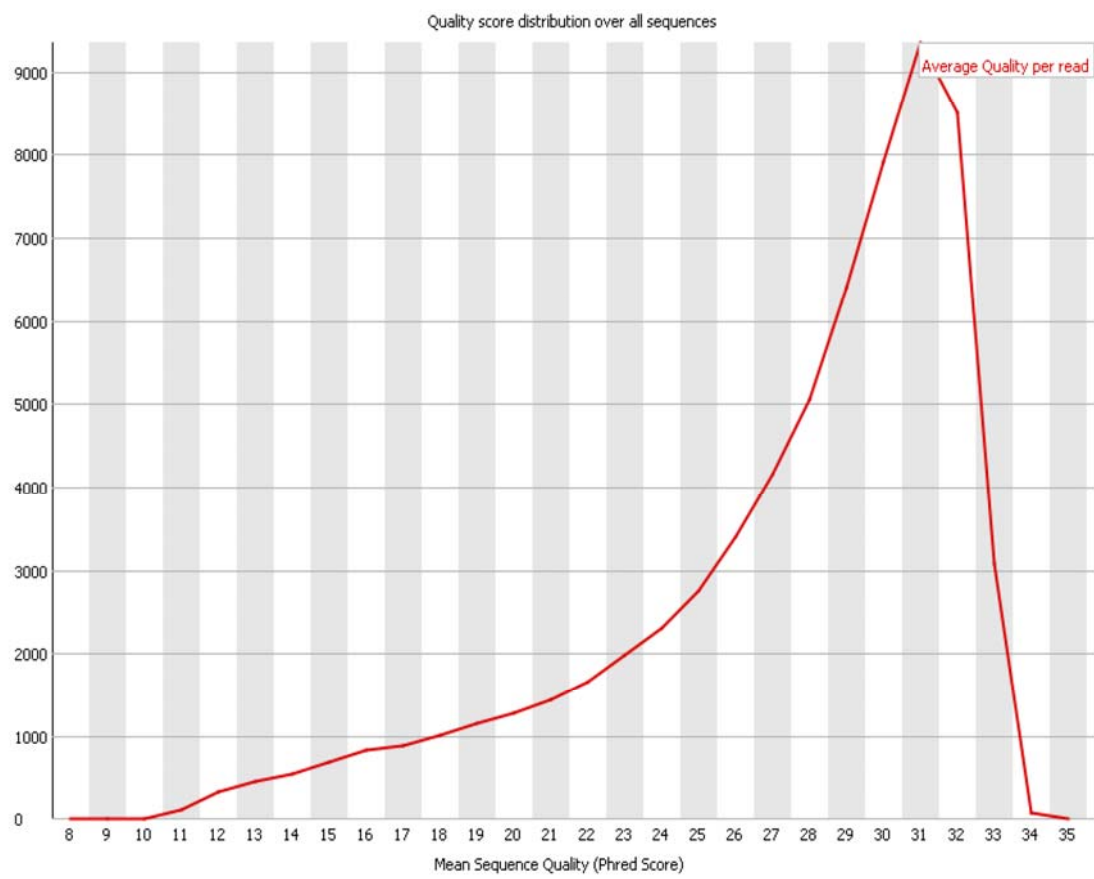
Per base sequence quality

R_2011_11_19_16_02_37_user_SMA-26-Zygophyllum_turbinatum_Zy... file:///C:/Documents and Settings/14411911/My Documents/Ion Torrent ...



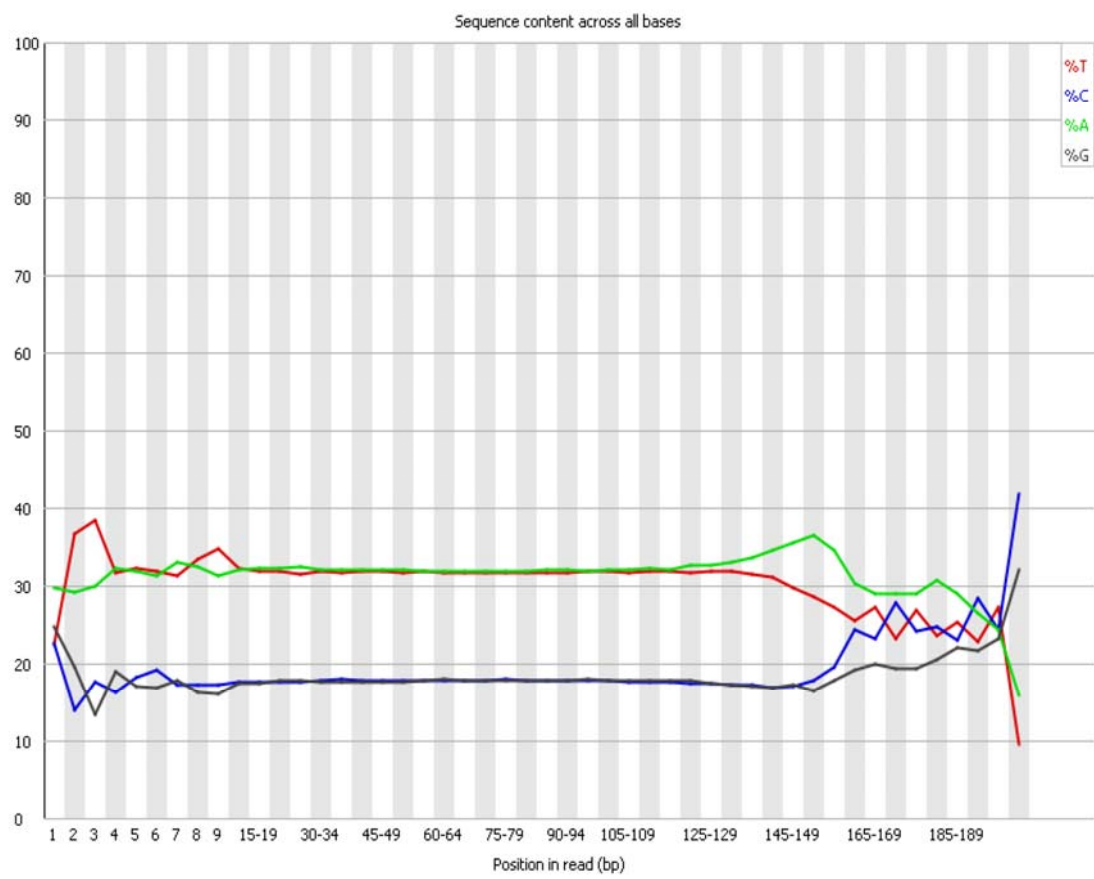
Per sequence quality scores

R_2011_11_19_16_02_37_user_SMA-26-Zygophyllum_turbinatum_Zy... file:///C:/Documents and Settings/14411911/My Documents/Ion Torrent ...



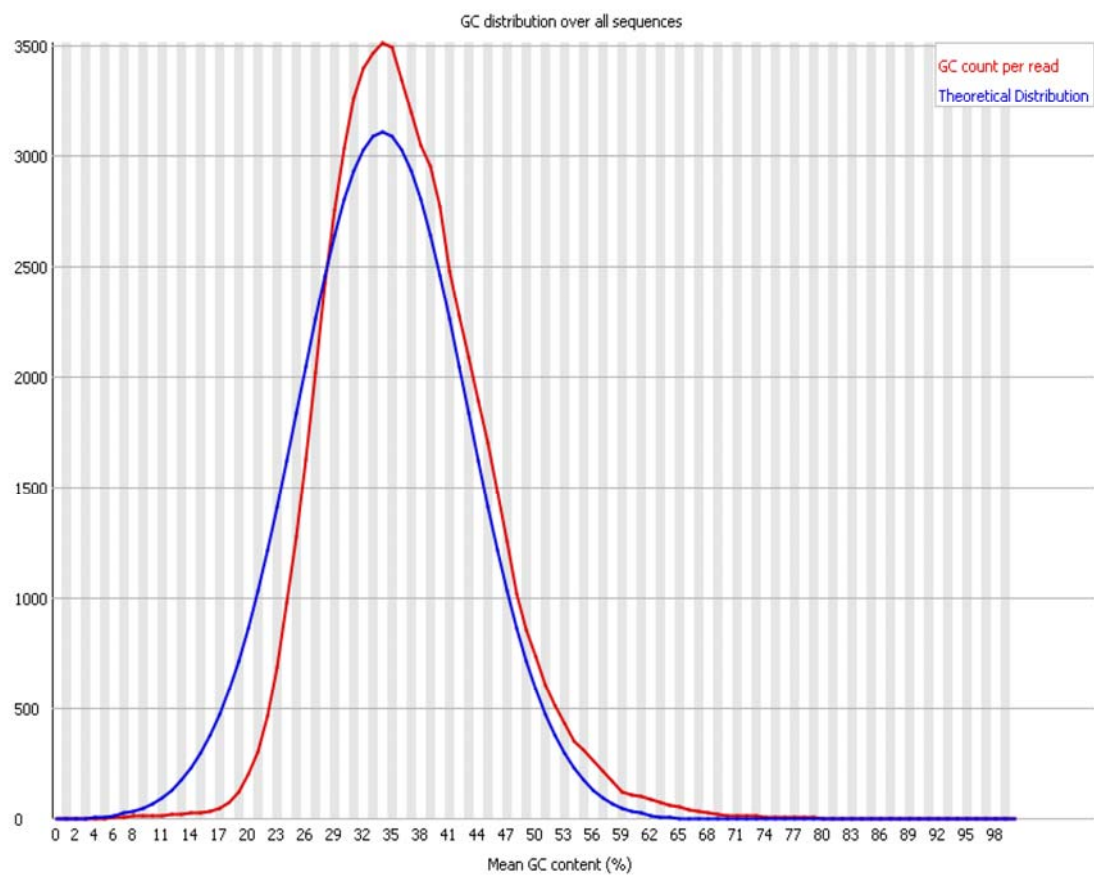
✖ Per base sequence content

R_2011_11_19_16_02_37_user_SMA-26-Zygophyllum_turbinatum_Zy... file:///C:/Documents and Settings/14411911/My Documents/Ion Torrent ...



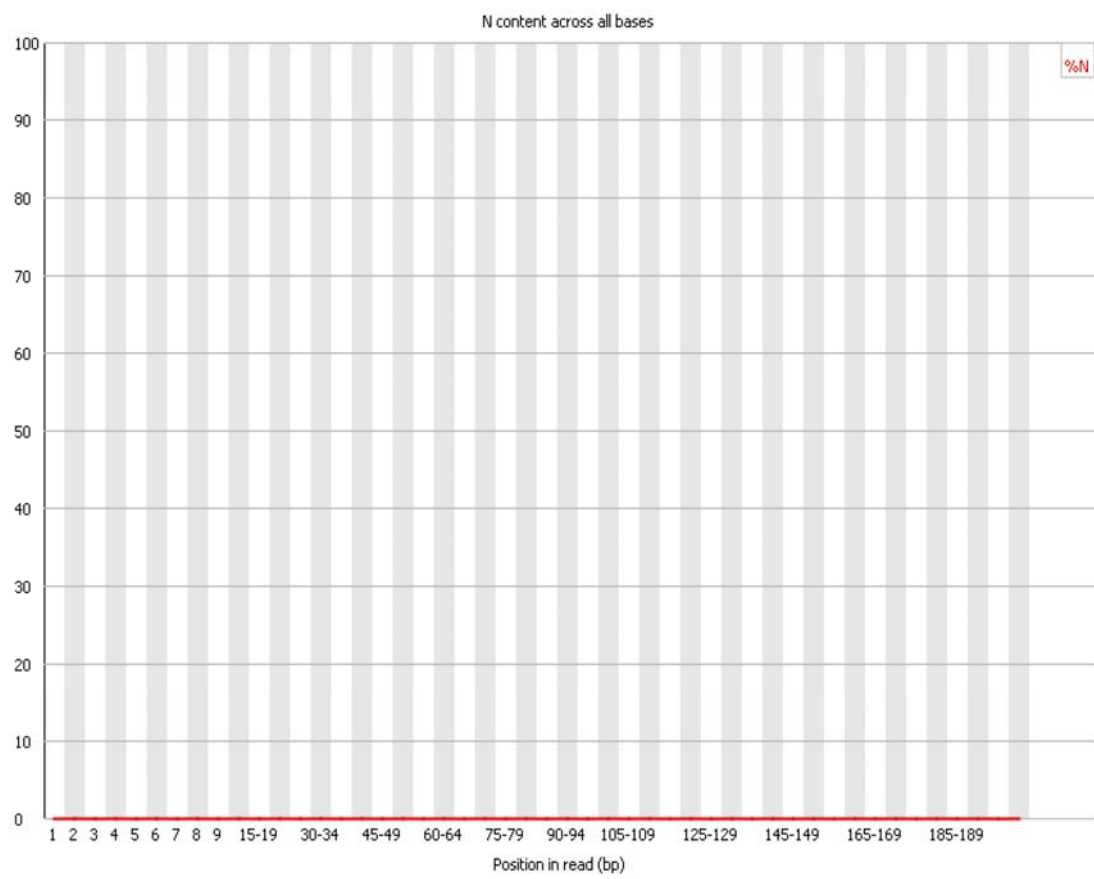
! Per sequence GC content

R_2011_11_19_16_02_37_user_SMA-26-Zygophyllum_turbinatum_Zy... file:///C:/Documents and Settings/14411911/My Documents/Ion Torrent ...



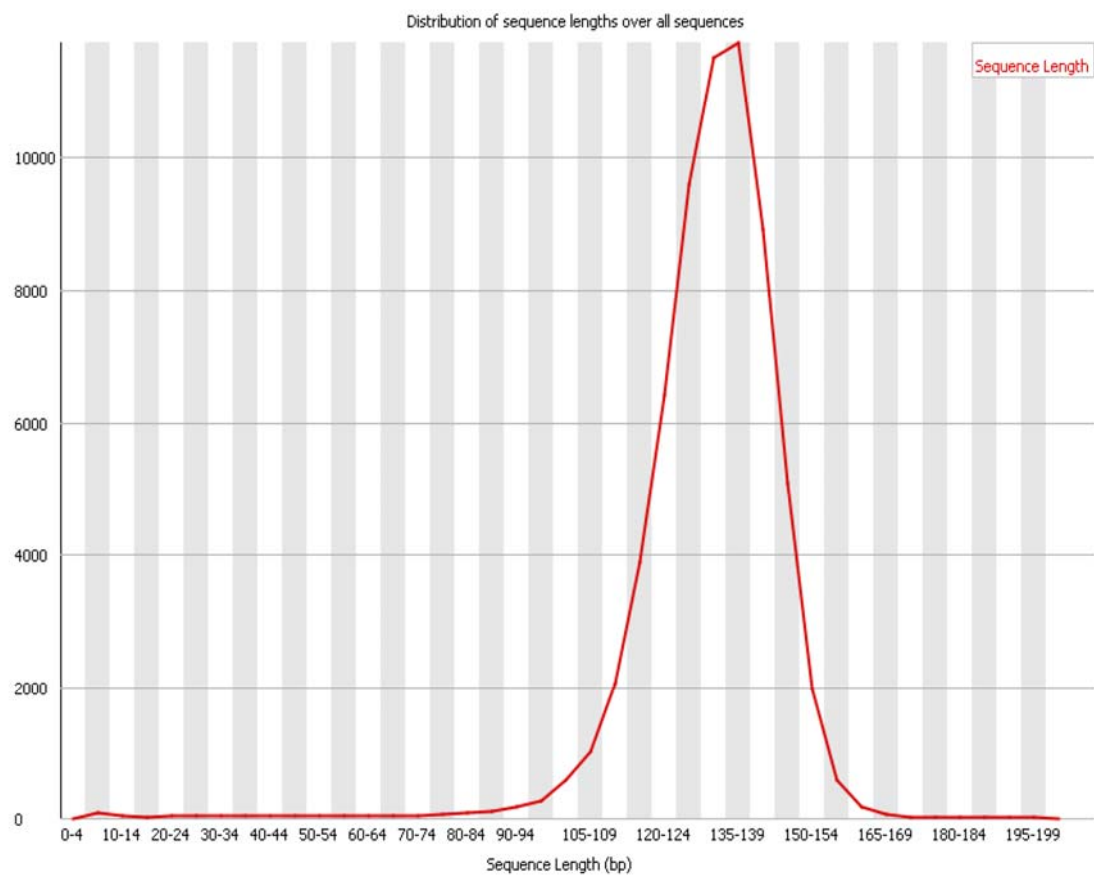
Per base N content

R_2011_11_19_16_02_37_user_SMA-26-Zygophyllum_turbinatum_Zy... file:///C:/Documents and Settings/14411911/My Documents/Ion Torrent ...



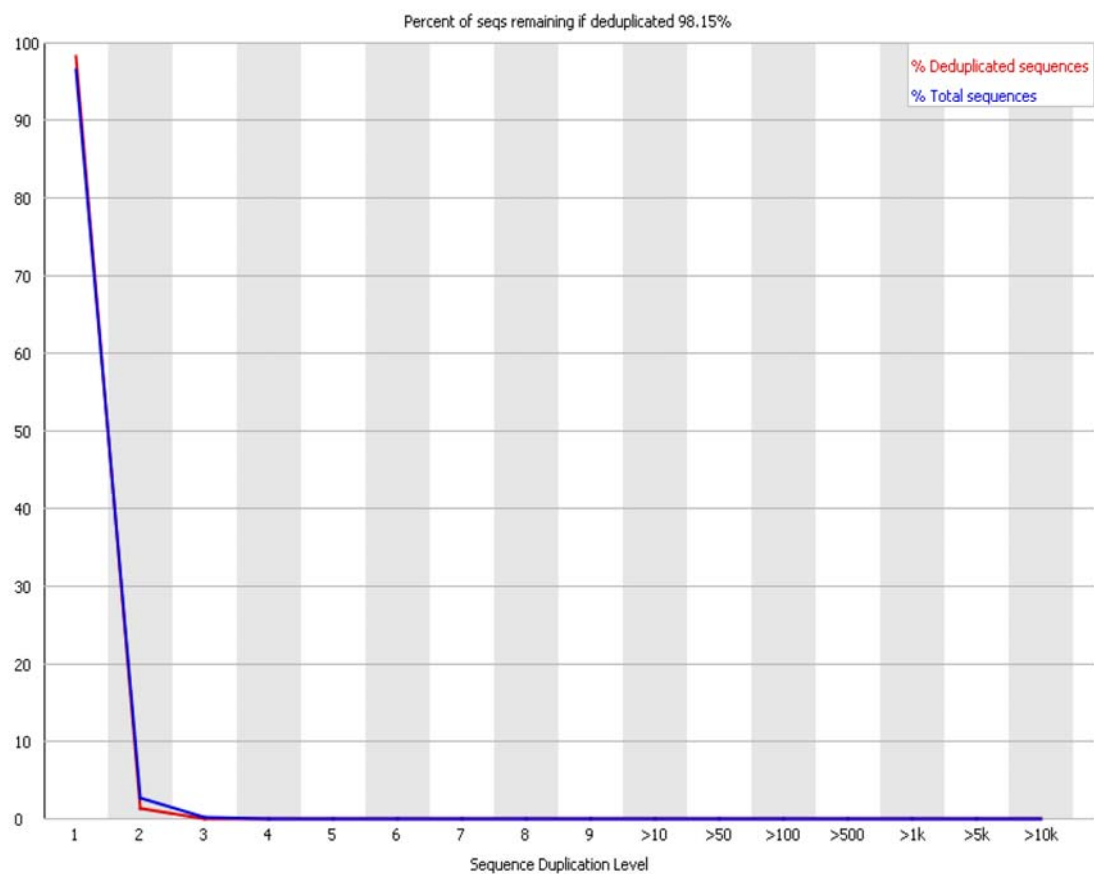
! Sequence Length Distribution

R_2011_11_19_16_02_37_user_SMA-26-Zygophyllum_turbinatum_Zy... file:///C:/Documents and Settings/14411911/My Documents/Ion Torrent ...



✔ Sequence Duplication Levels

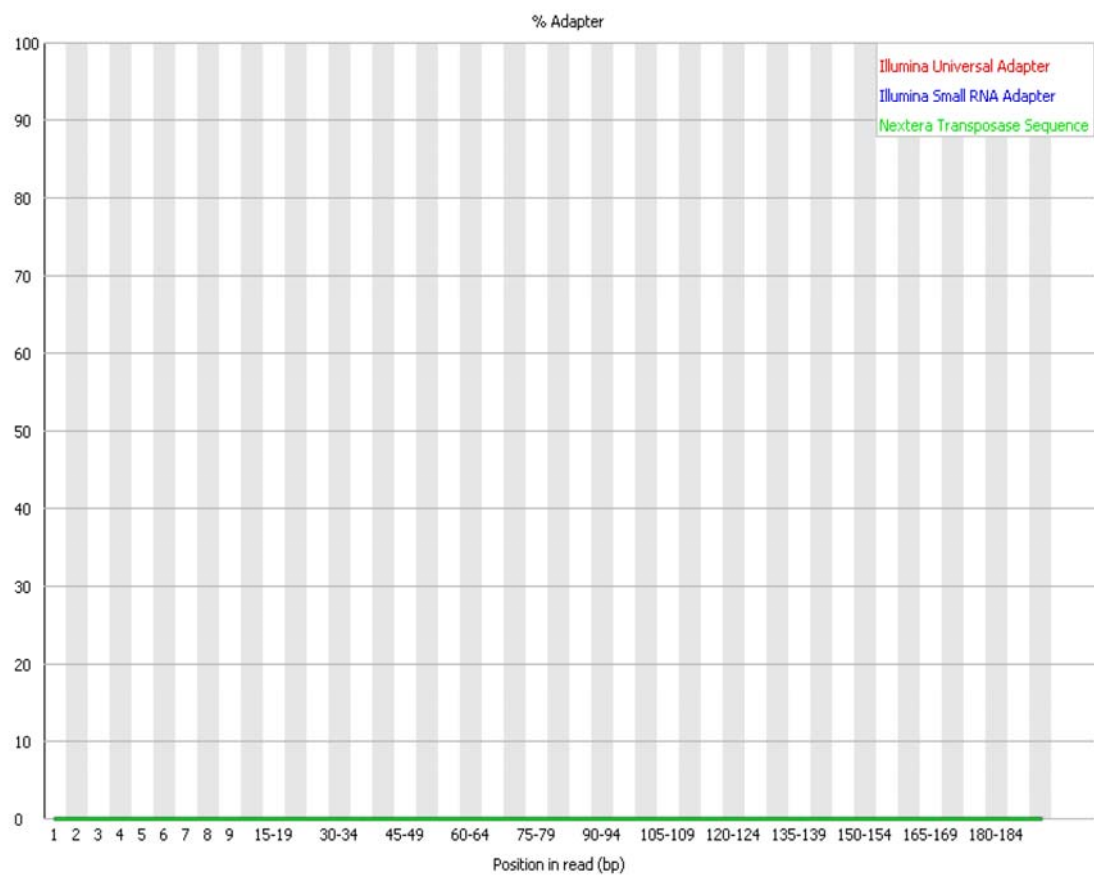
R_2011_11_19_16_02_37_user_SMA-26-Zygophyllum_turbinatum_Zy... file:///C:/Documents and Settings/14411911/My Documents/Ion Torrent ...



 **Overrepresented sequences**
No overrepresented sequences

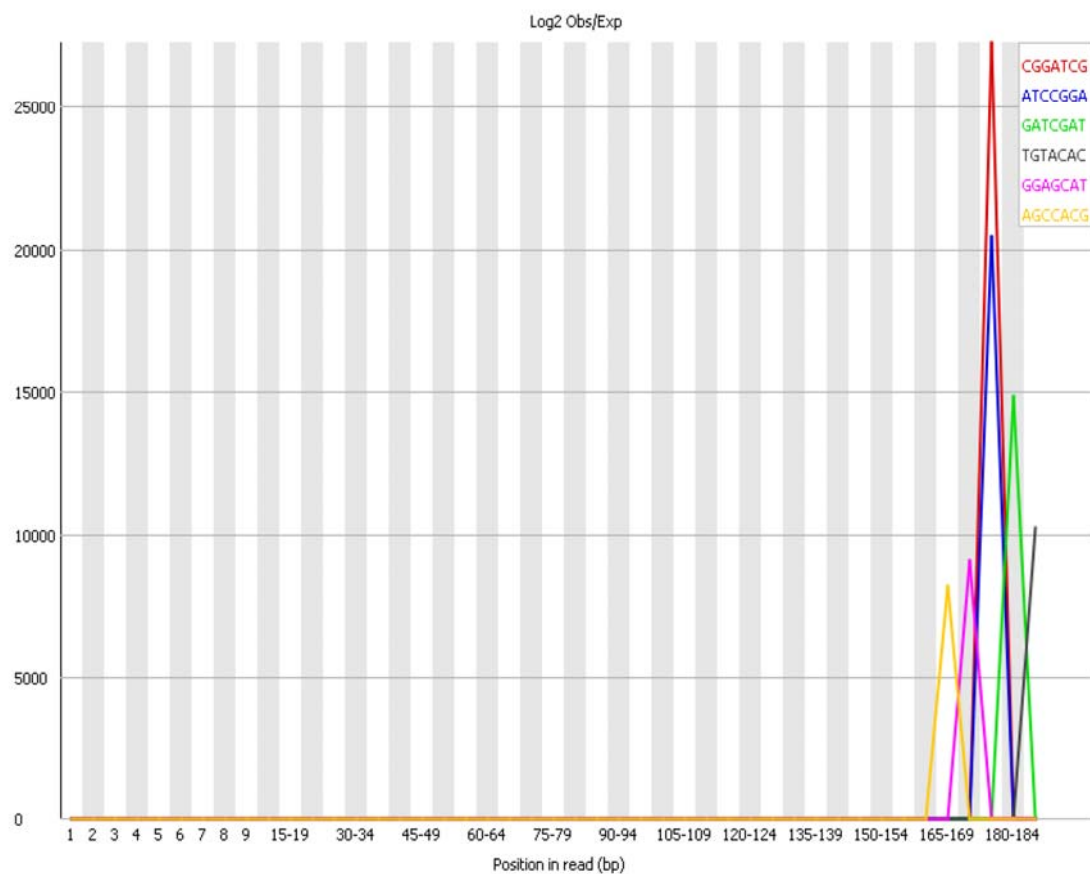
 **Adapter Content**

R_2011_11_19_16_02_37_user_SMA-26-Zygophyllum_turbinatum_Zy... file:///C:/Documents and Settings/14411911/My Documents/Ion Torrent ...



 **Kmer Content**

R_2011_11_19_16_02_37_user_SMA-26-Zygophyllum_turbinatum_Zy... file:///C:/Documents and Settings/14411911/My Documents/Ion Torrent ...



Sequence	Count	PValue	Obs/Exp Max	Max Obs/Exp Position
CGGATCG	5	9.207426E-6	27229.0	175-179
ATCCGGA	20	9.821413E-6	20421.75	175-179
GATCGAT	55	0.0	14852.182	180-184
TGTACAC	40	3.9285653E-5	10210.875	185-188
GGAGCAT	30	6.629292E-5	9076.334	170-174
AGCCACG	20	9.820933E-5	8168.7	165-169
CCGGATC	25	9.2072E-5	8168.7	175-179
GCCACGA	10	1.1048369E-4	8168.7	165-169
TCCGGAT	35	9.023166E-5	7779.7144	175-179
ATGTACA	110	0.0	7426.091	185-188
GTACACC	40	1.1785312E-4	6807.25	185-188
GGATCGA	20	1.4731158E-4	6807.25	175-179
ATCGATG	65	1.037387E-4	6283.615	180-184
ATAACGC	30	2.2096558E-4	5445.8003	160-164
CATCCGG	15	4.308565E-4	4189.077	170-174

R_2011_11_19_16_02_37_user_SMA-26-Zygophyllum_turbinatum_Zy... file:///C:/Documents and Settings/14411911/My Documents/Ion Torrent ...

Sequence	Count	PValue	Obs/Exp Max	Max Obs/Exp Position
GGGGAAT	20	4.4190593E-4	4084.35	155-159
TCCTTGG	40	3.9281807E-4	4084.35	165-169
TCGATGT	70	3.6091634E-4	3889.8572	180-184
GCATCCG	20	5.400984E-4	3713.0457	170-174
CGATGTA	55	4.4560668E-4	3713.0454	180-184

Produced by [FastQC](#) (version 0.11.2)

6-5: The Ion Torrent run report and FastQC report of *Tribulus terrestris*.

Tribulus, CAF - Stellenbosch - Torrent Browser

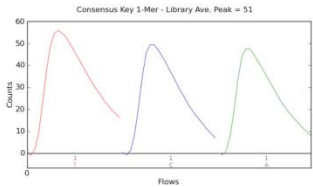
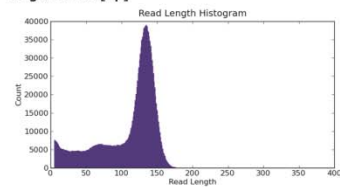
http://146.232.41.138/output/Home/Tribulus_205/Default_Re...

Report for Tribulus

Library Summary

Based on Predicted Per-Base Quality Scores - Independent of Alignment

Total Number of Bases [Mbp]	187.29
• Number of Q20 Bases [Mbp]	140.54
Total Number of Reads	1,731,969
Mean Length [bp]	108
Longest Read [bp]	200



Reference Genome Information

Genome Name	Corynocarpus laevigata chloroplast
Genome Size	159,202 bases
Genome Version	HQ207704.1
Index Version	tmap-f3

Based on Full Library Alignment to Provided Reference

	AQ20	Perfect
Total Number of Bases [Mbp]	0.15	0.13
Mean Length [bp]	52	45
Longest Alignment [bp]	160	153
Mean Coverage Depth	1.00x	0.80x
Percentage of Library Covered	100%	100%

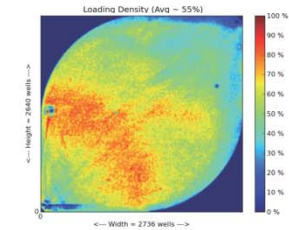
Read Alignment Distribution

Read Length [bp]	Reads	Unmapped	Excluded	Clipped	Perfect	1 mismatch	≥2 mismatches
50	1,103,800	1,093,937	5	0	914	1,213	7,731
100	876,568	868,184	5	1,457	147	306	6,469
150	64,767	63,741	1	517	2	6	500

Test Fragment Report

Ion Sphere™ Particle (ISP) Identification Summary

Total Addressable Wells	Count	Percentage
• Wells with ISPs	6,348,215	56%
• Live ISPs	3,542,752	90%
• Test Fragment ISPs	3,184,485	<1%
• Library ISPs	24,505	99%
Library ISPs / Percent Enrichment	Count	Percentage
• Filtered: Polyclonal	3,159,980	90%
• Filtered: Primer dimer	990,644	31%
• Filtered: Low quality	100	<1%
• Final Library Reads	437,267	14%
	1,731,969	55%



Report Information

Analysis Info

Run Name	R_2012_04_20_13_04_51_user_SMA-35
Run Date	2012-04-20 13:04:51
Analysis Name	Tribulus
Analysis Date	2012-08-24
Analysis Cycles	8
Analysis Flows	260
Project	chl-tribulus-terrestris

Sample	chl
Library	Claevigata_cp
PGM	SmartBlue
Chip Check	Passed
Chip Type	316D
Chip Data	single
Notes	
Barcode Set	
runID	QHJ8
Flow Order	TACGTACGTCTGAGCATCGATCGATGTACAGC
Library Key	TCAG

Software Version

Torrent_Suite	2.2
Datacollect	207
LiveView	318
Script	18.1.0
host	stellenboschpgm1
ion-alignment	2.2.4-1
ion-analysis	2.2.12-1
ion-dbreports	2.2.17-1
ion-gpu	1.2-1
ion-pipeline	2.2.12-1
ion-plugins	2.2.16-1
ion-torrentR	2.2.8-1
ion-tsups	1.0-1

File Links

- [Library Sequence \(SFF\)](#)
- [Library Sequence \(FASTQ\)](#)
- [Full Library Alignments \(BAM\)](#)
- [Full Library Alignments \(BAI\)](#)
- [Test Fragments \(SFF\)](#)
- [PDF of this Report](#)
- [Customer Support Archive](#)

Plugin Summary

Select Plugins To Run	Refresh Plugin Status
-----------------------	-----------------------

Request Support | Help/Documentation | Terms of Use
Copyright © 2012 Life Technologies Corporation
This product should be used for research use only












R_2012_04_20_13_04_51_user_SMA-35_Auto_SMA-35_37.fastq Fa... file:///C:/Documents and Settings/14411911/My Documents/Ion Torrent ...

FastQC Report

Summary

Fri 20 Feb 2015

R_2012_04_20_13_04_51_user_SMA-35_Auto_SMA-35_37.fastq

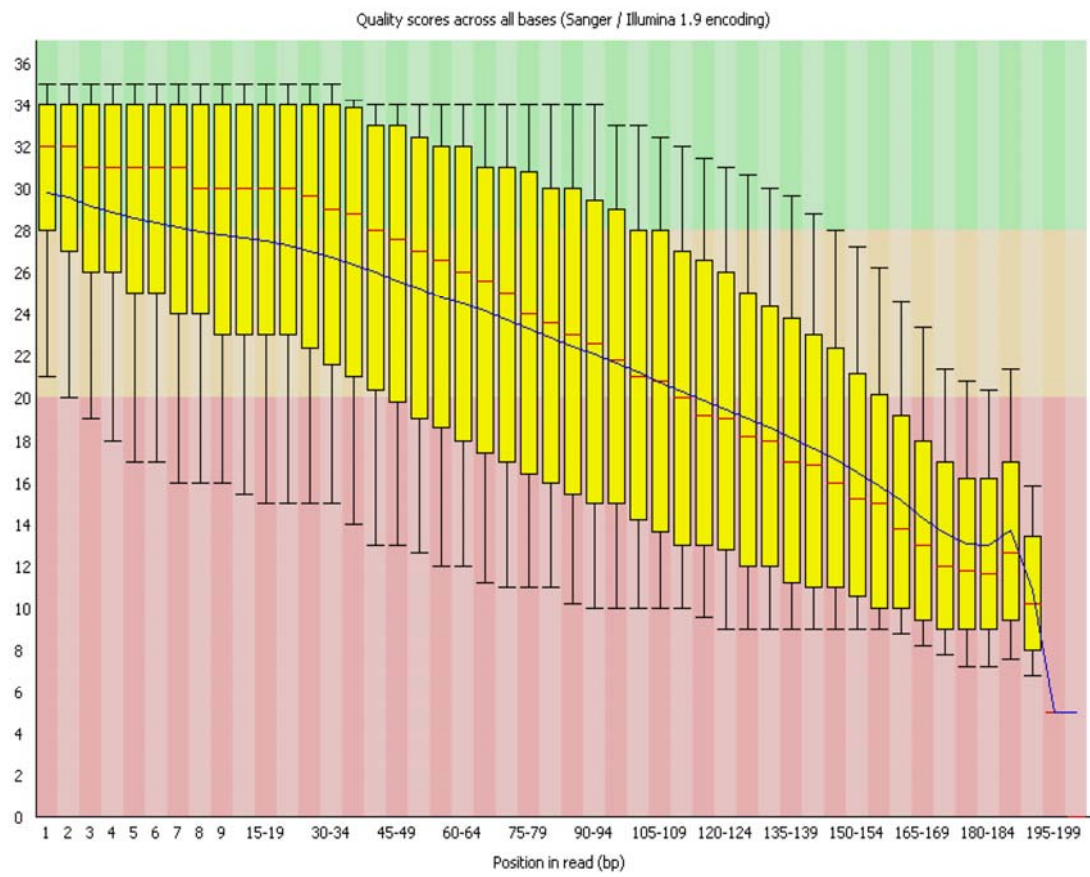
-  [Basic Statistics](#)
-  [Per base sequence quality](#)
-  [Per sequence quality scores](#)
-  [Per base sequence content](#)
-  [Per sequence GC content](#)
-  [Per base N content](#)
-  [Sequence Length Distribution](#)
-  [Sequence Duplication Levels](#)
-  [Overrepresented sequences](#)
-  [Adapter Content](#)
-  [Kmer Content](#)

Basic Statistics

Measure	Value
Filename	R_2012_04_20_13_04_51_user_SMA-35_Auto_SMA-35_37.fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	1695931
Sequences flagged as poor quality	0
Sequence length	5-203
%GC	39

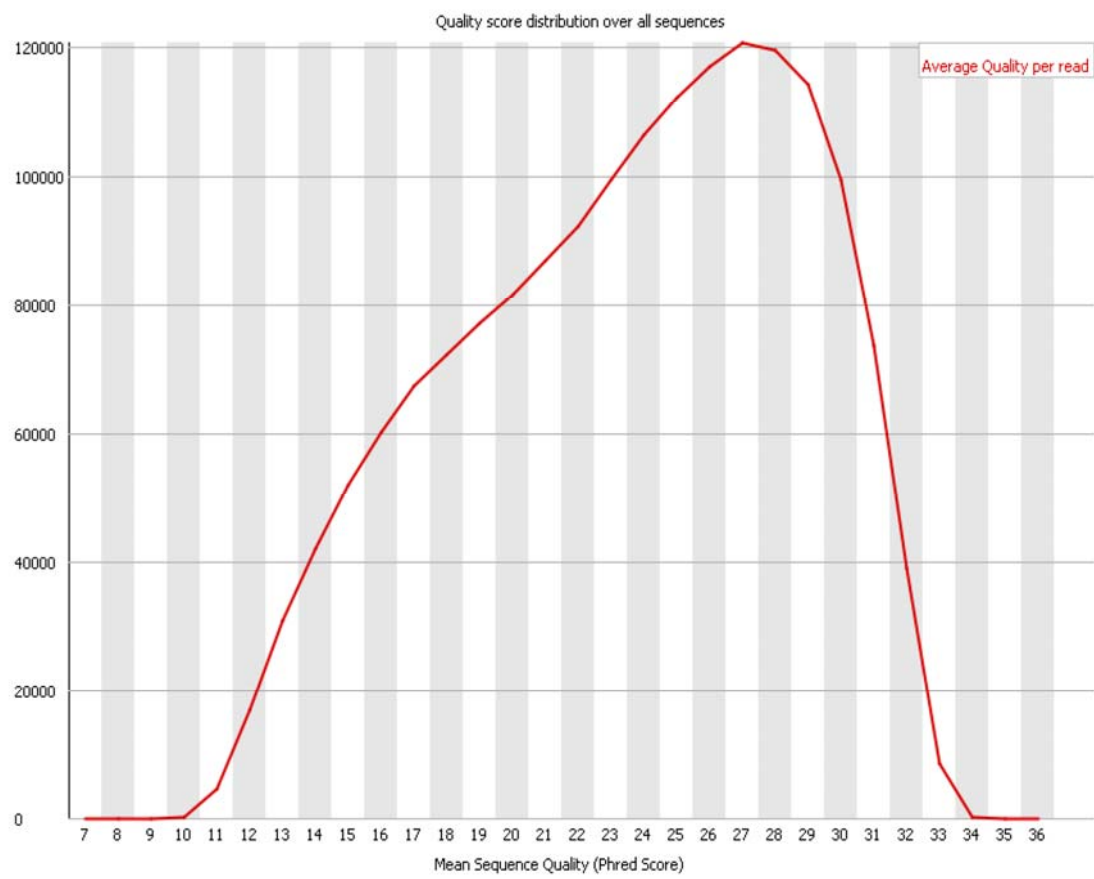
Per base sequence quality

R_2012_04_20_13_04_51_user_SMA-35_Auto_SMA-35_37.fastq Fa... file:///C:/Documents and Settings/14411911/My Documents/Ion Torrent ...



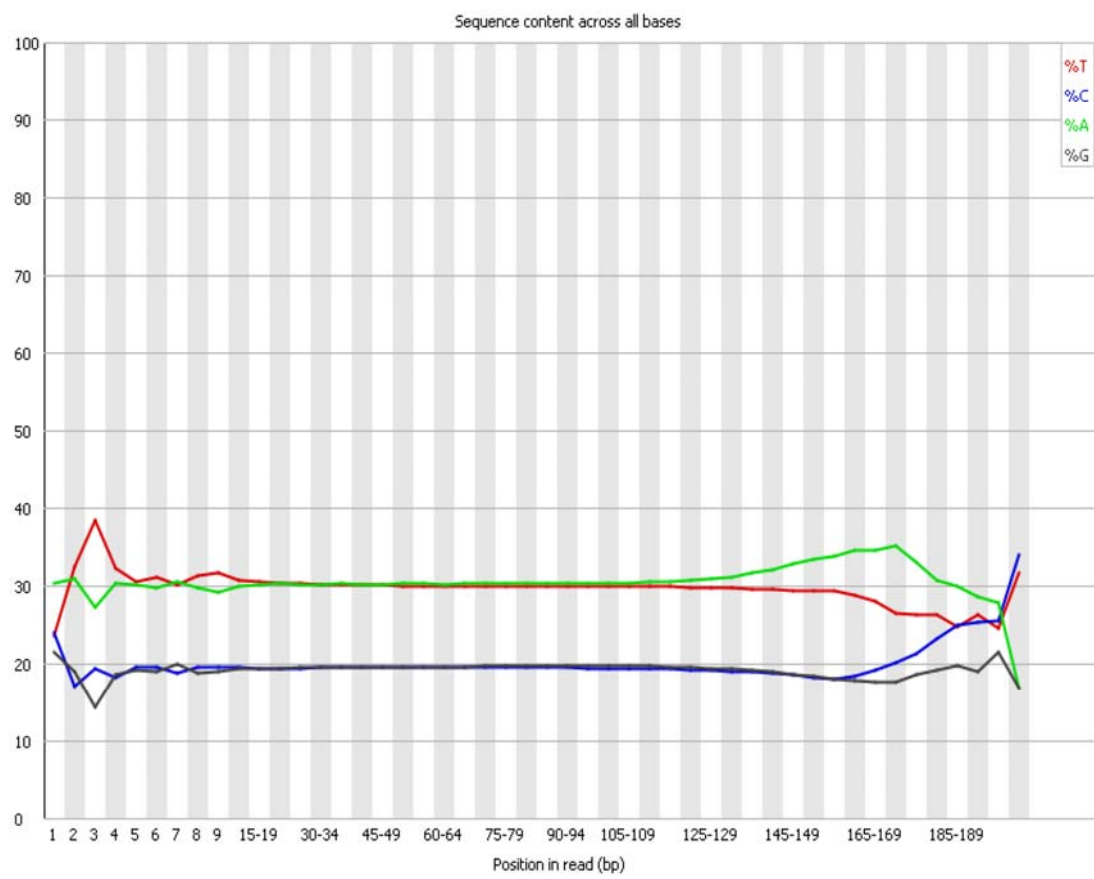
! Per sequence quality scores

R_2012_04_20_13_04_51_user_SMA-35_Auto_SMA-35_37.fastq Fa... file:///C:/Documents and Settings/14411911/My Documents/Ion Torrent ...



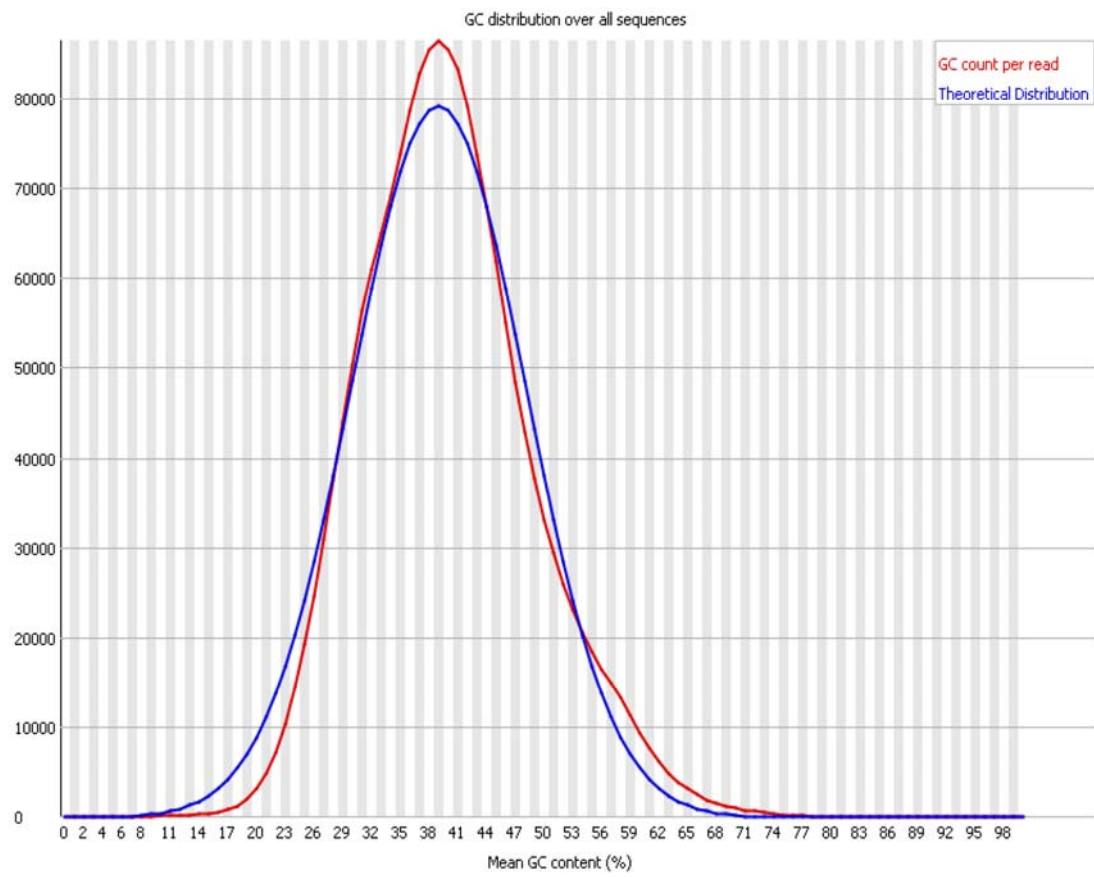
✖ Per base sequence content

R_2012_04_20_13_04_51_user_SMA-35_Auto_SMA-35_37.fastq Fa... file:///C:/Documents and Settings/14411911/My Documents/Ion Torrent ...



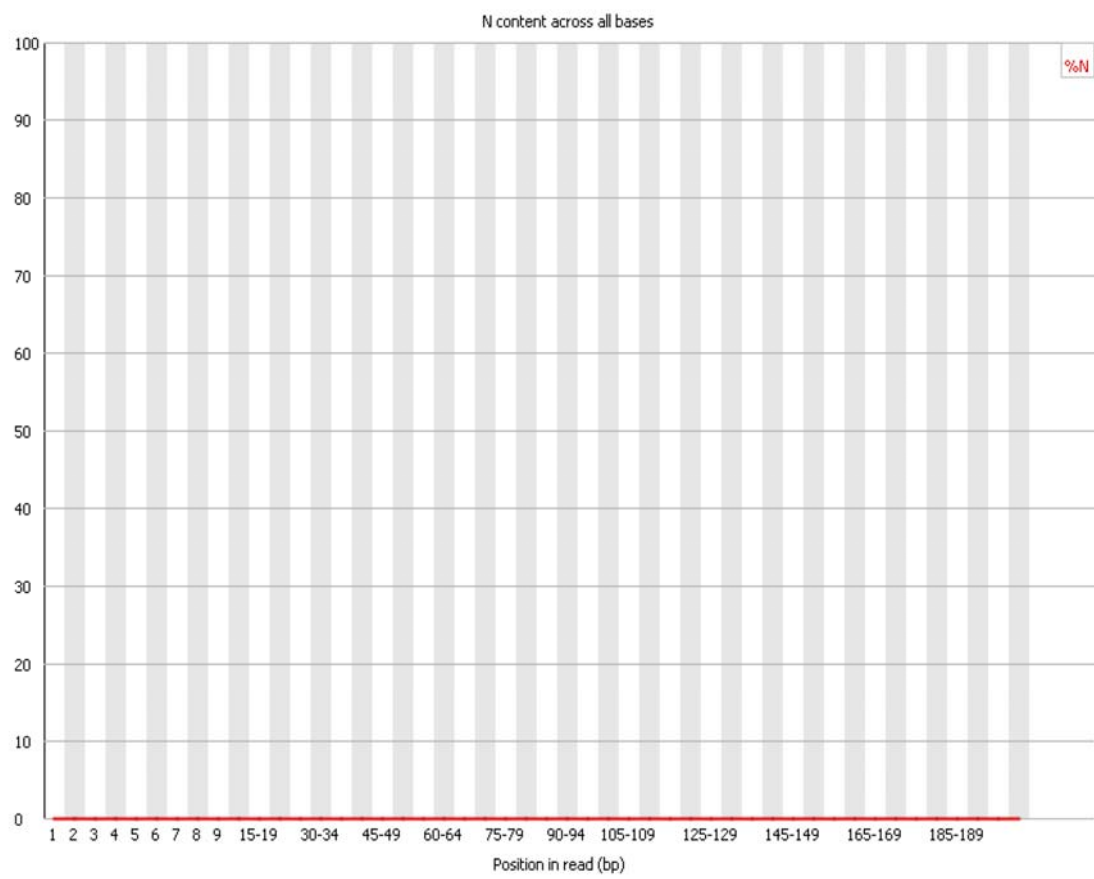
Per sequence GC content

R_2012_04_20_13_04_51_user_SMA-35_Auto_SMA-35_37.fastq Fa... file:///C:/Documents and Settings/14411911/My Documents/Ion Torrent ...



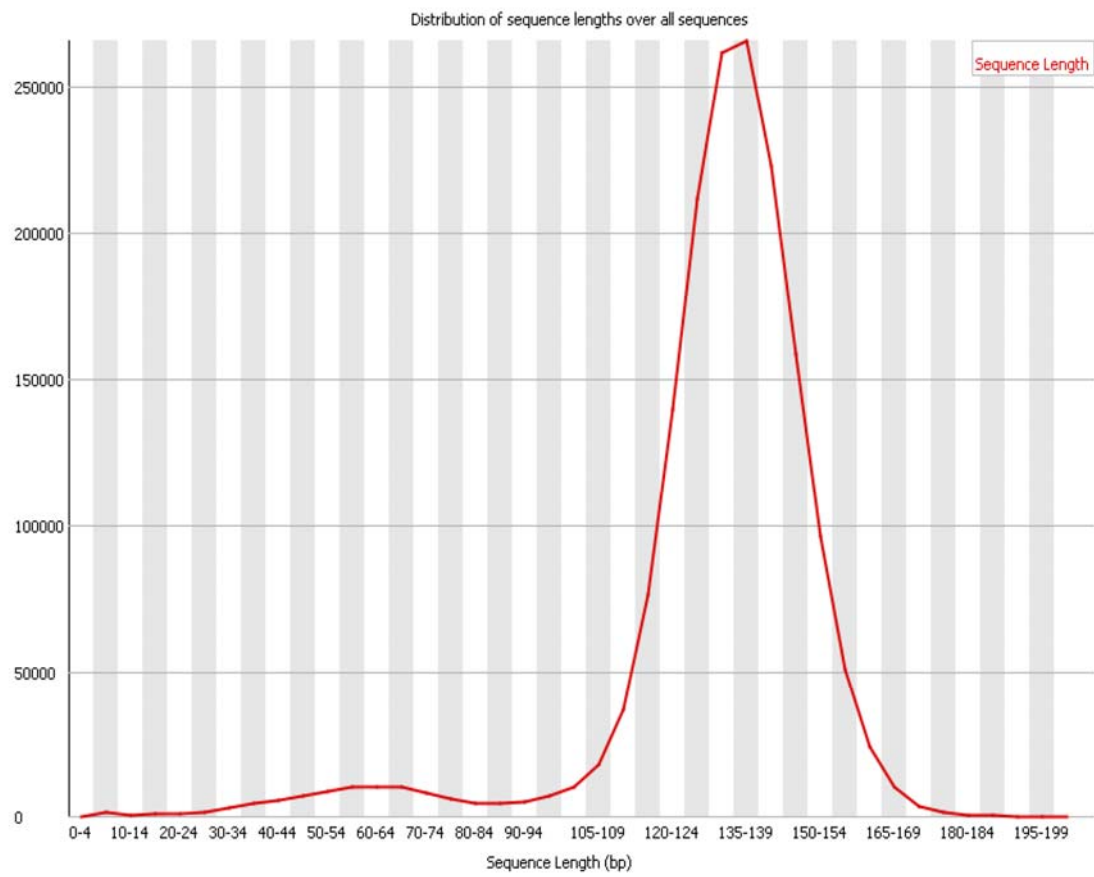
✓ Per base N content

R_2012_04_20_13_04_51_user_SMA-35_Auto_SMA-35_37.fastq Fa... file:///C:/Documents and Settings/14411911/My Documents/Ion Torrent ...



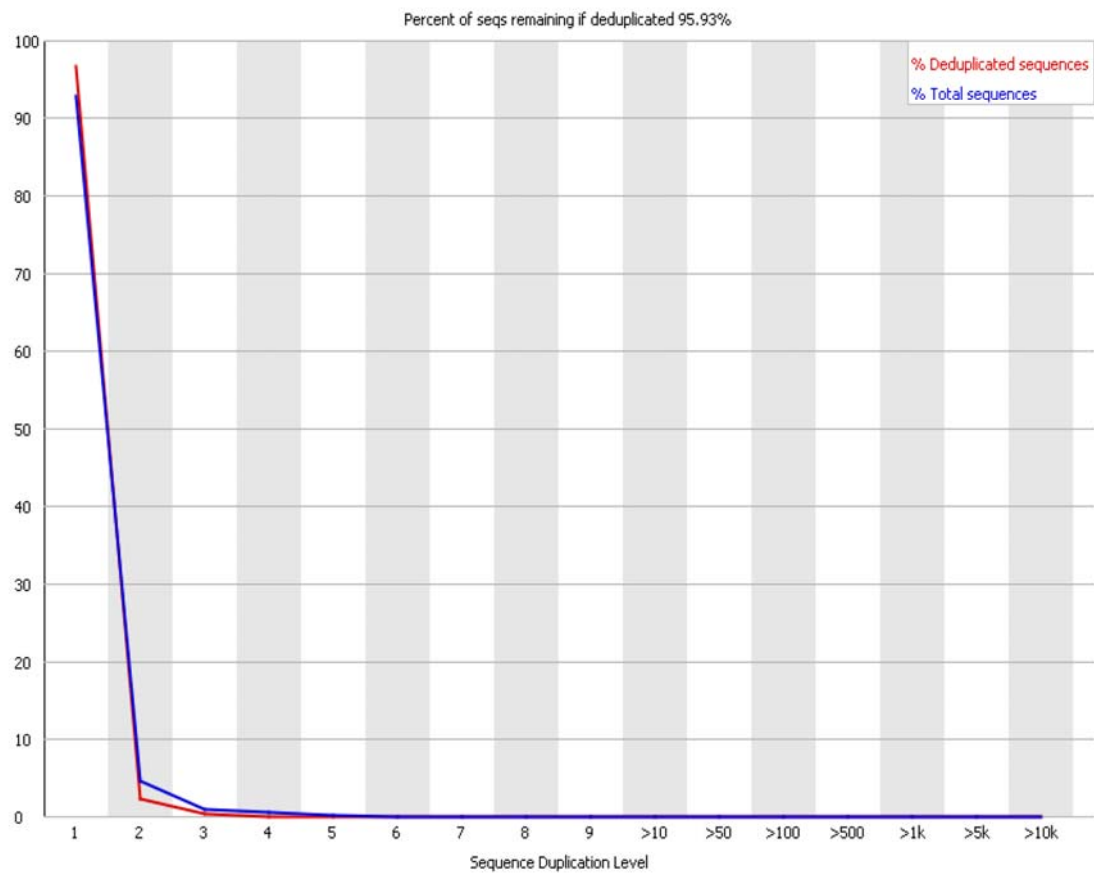
🚨 Sequence Length Distribution

R_2012_04_20_13_04_51_user_SMA-35_Auto_SMA-35_37.fastq Fa... file:///C:/Documents and Settings/14411911/My Documents/Ion Torrent ...



✔ Sequence Duplication Levels

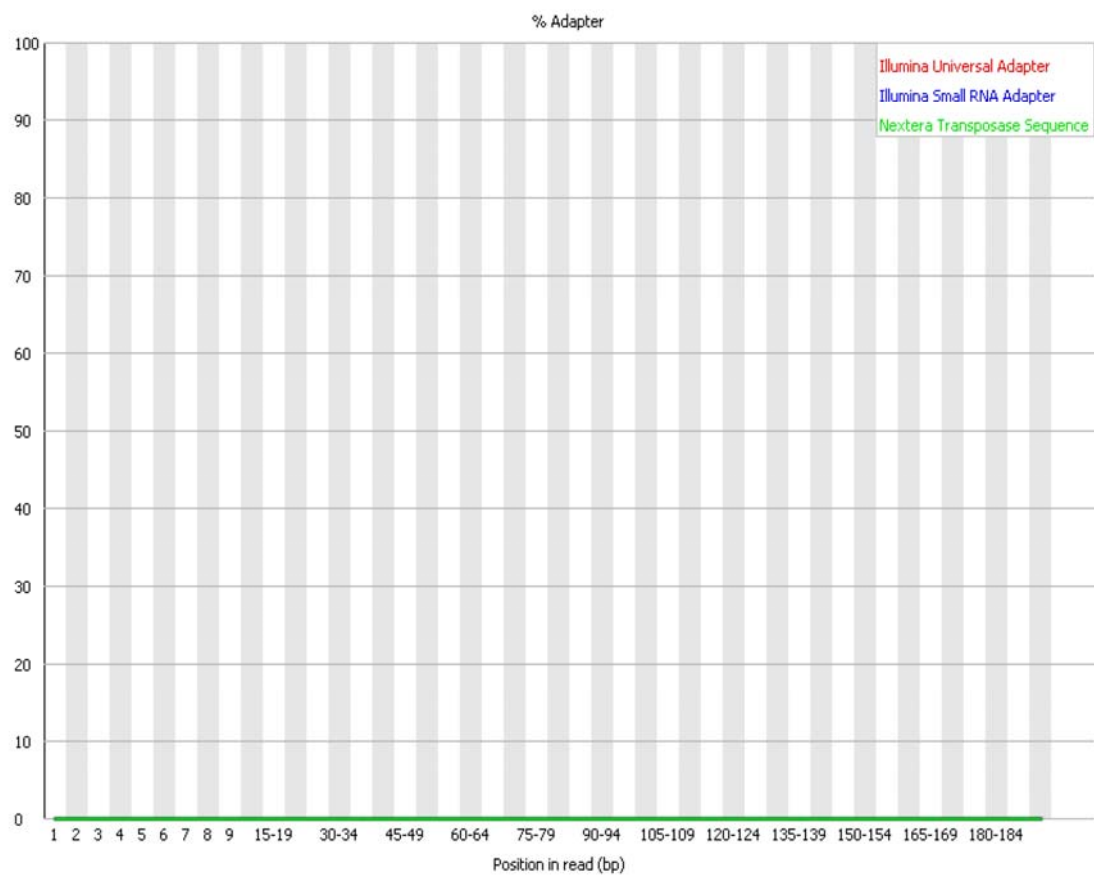
R_2012_04_20_13_04_51_user_SMA-35_Auto_SMA-35_37.fastq Fa... file:///C:/Documents and Settings/14411911/My Documents/Ion Torrent ...



 **Overrepresented sequences**
No overrepresented sequences

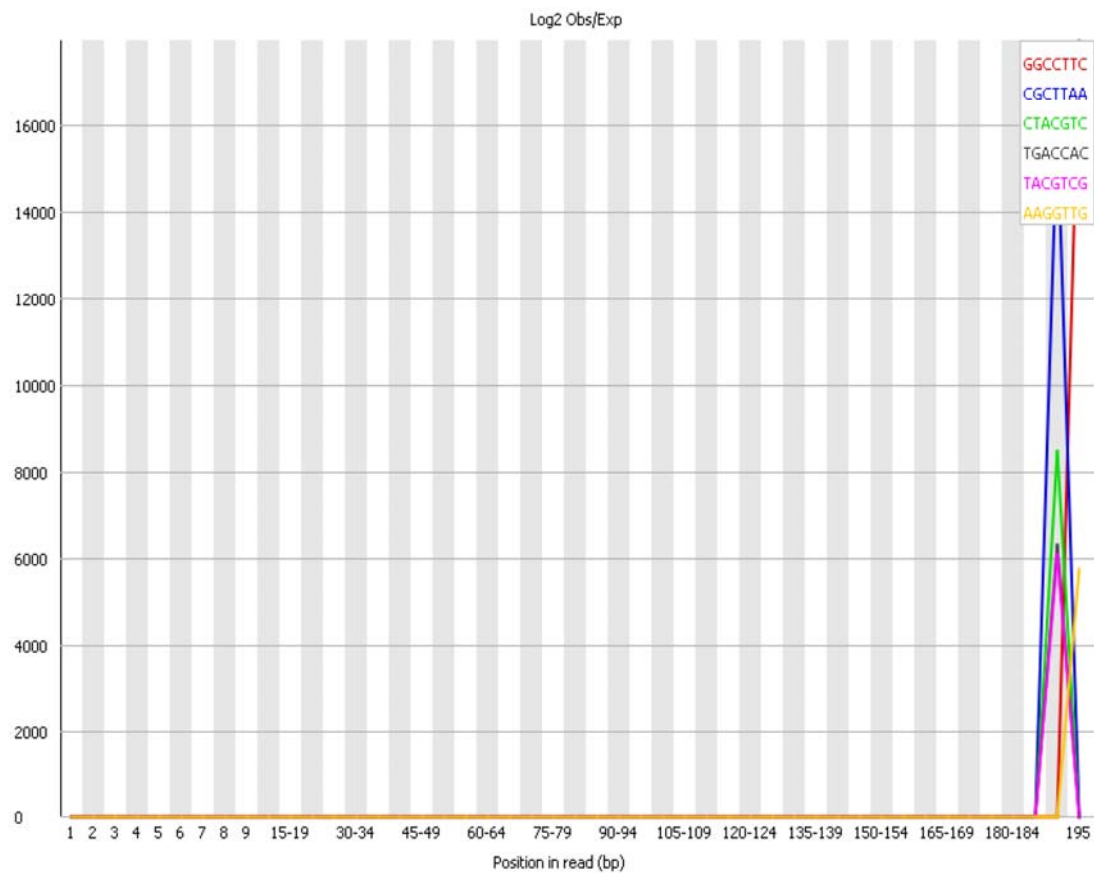
 **Adapter Content**

R_2012_04_20_13_04_51_user_SMA-35_Auto_SMA-35_37.fastq Fa... file:///C:/Documents and Settings/14411911/My Documents/Ion Torrent ...



Kmer Content

R_2012_04_20_13_04_51_user_SMA-35_Auto_SMA-35_37.fastq Fa... file:///C:/Documents and Settings/14411911/My Documents/Ion Torrent ...



Sequence	Count	PValue	Obs/Exp Max	Max Obs/Exp Position
GGCCTTC	590	1.2730328E-5	17937.428	195
CGCTTAA	330	2.3894923E-5	16034.974	190-194
CTACGTC	625	8.570966E-5	8466.466	190-194
TGACCAC	840	1.5481835E-4	6299.454	190-194
TACGTCG	435	1.9374763E-4	6082.2314	190-194
AAGGTTG	1840	1.2381442E-4	5751.6753	195
GACCACT	825	6.9684116E-4	3206.9946	190-194
CGGGGCT	185	7.883764E-4	3178.1025	175-179
TTCCCGG	725	8.6485804E-4	2919.471	185-189
CTCGCCA	780	0.0010010411	2713.6108	185-189
GATCCTG	785	0.0010139148	2696.327	185-189
ACGTCGG	780	0.0012234647	2466.919	190-194
TCCCGGA	575	0.0014505823	2300.6702	185-189
CGTCGGT	660	0.0014492939	2290.7107	190-194
GCTTAAG	1155	0.0013657236	2290.7104	190-194

R_2012_04_20_13_04_51_user_SMA-35_Auto_SMA-35_37.fastq Fa... file:///C:/Documents and Settings/14411911/My Documents/Ion Torrent ...

Sequence	Count	PValue	Obs/Exp Max	Max Obs/Exp Position
CCTACGT	305	0.0015814928	2238.6213	185-189
CTCTTCG	1010	0.0016783414	2095.66	185-189
GCCTACG	415	0.002046345	1961.6466	185-189
TGTTTGG	2725	0.0016290909	1941.85	190-194
CTGCGTT	575	0.0020669675	1937.4065	180-184

Produced by [FastQC](#) (version 0.11.2)

6-6: The Ion Torrent run report and FastQC report of *Fagonia rangei*.

Auto_SMA-68_104, CAF - Stellenbosch - Torrent Browser

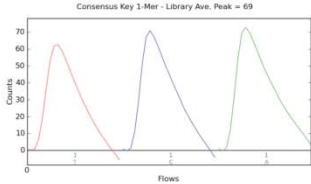
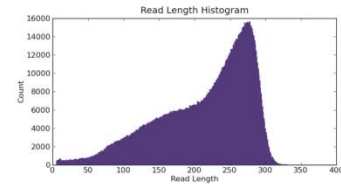
http://146.232.41.138/output/Home/Auto_SMA-68_104_234...

Report for Auto_SMA-68_104

Library Summary

Based on Predicted Per-Base Quality Scores - Independent of Alignment

Total Number of Bases [Mbp]	363.58
• Number of Q20 Bases [Mbp]	272.26
Total Number of Reads	1,707,924
Mean Length [bp]	213
Longest Read [bp]	392



Reference Genome Information

Based on Full Library Alignment to Provided Reference

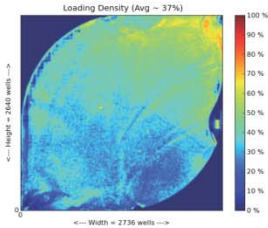
Read Alignment Distribution

./alignTable.txt not found

Test Fragment Report

Ion Sphere™ Particle (ISP) Identification Summary

Total Addressable Wells	Count	Percentage
• Wells with ISPs	6,348,218	37%
• Live ISPs	2,328,188	95%
• Test Fragment ISPs	2,214,169	10%
• Library ISPs	214,261	90%
Library ISPs / Percent Enrichment	Count	Percentage
• Filtered: Polyclonal	1,999,908	95%
• Filtered: Primer dimer	205,371	10%
• Filtered: Low quality	325	< 1%
• Final Library Reads	86,288	4%
	1,707,924	85%



Report Information

Analysis Info

Run Name	R_2012_08_31_11_29_40_user_SMA-68
Run Date	2012-08-31 11:29:40
Analysis Name	Auto_SMA-68_104
Analysis Date	2012-10-05
Analysis Cycles	16
Analysis Flows	520
Project	chl-fagonia.rangei
Sample	fagonia.rangei-chl
Library	none
PGM	SmartBlue
Chip Check	Passed
Chip Type	316D
Chip Data	single
Notes	
Barcode Set	
runID	NS7E1
Flow Order	TACGTACGTCGTGACATCGATCGATGTACAGC
Library Key	TCAG

Software Version

Torrent_Suite	2.2
Datacollect	210

Auto_SMA-68_104, CAF - Stellenbosch - Torrent Browser

http://146.232.41.138/output/Home/Auto_SMA-68_104_234...

Graphics	18
LiveView	345
OS	19
Script	18.1.6
host	stellenboschpgm1
ion-alignment	2.2.4-1
ion-analysis	2.2.12-1
ion-dbreports	2.2.17-1
ion-gpu	1.2-1
ion-pipeline	2.2.12-1
ion-plugins	2.2.16-1
ion-torrentR	2.2.8-1
ion-tsups	1.0-1

File Links

[Library Sequence \(SFF\)](#)
[Library Sequence \(FASTQ\)](#)
[Full Library Alignments \(BAM\)](#)
[Full Library Alignments \(BAI\)](#)
[Test Fragments \(SFF\)](#)
[PDF of this Report](#)
[Customer Support Archive](#)

Plugin Summary

Select Plugins To Run	Refresh Plugin Status
-----------------------	-----------------------

[Request Support](#) | [Help/Documentation](#) | [Terms of Use](#)
 Copyright © 2012 Life Technologies Corporation
 This product should be used for research use only












R_2012_08_31_11_29_40_user_SMA-68_Auto_SMA-68_104.fastq F... file:///C:/Documents and Settings/14411911/My Documents/Ion Torrent ...

FastQC Report

Summary

Fri 20 Feb 2015

R_2012_08_31_11_29_40_user_SMA-68_Auto_SMA-68_104.fastq

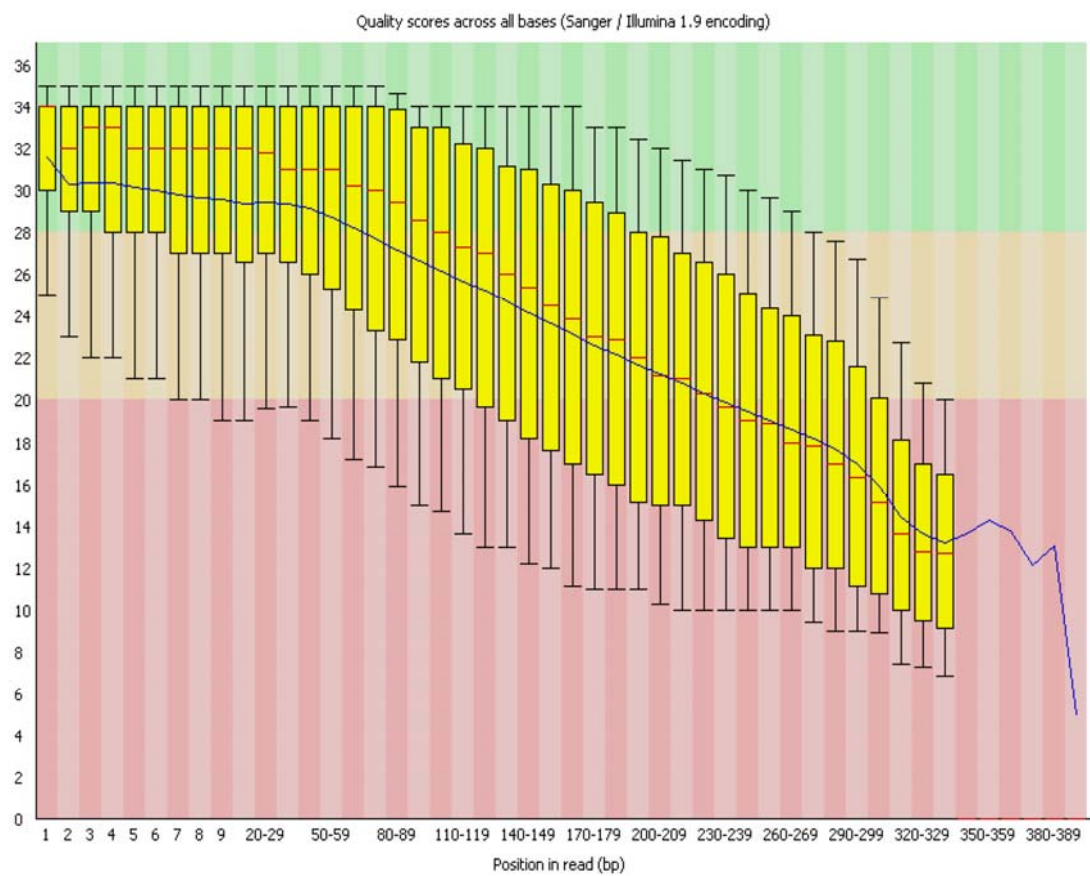
-  [Basic Statistics](#)
-  [Per base sequence quality](#)
-  [Per sequence quality scores](#)
-  [Per base sequence content](#)
-  [Per sequence GC content](#)
-  [Per base N content](#)
-  [Sequence Length Distribution](#)
-  [Sequence Duplication Levels](#)
-  [Overrepresented sequences](#)
-  [Adapter Content](#)
-  [Kmer Content](#)

Basic Statistics

Measure	Value
Filename	R_2012_08_31_11_29_40_user_SMA-68_Auto_SMA-68_104.fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	1707924
Sequences flagged as poor quality	0
Sequence length	5-392
%GC	35

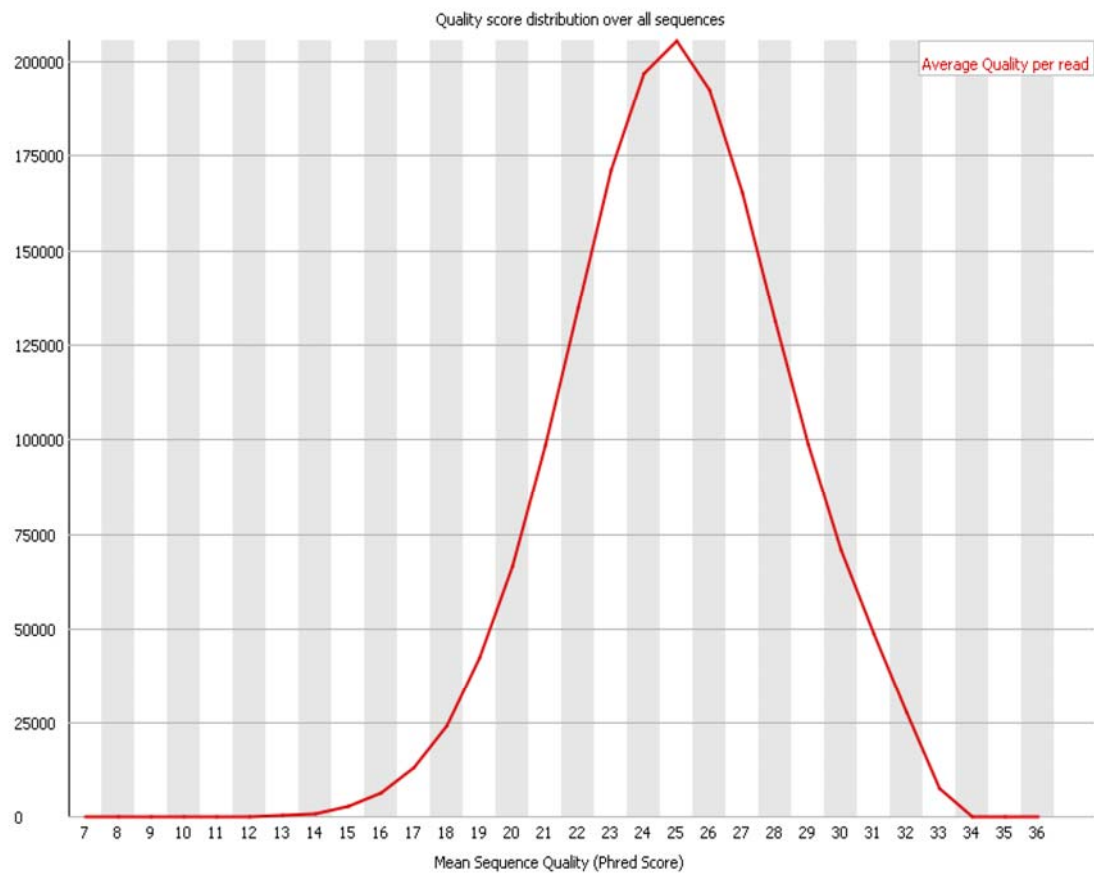
Per base sequence quality

R_2012_08_31_11_29_40_user_SMA-68_Auto_SMA-68_104.fastq F... file:///C:/Documents and Settings/14411911/My Documents/Ion Torrent ...



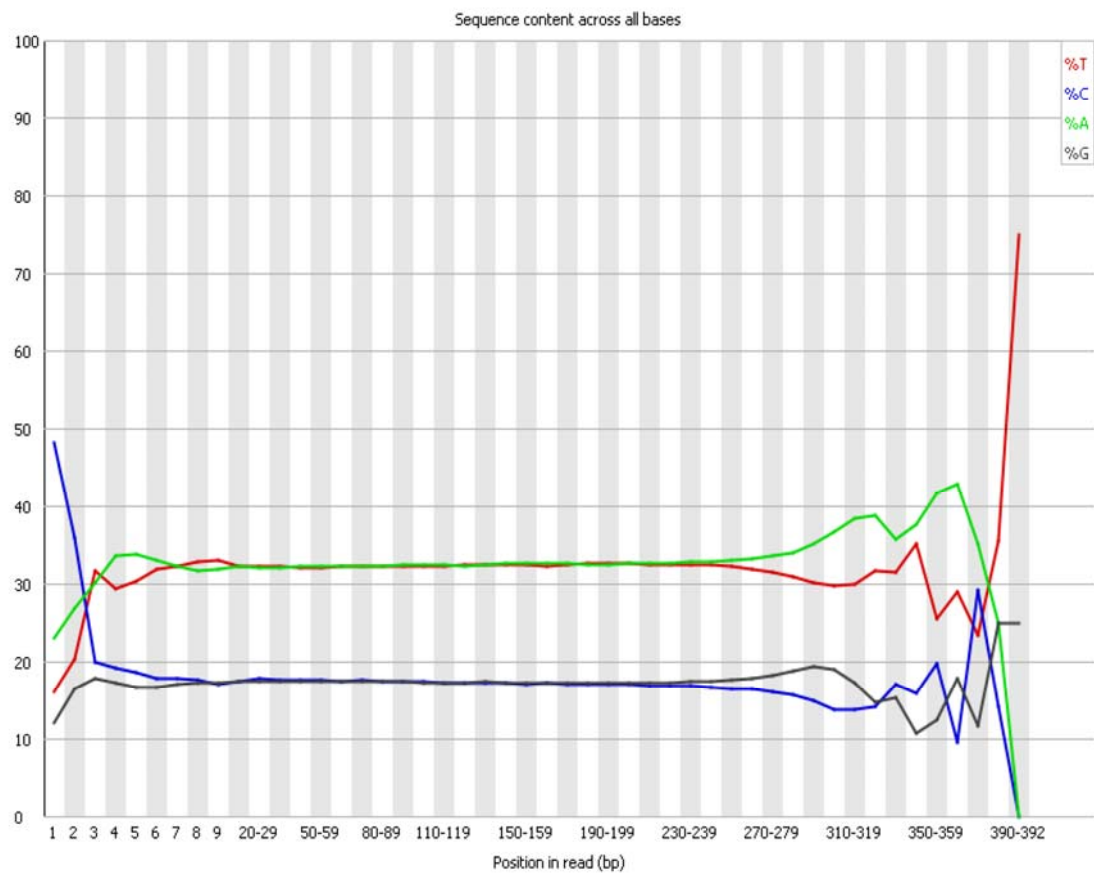
! Per sequence quality scores

R_2012_08_31_11_29_40_user_SMA-68_Auto_SMA-68_104.fastq F... file:///C:/Documents and Settings/14411911/My Documents/Ion Torrent ...



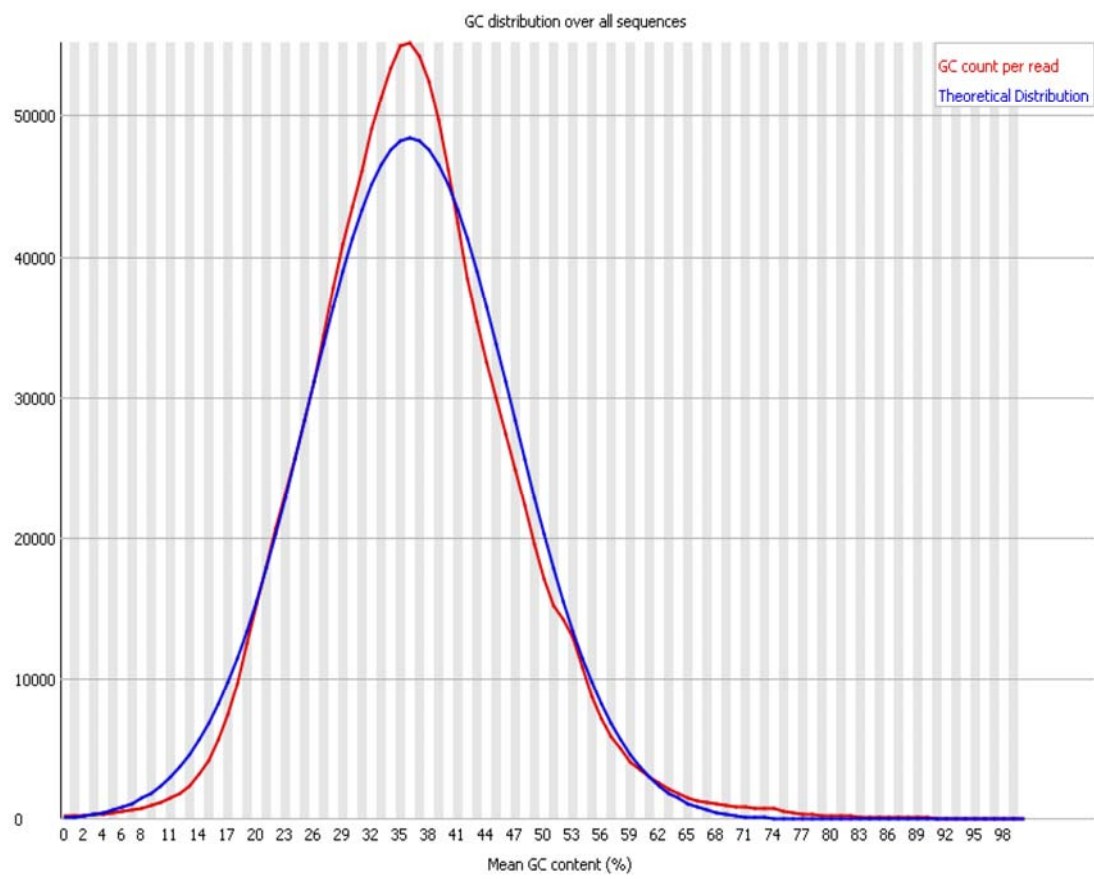
✖ Per base sequence content

R_2012_08_31_11_29_40_user_SMA-68_Auto_SMA-68_104.fastq F... file:///C:/Documents and Settings/14411911/My Documents/Ion Torrent ...



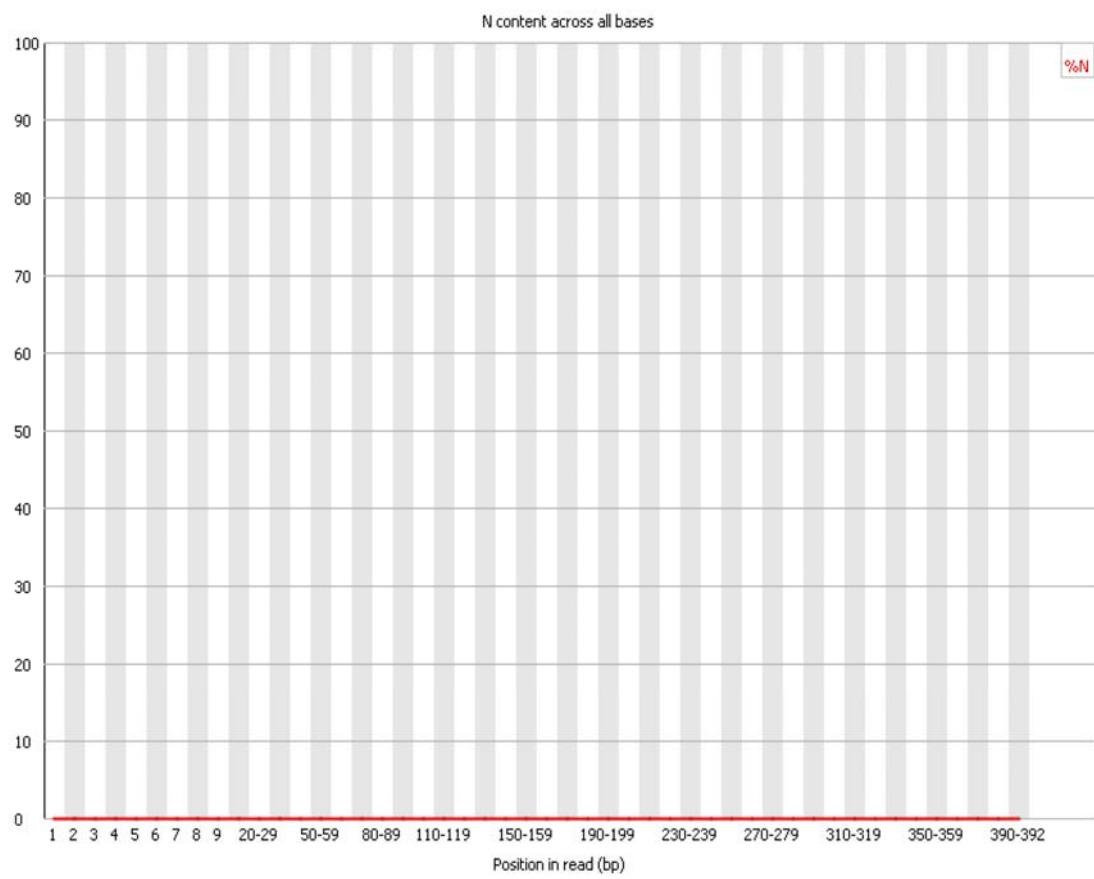
Per sequence GC content

R_2012_08_31_11_29_40_user_SMA-68_Auto_SMA-68_104.fastq F... file:///C:/Documents and Settings/14411911/My Documents/Ion Torrent ...



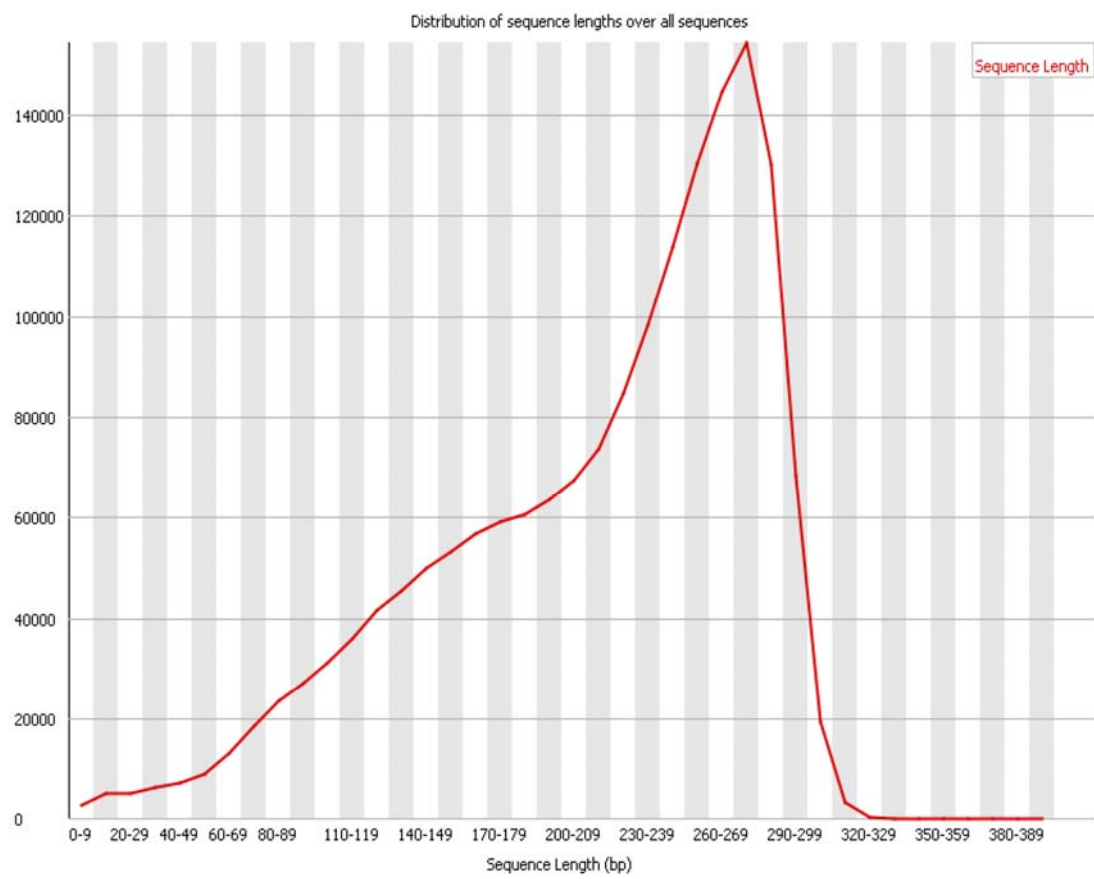
Per base N content

R_2012_08_31_11_29_40_user_SMA-68_Auto_SMA-68_104.fastq F... file:///C:/Documents and Settings/14411911/My Documents/Ion Torrent ...



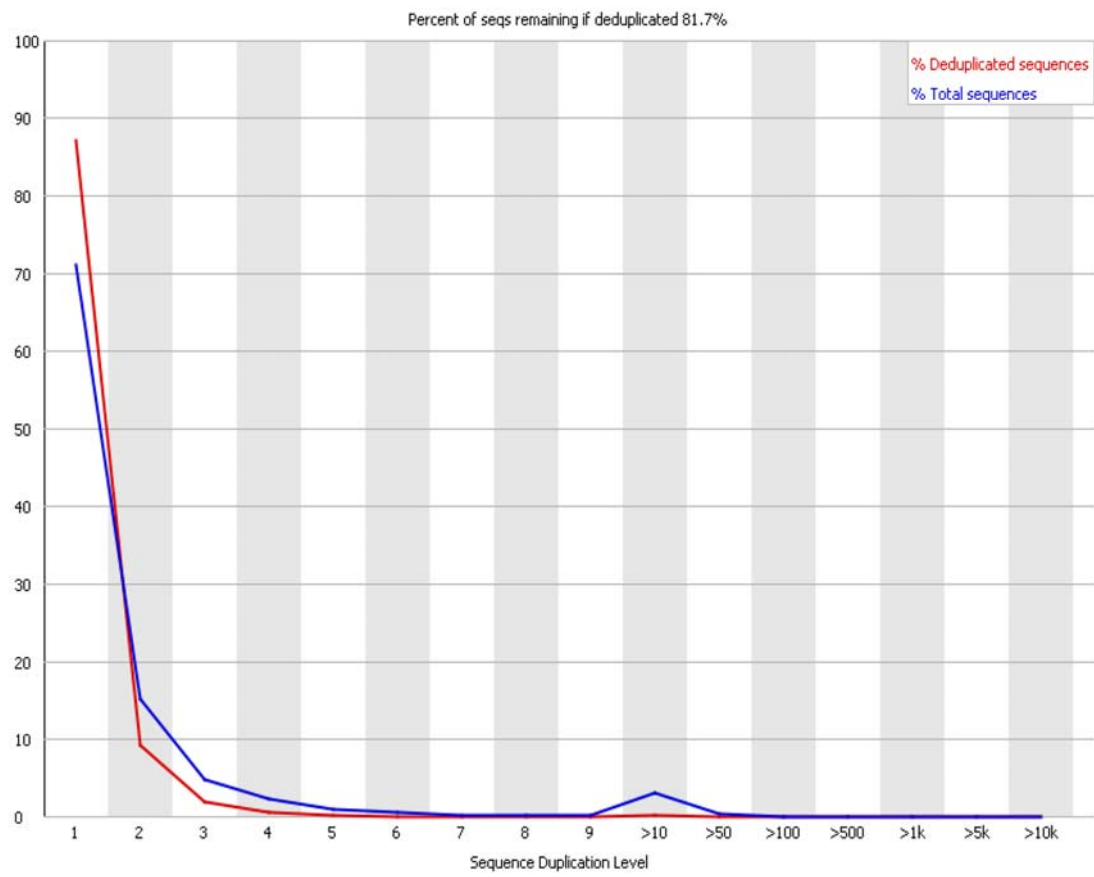
🚫 Sequence Length Distribution

R_2012_08_31_11_29_40_user_SMA-68_Auto_SMA-68_104.fastq F... file:///C:/Documents and Settings/14411911/My Documents/Ion Torrent ...



✔ Sequence Duplication Levels

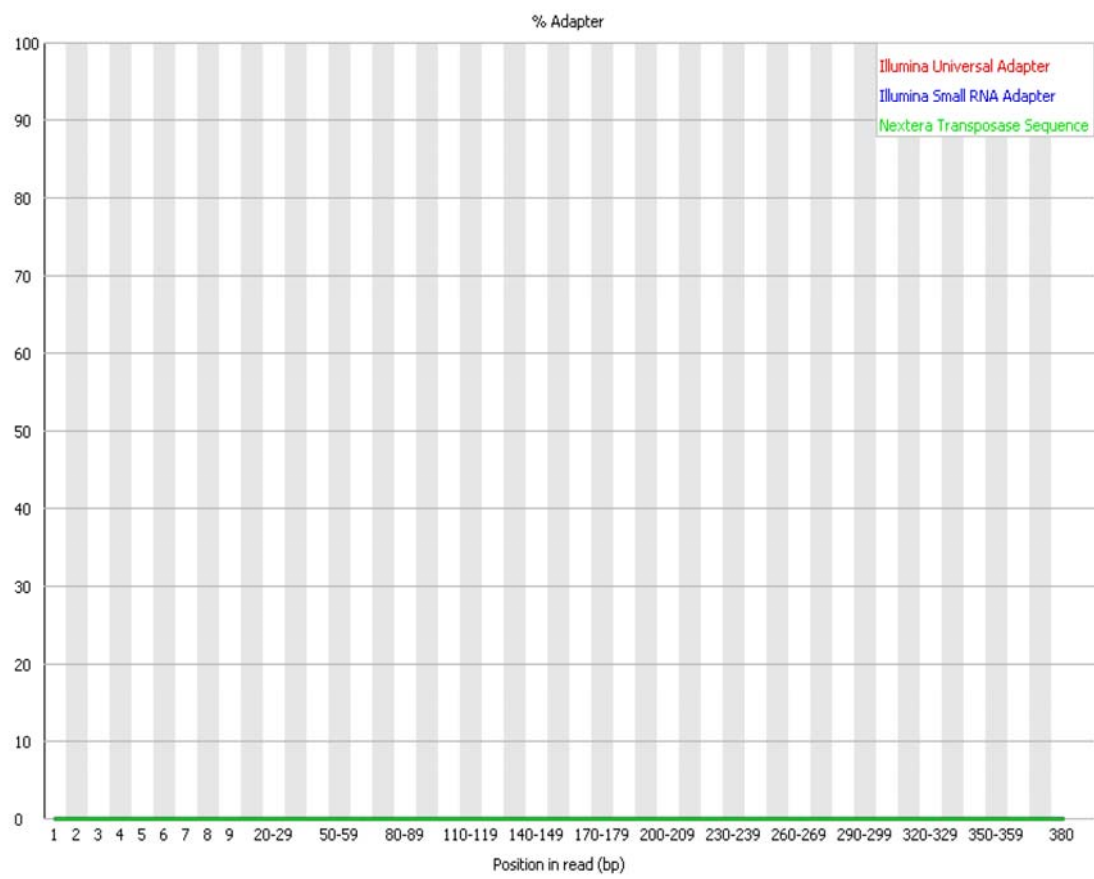
R_2012_08_31_11_29_40_user_SMA-68_Auto_SMA-68_104.fastq F... file:///C:/Documents and Settings/14411911/My Documents/Ion Torrent ...



 **Overrepresented sequences**
No overrepresented sequences

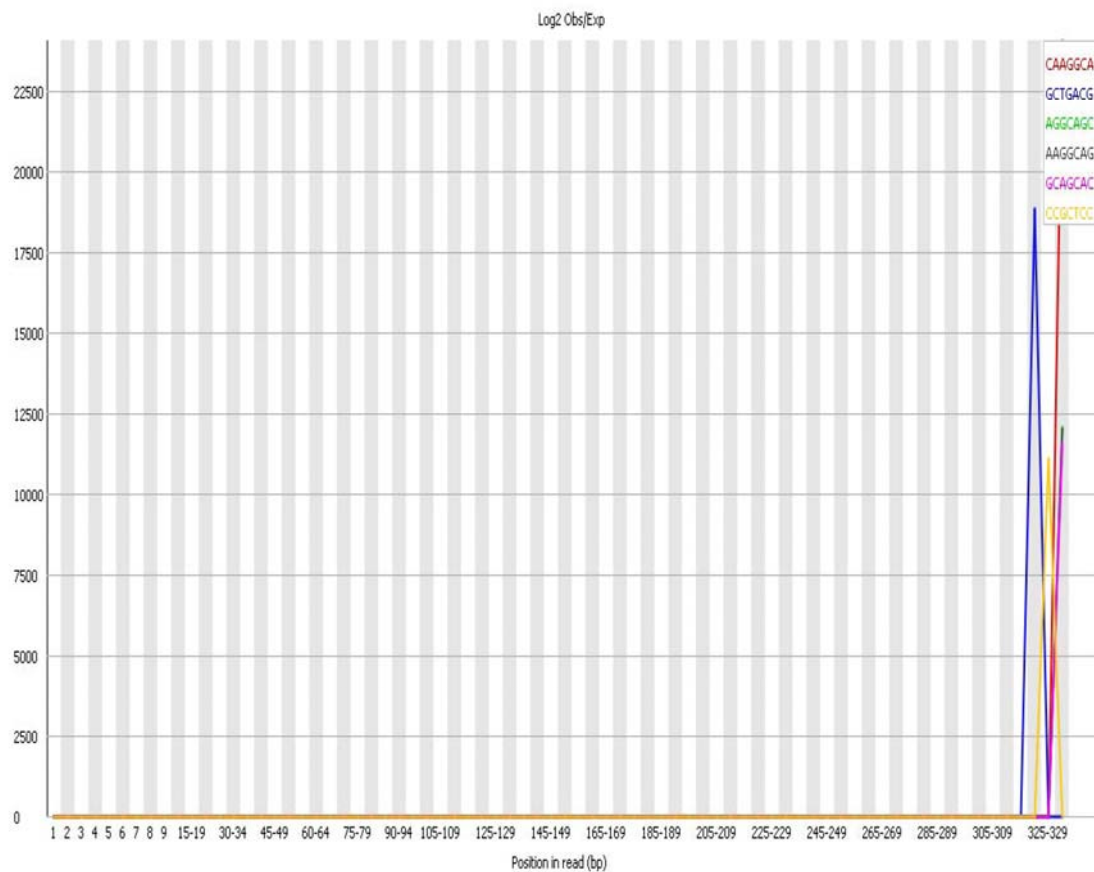
 **Adapter Content**

R_2012_08_31_11_29_40_user_SMA-68_Auto_SMA-68_104.fastq F... file:///C:/Documents and Settings/14411911/My Documents/Ion Torrent ...



 **Kmer Content**

R_2012_08_31_11_29_40_user_SMA-68_Auto_SMA-68_104.fastq F... file:///C:/Documents and Settings/14411911/My Documents/Ion Torrent ...



Sequence	Count	PValue	Obs/Exp Max	Max Obs/Exp Position
CAAGGCA	1470	0.0	24029.332	330-334
GCTGACG	375	1.8465247E-5	18838.998	320-324
AGGCAGC	975	3.744763E-5	12076.28	330-334
AAGGCAG	1465	2.8182343E-5	12055.672	330-334
GCAGCAC	610	4.885922E-5	11581.351	330-334
CCGCTCC	635	5.294606E-5	11125.392	325-329
GGCAGCA	985	7.643801E-5	8965.259	330-334
CGTCCC	445	9.3604955E-5	8819.755	325-329
CGATGCA	820	8.828956E-5	8615.3955	325-329
AACCGCG	195	1.497806E-4	7245.7686	310-314
GCTCCCG	395	1.5979199E-4	6878.894	325-329
GACCGCT	260	1.6864124E-4	6792.907	320-324
TTACGCG	230	2.0837011E-4	6143.1514	310-314
AGACCGC	405	2.2612992E-4	5814.5054	320-324
CTGACGA	670	2.652284E-4	5272.1074	320-324

R_2012_08_31_11_29_40_user_SMA-68_Auto_SMA-68_104.fastq F... file:///C:/Documents and Settings/14411911/My Documents/Ion Torrent ...

Sequence	Count	PValue	Obs/Exp Max	Max Obs/Exp Position
TACGCGG	165	3.8639773E-4	4554.8833	310-314
ATGCCCA	1665	3.639899E-4	4243.0176	320-324
CACTAAG	1695	3.7722423E-4	4167.92	325-329
GGAGGCC	735	5.532289E-4	3696.8203	315-319
AAGACCG	1005	5.967337E-4	3514.7385	320-324

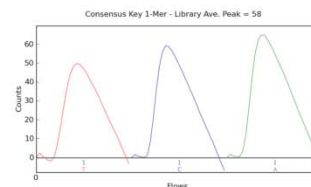
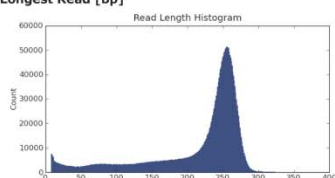
Produced by [FastQC](#) (version 0.11.2)

6-7: The Ion Torrent run report and FastQC report of *Zygophyllum stapffii*.

Zygo_sma85, CAF - Stellenbosch - Torrent Browser

http://146.232.41.138/output/Home/Zygo_sma85_279/**Report for Zygo_sma85****Library Summary****Based on Predicted Per-Base Quality Scores - Independent of Alignment**

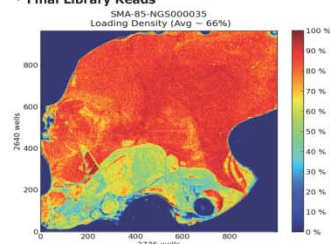
Total Number of Bases [Mbp]	578.03
• Number of Q20 Bases [Mbp]	494.42
Total Number of Reads	2,716,726
Mean Length [bp]	212
Longest Read [bp]	384

**Reference Genome Information**

Based on Full Library Alignment to Provided Reference

Test Fragment Report**Ion Sphere™ Particle (ISP) Identification Summary**

Total Addressable Wells	Count	Percentage
• Wells with ISPs	6,348,241	66%
• Live ISPs	4,204,443	94%
• Test Fragment ISPs	3,938,105	1%
• Library ISPs	44,479	99%
	3,893,626	
Library ISPs / Percent Enrichment	Count	Percentage
• Filtered: Polyclonal	3,893,626	94%
• Filtered: Primer dimer	883,547	23%
• Filtered: Low quality	9	<1%
• Final Library Reads	293,344	8%
	2,716,726	70%

**Report Information****Analysis Info**

Run Name	R_2013_02_14_10_51_11_user_SMA-85-NGS000035
Run Date	2013-02-14 08:51:11+00:00
Run Cycles	15
Run Flows	500
Project	NGS000035
Sample	Zygophyllum_stapffii
Library	none
PGM	SmartBlue
Flow Order	TACGTACGTCTGAGCATCGATCGATGTACAGC
Library Key	TCAG
TF Key	ATCG
Chip Check	Passed
Chip Type	316D
Chip Data	single
Notes	Prof Bellstedt
Barcode Set	
Analysis Name	Zygo_sma85
Analysis Date	2013-02-16
Analysis Flows	0
runID	IS18B

Software Version

Torrent_Suite	3.2.1
Datacollect	210
Graphics	31
LiveView	395
OS	19
Script	20.1.4
host	stellenboschpgm1
ion-alignment	3.2.1-1
ion-analysis	3.2.5-1
ion-dbreports	3.2.15-1
ion-gpu	3.0.0-1

Zygo_sma85, CAF - Stellenbosch - Torrent Browser

http://146.232.41.138/output/Home/Zygo_sma85_279/

ion-pipeline 3.2.10-1
ion-plugins 3.2.8-1
ion-tsups 1.0-1

File Links

[Library Sequence \(SFF\)](#)
[Library Sequence \(FASTQ\)](#)
[Full Library Alignments \(BAM\)](#)
[Full Library Alignments \(BAI\)](#)
[Test Fragments \(SFF\)](#)
[PDF of this Report](#)

Plugin Summary

Select Plugins To Run	Refresh Plugin Status	
Assembler — v2.2.2		Completed
• Assembler.html		
FastQC — v1.0.1		Completed
• fastqc_report.html		

[Request Support](#) | [Help/Documentation](#) | [Terms of Use](#)
Copyright © 2012 Life Technologies Corporation
This product should be used for research use only












R_2013_02_14_10_51_11_user_SMA-85-NGS000035_Zygo_sma85.fa... file:///C:/Documents and Settings/14411911/My Documents/Ion Torrent ...

FastQC Report

Summary

Fri 20 Feb 2015

R_2013_02_14_10_51_11_user_SMA-85-NGS000035_Zygo_sma85.fastq

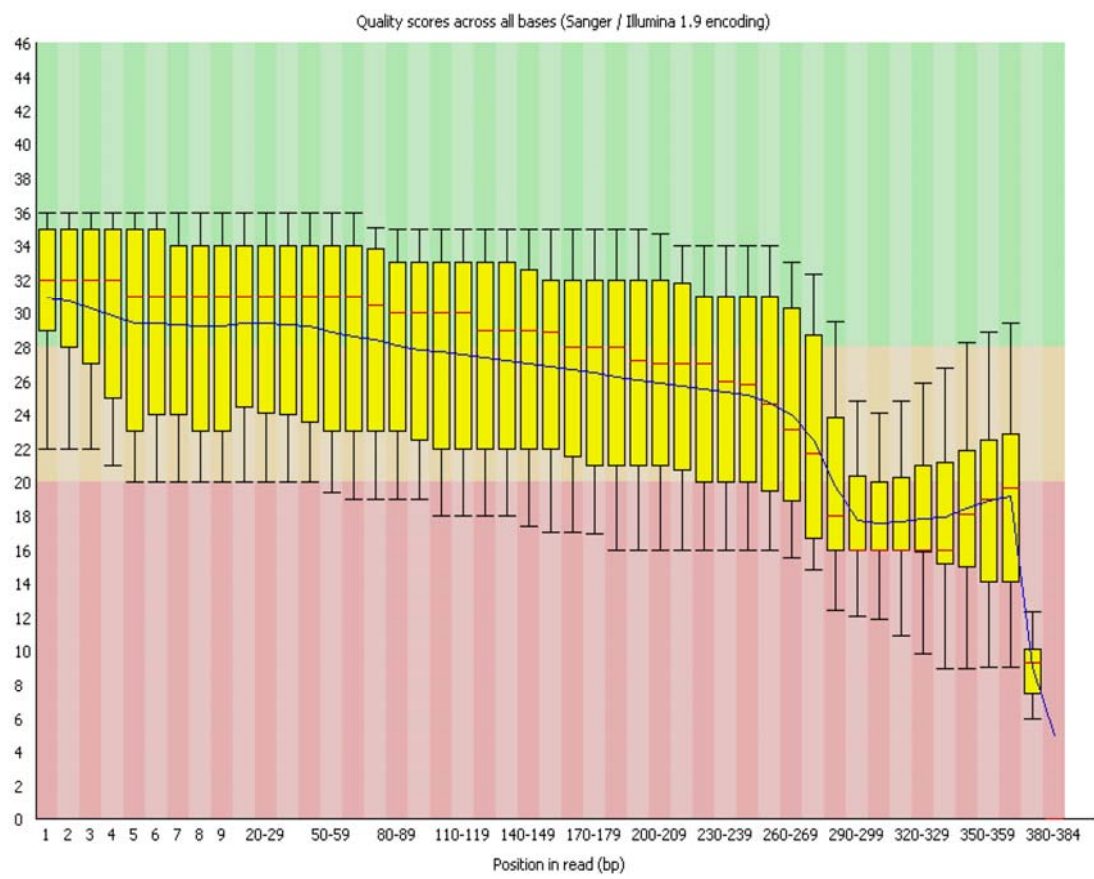
-  [Basic Statistics](#)
-  [Per base sequence quality](#)
-  [Per sequence quality scores](#)
-  [Per base sequence content](#)
-  [Per sequence GC content](#)
-  [Per base N content](#)
-  [Sequence Length Distribution](#)
-  [Sequence Duplication Levels](#)
-  [Overrepresented sequences](#)
-  [Adapter Content](#)
-  [Kmer Content](#)

Basic Statistics

Measure	Value
Filename	R_2013_02_14_10_51_11_user_SMA-85-NGS000035_Zygo_sma85.fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	2716726
Sequences flagged as poor quality	0
Sequence length	8-384
%GC	34

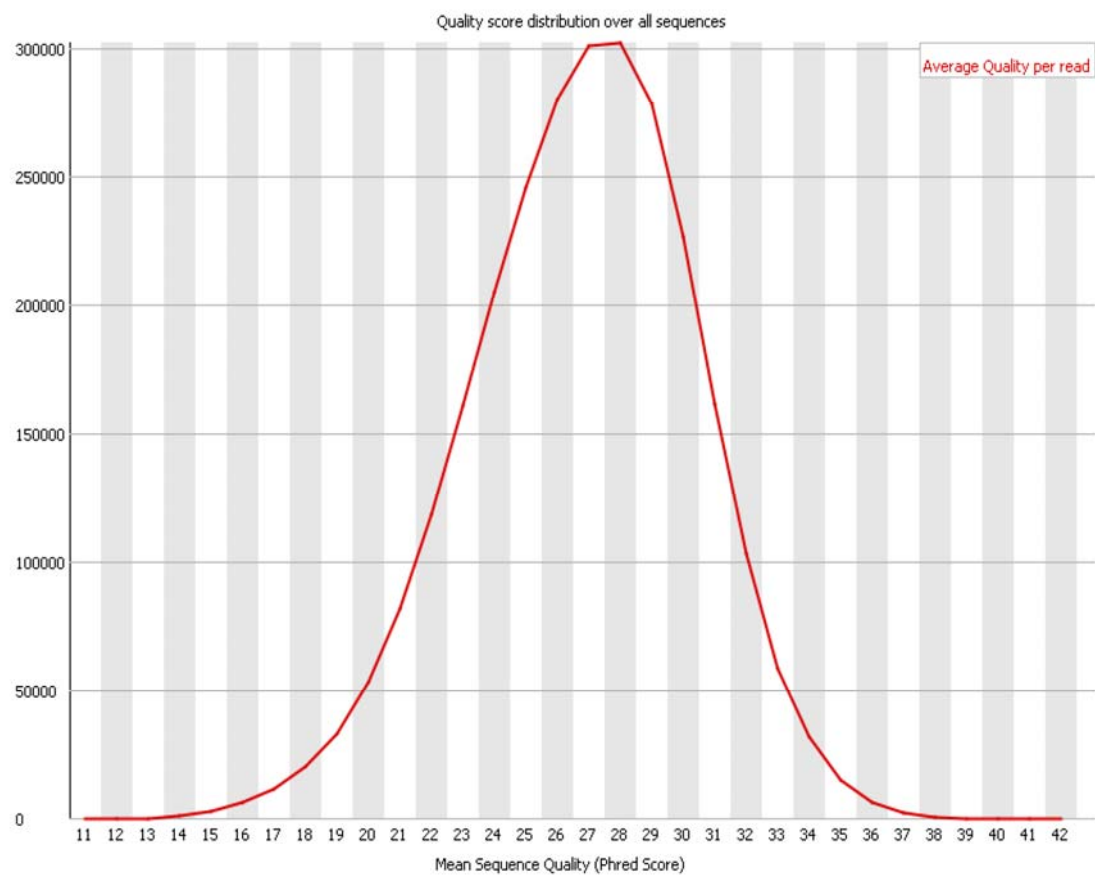
Per base sequence quality

R_2013_02_14_10_51_11_user_SMA-85-NGS000035_Zygo_sma85.fa... file:///C:/Documents and Settings/14411911/My Documents/Ion Torrent ...



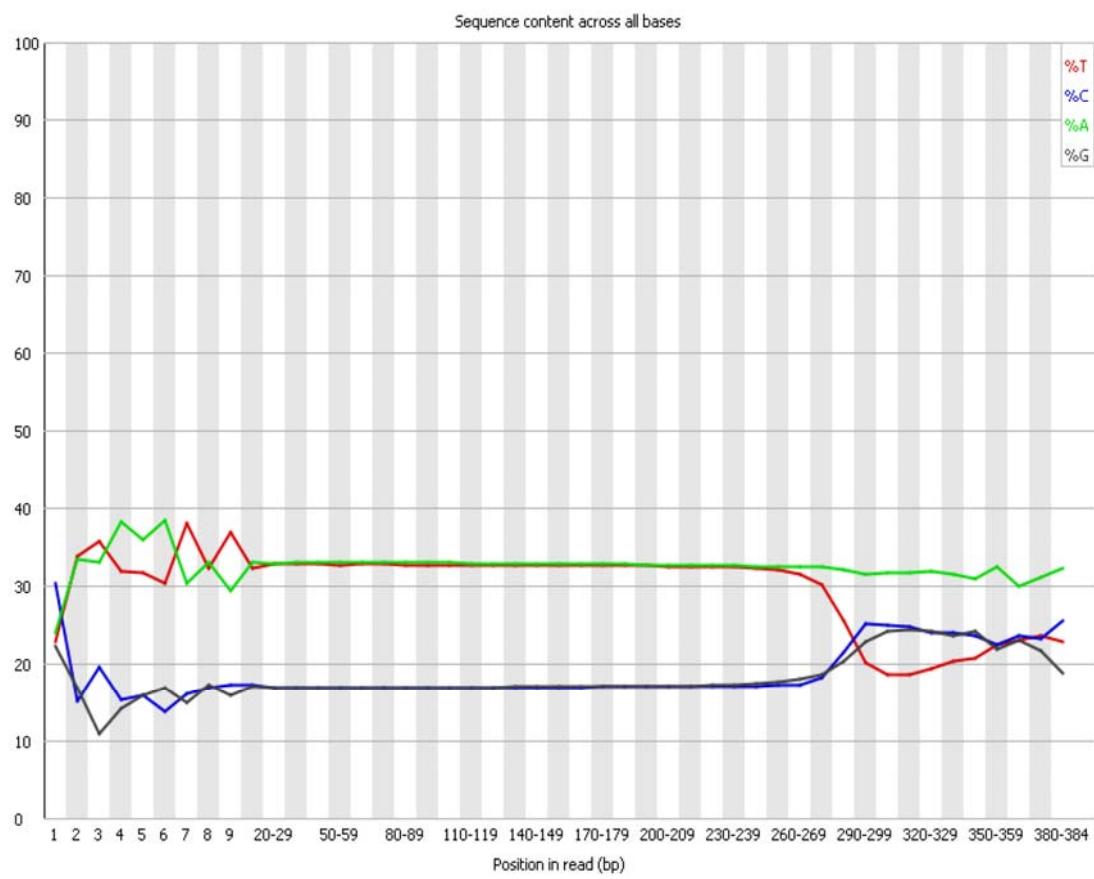
Per sequence quality scores

R_2013_02_14_10_51_11_user_SMA-85-NGS000035_Zygo_sma85.fa... file:///C:/Documents and Settings/14411911/My Documents/Ion Torrent ...



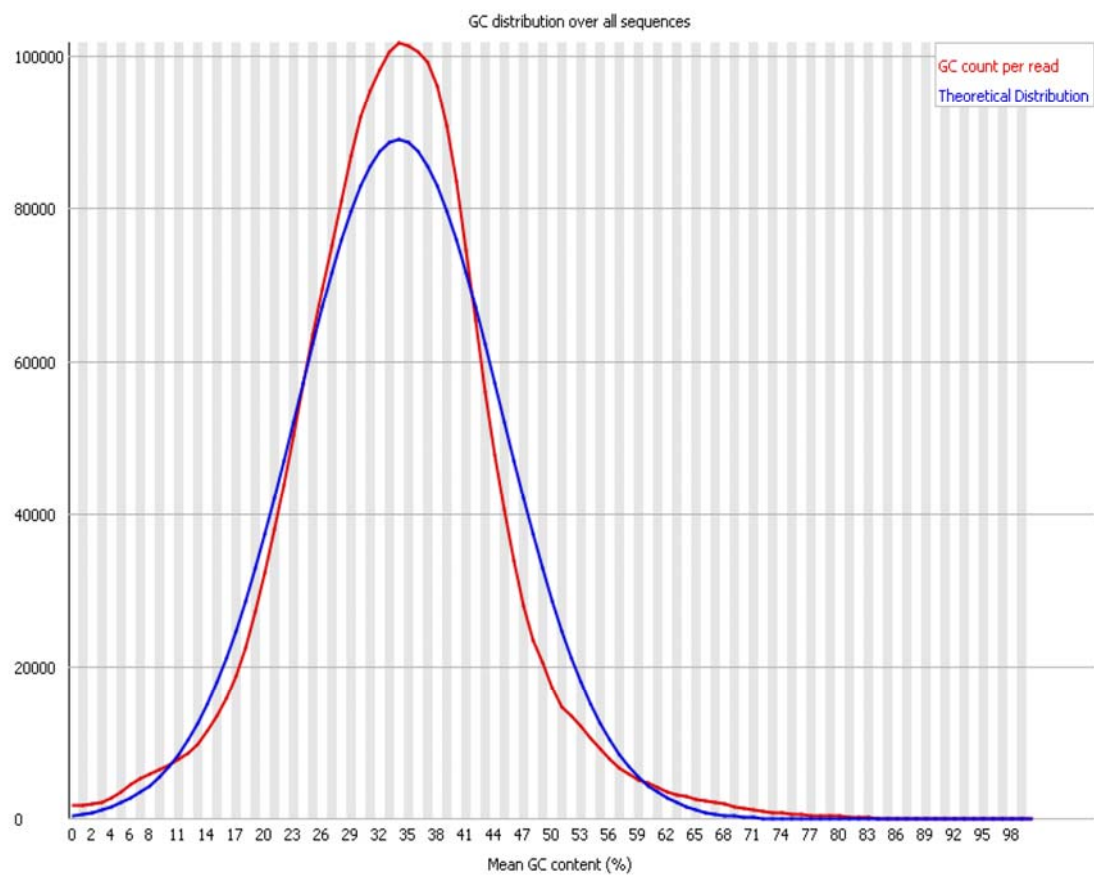
✖ Per base sequence content

R_2013_02_14_10_51_11_user_SMA-85-NGS000035_Zygo_sma85.fa... file:///C:/Documents and Settings/14411911/My Documents/Ion Torrent ...



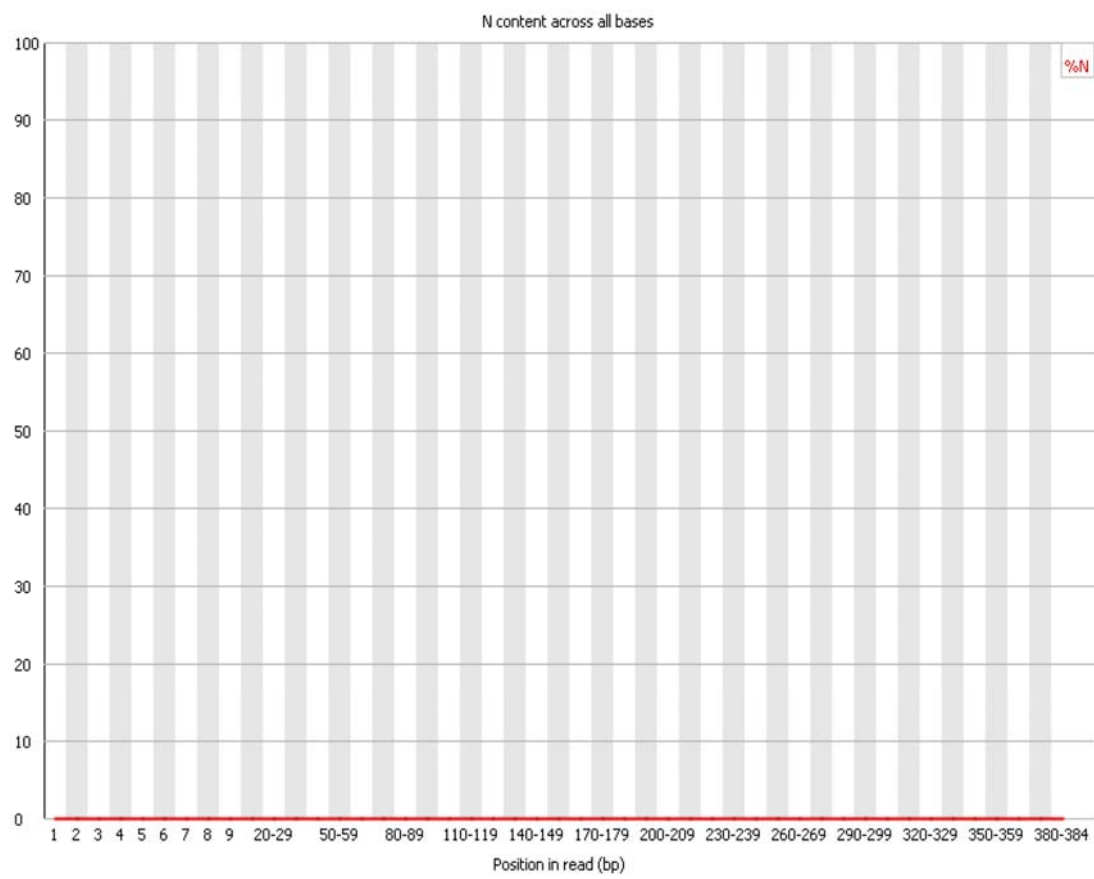
! Per sequence GC content

R_2013_02_14_10_51_11_user_SMA-85-NGS000035_Zygo_sma85.fa... file:///C:/Documents and Settings/14411911/My Documents/Ion Torrent ...



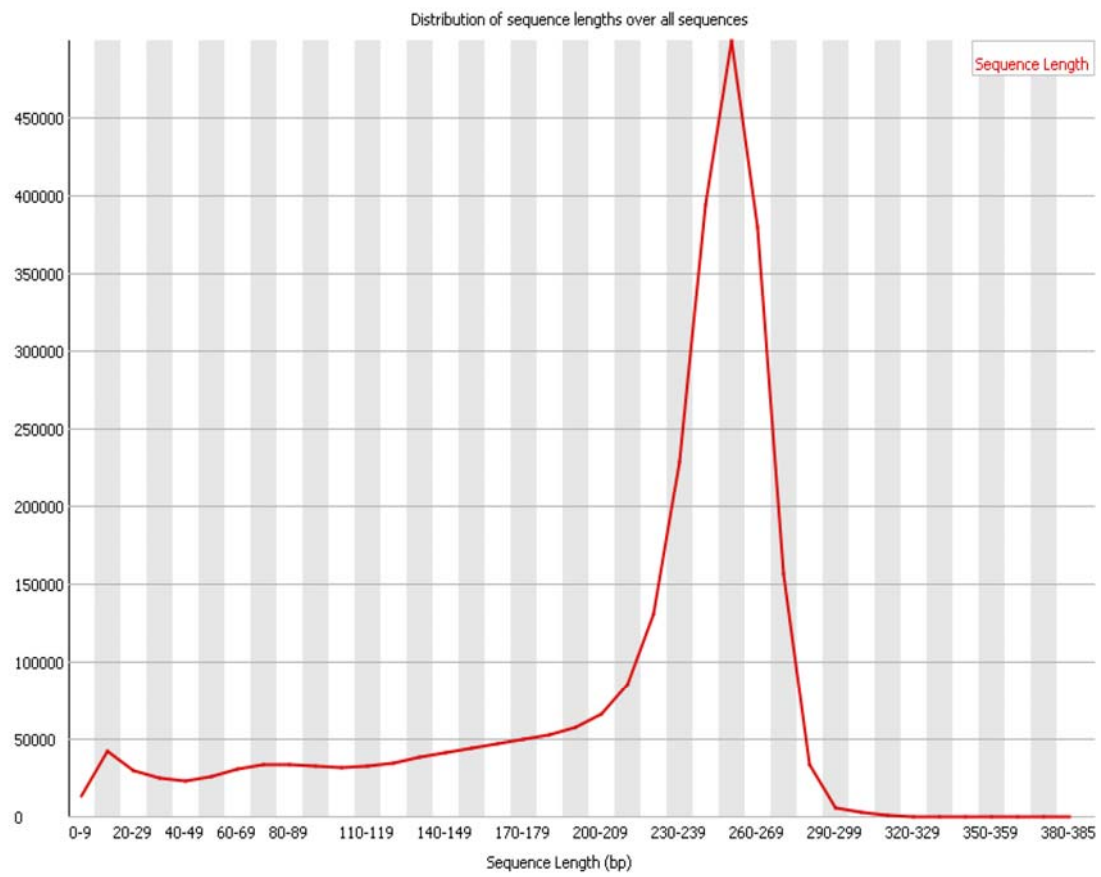
✓ Per base N content

R_2013_02_14_10_51_11_user_SMA-85-NGS000035_Zygo_sma85.fa... file:///C:/Documents and Settings/14411911/My Documents/Ion Torrent ...



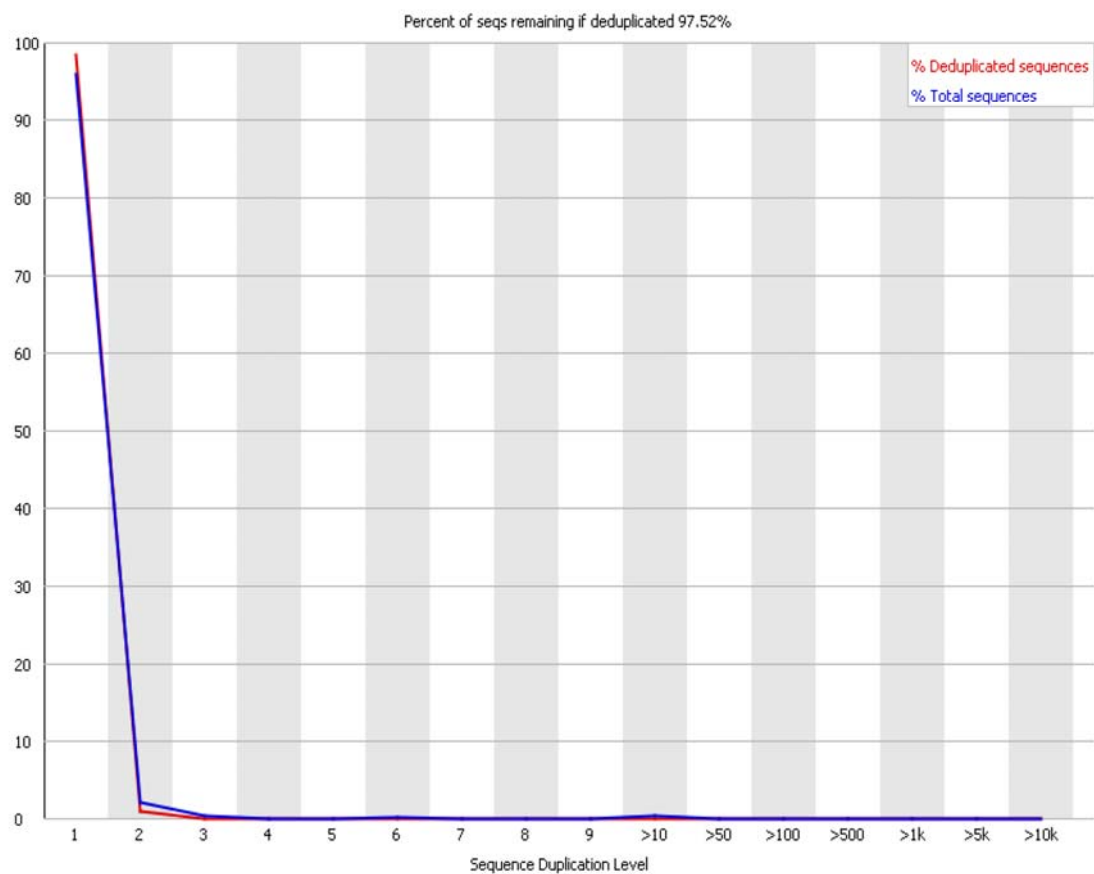
! Sequence Length Distribution

R_2013_02_14_10_51_11_user_SMA-85-NGS000035_Zygo_sma85.fa... file:///C:/Documents and Settings/14411911/My Documents/Ion Torrent ...



✔ Sequence Duplication Levels

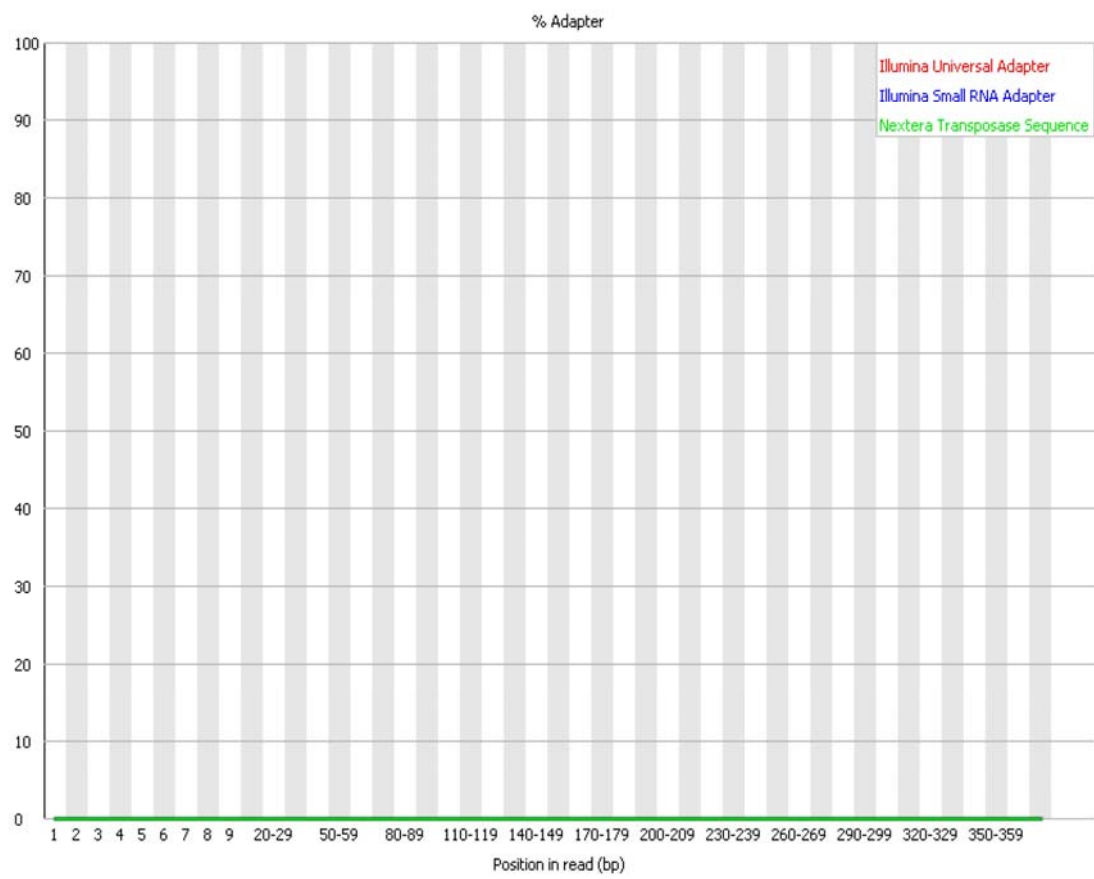
R_2013_02_14_10_51_11_user_SMA-85-NGS000035_Zygo_sma85.fa... file:///C:/Documents and Settings/14411911/My Documents/Ion Torrent ...



 **Overrepresented sequences**
No overrepresented sequences

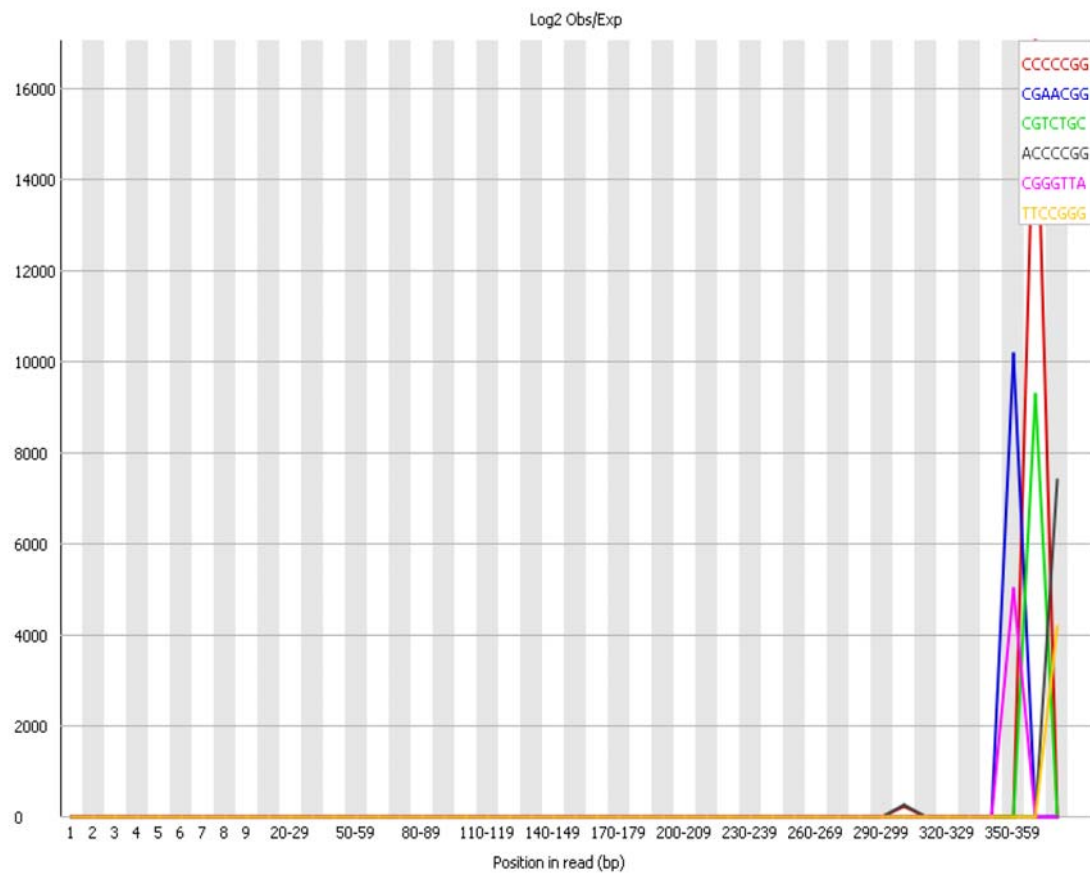
 **Adapter Content**

R_2013_02_14_10_51_11_user_SMA-85-NGS000035_Zygo_sma85.fa... file:///C:/Documents and Settings/14411911/My Documents/Ion Torrent ...



Kmer Content

R_2013_02_14_10_51_11_user_SMA-85-NGS000035_Zygo_sma85.fa... file:///C:/Documents and Settings/14411911/My Documents/Ion Torrent ...



Sequence	Count	PValue	Obs/Exp Max	Max Obs/Exp Position
CCCCCGG	300	2.5702422E-5	17021.768	360-369
CGAACGG	460	7.2513474E-5	10176.056	350-359
CGTCTGC	550	8.638639E-5	9284.601	360-369
ACCCCGG	280	1.428795E-4	7430.1367	370-377
CGGGTTA	930	2.9637714E-4	5033.318	350-359
TTCCGGG	955	4.3089542E-4	4201.3335	370-377
TTTCCGG	1805	4.7363018E-4	3890.016	370-377
ACCGGTT	1225	5.142044E-4	3821.213	350-359
AGGGCCC	710	6.045584E-4	3596.148	360-369
AAGGGCC	1425	8.896462E-5	3583.5295	360-369
GGGCCCT	510	6.6977035E-4	3441.9014	360-369
AACCCCG	735	7.740753E-4	3184.344	370-377
GCCCCAC	1030	8.426821E-4	3029.7642	330-339
GGGTTCC	1065	9.0091873E-4	2930.1946	330-339
CCCCGGG	275	0.0010034825	2836.9612	360-369

R_2013_02_14_10_51_11_user_SMA-85-NGS000035_Zygo_sma85.fa... file:///C:/Documents and Settings/14411911/My Documents/Ion Torrent ...

Sequence	Count	PValue	Obs/Exp Max	Max Obs/Exp Position
CGGTTTT	1870	9.984858E-4	2730.7646	360-369
CCGGTTT	935	0.0011482205	2612.036	350-359
TCCGGGA	1095	0.0011826685	2564.9236	370-377
CAATTCG	2015	0.0011593166	2534.258	360-369
TCCGAAT	2845	0.0011765689	2468.0066	370-377

Produced by [FastQC](#) (version 0.11.2)