Events That Shape Genomes

Stephen A. Schlebusch

University of Cape Town

Thesis presented for the degree

DOCTOR OF PHILOSOPHY

In the Department of Molecular and Cell Biology

University of Cape Town

November 2018

The copyright of this thesis vests in the author. No quotation from it or information derived from it is to be published without full acknowledgement of the source. The thesis is to be used for private study or non-commercial research purposes only.

Published by the University of Cape Town (UCT) in terms of the non-exclusive license granted to UCT by the author.

University of Cape Town

Acknowledgements

First and foremost, I would like to thank my supervisor Nicola Illing, whose guidance over the years has shaped me into the researcher I am today. She consistently believed in me, from when I was just beginning in her lab, through the good times and the bad; for that I am immensely grateful.

I'd also like to thank my co-supervisor, Jeff Wall. It was immensely comforting to have someone with Jeff's expertise to present ideas and findings to. Jeff also very generously shared interesting and impactful research, including the entirety of what became Chapter 3.

A big thank you needs to be given to Nicola Mulder, Gerrit Botha, Ayton Meintjes and Suresh Maslamoney, who kindly allowed me to use their computational resources and also helped set up my own. It is not an exaggeration to say that this PhD would not have been possible without their generosity.

Then I'd like to thank: The NRF for providing me with financial support, Cornelia Klak, who allowed me to "harvest" her Ruschioideae plants as well as helped conceptualise that project, Jessica Proctor who did the deceptively painstaking task of growing *Xerophyta* plants and making sure they never dried out, Rafe Lyall, who was a thoughtful and funny brainstorming companion to work with, the other members of Lab 425, specifically Mandy Mason, Dorit Hockman, Zoe Gill, Evan Milborrow and Ash Parker, who made my time at UCT more fun than it had any right to be and my friends Aleyo Chabeda and Frans Cronje, who kept me sane throughout these years.

Finally I'd like to thank my parents, Joy and Johan, for their unwavering love and support.

Abstract

The invention and development of Next Generation Sequencing has opened up new possibilities for exploring the genomes of non-model organisms. For this thesis, a diverse range of non-model species from both plants and animals were used to identify and answer questions of evolutionary interest in four case studies. In doing so, a wide assortment of methodologies were used and developed, taking full advantage of the versatility that whole genome sequencing can provide.

The genome of the Natal Long Fingered Bat, *Miniopterus natalensis*, was assembled to investigate the genetic mechanisms responsible for the evolution of the bat wing. The assembled genome was required to facilitate RNA-seq and ChIP-seq analysis. In addition to the genome assembly and annotation, dN/dS analysis and lncRNA prediction were also conducted. This resulted in a high quality genome assembly with just over 24000 genes being annotated and 227 putative lncRNAs being identified. None of the genetic pathways highlighted by the RNA-seq analysis showed any elevated dN/dS signal, suggesting this was not the loci of evolutionary change.

The Amboseli National Park in Kenya has a local population of Yellow baboons (*Papio cynocephalus*) that has recently come into contact and hybridised with a population of Olive baboons (*Papio anubis*). A genome assembly of *P. cynocephalus* was created and used to align low coverage sequencing from 45 baboons, including admixed individuals along with unadmixed individuals from each species. By identifying SNPs that were predictive of the species, hybrid individuals were confirmed and evidence for previous admixture events discovered, such as *P. anubis* SNPs already at fixation in the P. cynocephalus population at Amboseli.

The Ruschioideae are a clade of plants that encompasses the prolific tribe, the Ruschieae, which is comprised of approximately 1500 recently diverged species. An exploratory analysis sequenced two Ruschieae genomes (*Polymita steenbokensis* and *Faucaria felina*) along with a sister taxon (*Cleretum herrei*) from a neighbouring tribe (Dorotheantheae). The three plants were compared to each other in order to try and identify any genetic signatures that could be influencing the rapid speciation. The two Ruschieae species were found to have increased levels of non-tandem duplication within the genome as well as on going transposable element activity when compared *to C. herrei*.

Xerohpyta humilis is a desiccation tolerant plant. In order to further facilitate research into how this is possible, the genome was sequenced and assembled. Irregular data led to the discovery that the plant had a genome duplication as well as a large amount of somatic mutations in its genome.

Further analysis confirmed that this pattern of somatic mutations was only present in plants that had undergone multiple cycles of desiccation and rehydration.

These apparently disparate topics explored the possibilities and limitations for whole genome sequencing in the study of non-model organisms. Mechanisms of genetic change were examined at the genomic scale, from adaptation and hybridisation to various forms of duplication and mutation. In this way, a large variety of events responsible for the evolutionary change of genomes in plants and animals were analysed in a diverse set of systems.

Plagiarism Declaration

I hereby declare that the work within this thesis is my own original work, unless otherwise acknowledged.

Stephen Schlebusch

Table of Contents

Chapter 1: Introduction - Whole Genome Sequencing and *de novo* Assembly

The full genome sequence of an organism has always been a sought after resource for any geneticist. If a researcher only had access to that information, they would know all there was to know about the organism. At least that was the naïve expectation before assembled genome sequences actually started becoming available. This was driven home with the completion of the human genome project in 2003 (Schmutz et al, 2004) and the necessity of projects such as ENCODE (The ENCODE Project Consortium, 2011). After more than a decade of international collaboration and immense expense, the world was left with a genome that was predominantly filled with sequence of unknown function. In this chapter, the predominant method for whole genome sequencing and *de novo* assembly will be described. This will include several difficulties associated with *de novo* assembly using modern technology. The following chapters will then describe my efforts to make sense of the *de novo* assemblies of several non-model organisms' genomes.

The advent of Next Generation Sequencing (NGS) Technology has made many more genome assemblies possible. NGS first became available in 2005, with the description of a method to visualise the sequencing process of multiple DNA fragments in parallel (Margulies et al, 2005). This made it possible to sequence millions of DNA fragments at once and would signal the start of a new era in genetics. Collectively, companies using these methods rapidly drove the price of sequencing down by 4 orders of magnitude over just 10 years (Figure 1.1). This means that whole genome sequencing is no longer the massive investment it was in the past and can now be completed by individual laboratories.

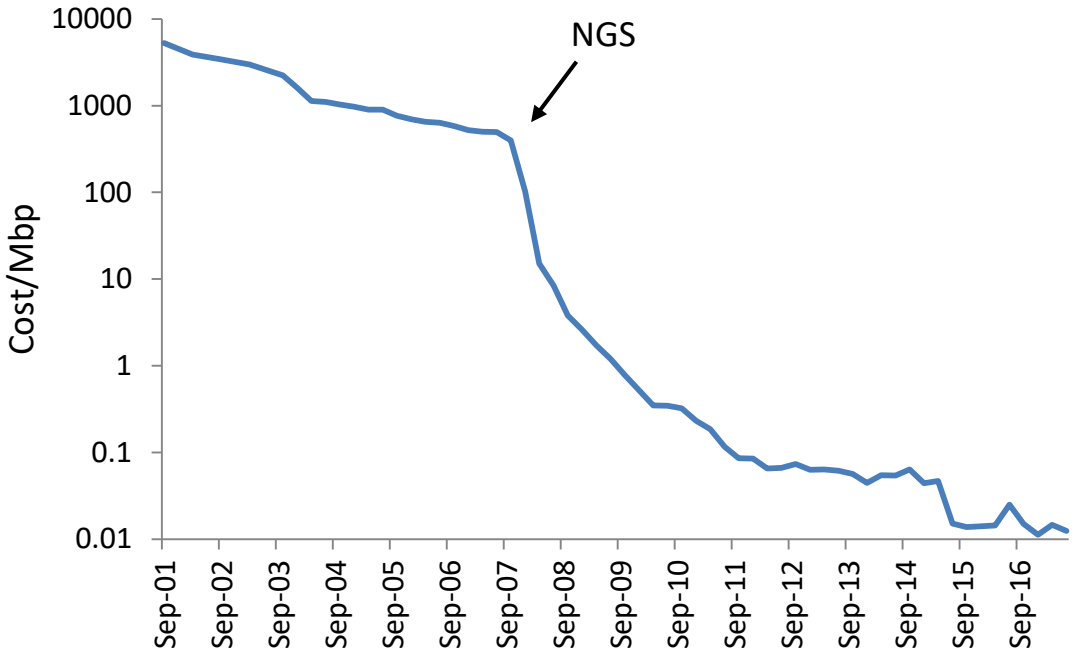


Figure 1.1: Change in the cost of sequencing per Mbp over time. Note that the cost is on a logarithmic scale. The introduction of Next Generation Sequencing is also displayed on the graph. The data was obtained from genome.gov/sequencing costs.

The massive gains in capability afforded by these technologies did not come without some downsides, however. The main contributor in the advancement of NGS technology has been Illumina, which has created a high output, low cost sequencing method (Schlebusch and Illing, 2012). Unfortunately, the length of uninterrupted sequenced DNA, known as the "read", is a lot shorter than the older Sanger sequencing read (Sanger et al, 1977). Sanger reads are relatively long at 1 kbp, while Illumina's NGS reads are generally about 100-150bp[1]. This difference in read length becomes very notable when considering that repetitive elements in genomes are often a lot longer than this short read length. This means that reads that originate from these repetitive regions are ambiguous in origin; and assembly of overlapping reads is highly fragmented (Figure 1.2).

[1] There are Illumina platforms with longer read lengths than this (e.g. MiSeq), but these are not normally used for the bulk of whole genome sequencing due to their higher cost per base pair.

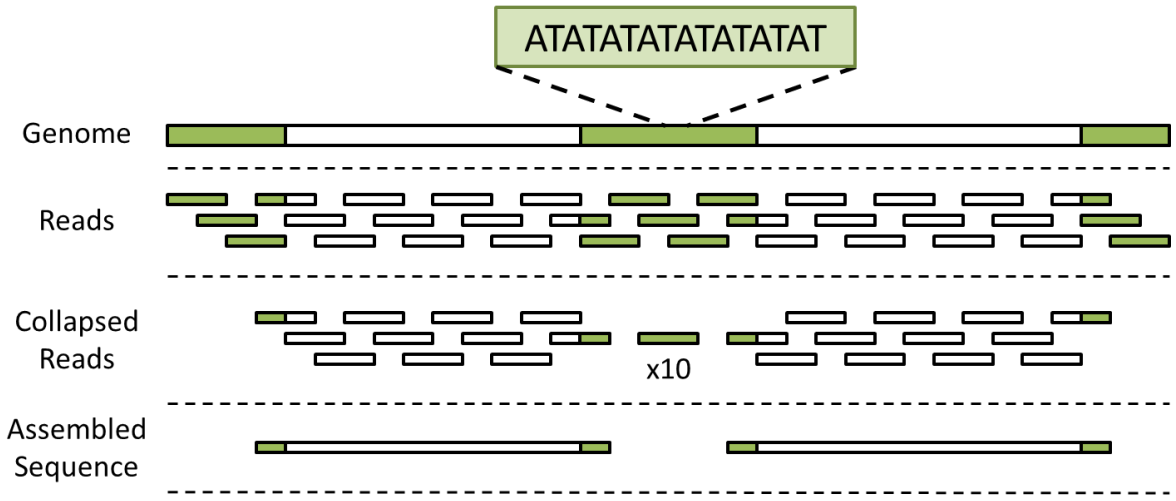


Figure 1.2: Using short reads to assemble sequences that include repetitive elements. Even when sequenced to completion, repetitive elements (in green) prevent the contiguous assembly of the genomic region. Each assembled sequence eventually ends when the reads become too ambiguous.

The problem of sequencing over repetitive regions can partially be solved by sequencing a read on either side of a DNA fragment of known size (known as the insert size). This is accomplished by randomly fragmenting the genomic DNA to the desired length before ligating the sequencing primers on the ends of each fragment to form a library. These reads can span a repeat element and resolve ambiguity in the assembly process. This is a central consideration of most genome sequencing strategies. Repeat elements vary in size, and as such, it is better to have libraries constructed from varying insert sizes if possible.

There are two methods that Illumina uses to create DNA fragments for sequencing, depending on the desired insert size. For smaller insert sizes (<1000bp) the fragments can be sequenced from either side using the attached primers. This is known as paired end sequencing. For longer fragments, mate-pair sequencing can be used. Mate-pair sequencing can sequence from either side of a much longer insert size (up to 20 kbp) by incorporating biotin onto the ends of long DNA fragments, and circularising the DNA (Mardis, 2011). This brings the two distant ends of the fragment in contact with each other. The circularised long fragments are then re-fragmented and the piece with the two distant ends is isolated using the biotin label (Figure 1.3). This new fragment can then be sequenced as normal.

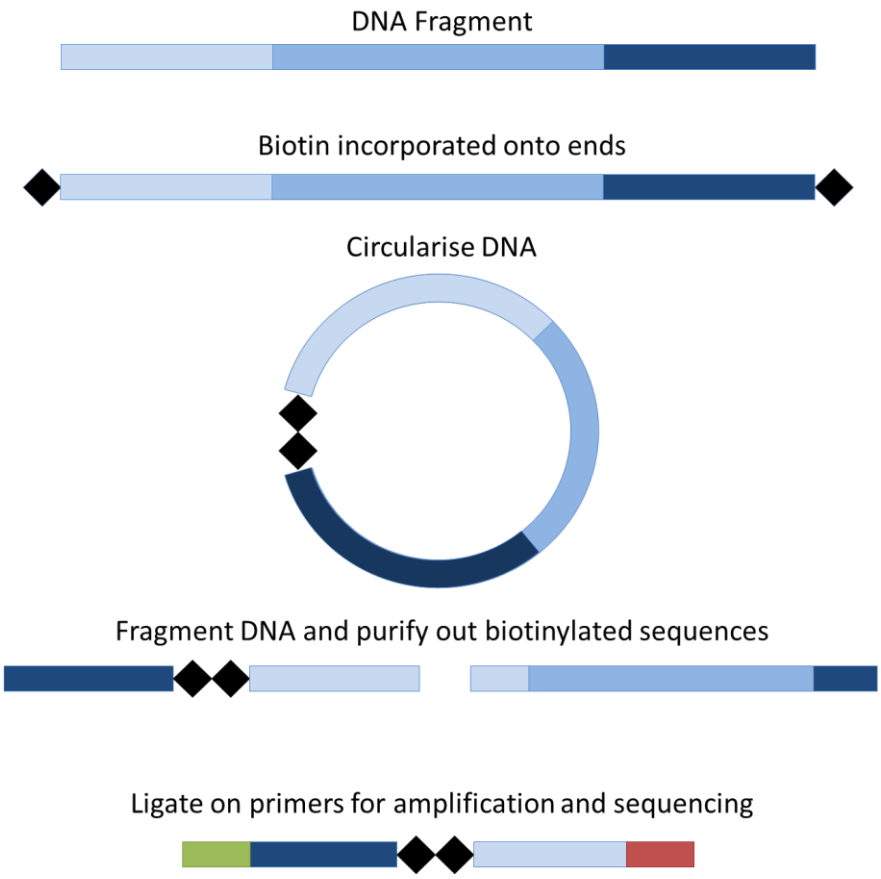


Figure 1.3: Creation of a mate-pair library for Illumina sequencing. The initial DNA fragment represents a range of about 1 kbp to 20 kbp. Figure adapted from Schlebusch and Illing (2012).

Illumina sequencing begins by affixing the DNA fragments to a specially prepared glass slide called the flow cell (Fedurco et al, 2006). This slide is covered in a lawn of small oligo primers for the DNA library to bind onto. Once bound, the DNA undergoes a process of specialised PCR called "bridge amplification" (Figure 1.4). This results in clusters of DNA fragments spread over the flow cell, where each cluster is made up of duplicates of a single founding fragment.

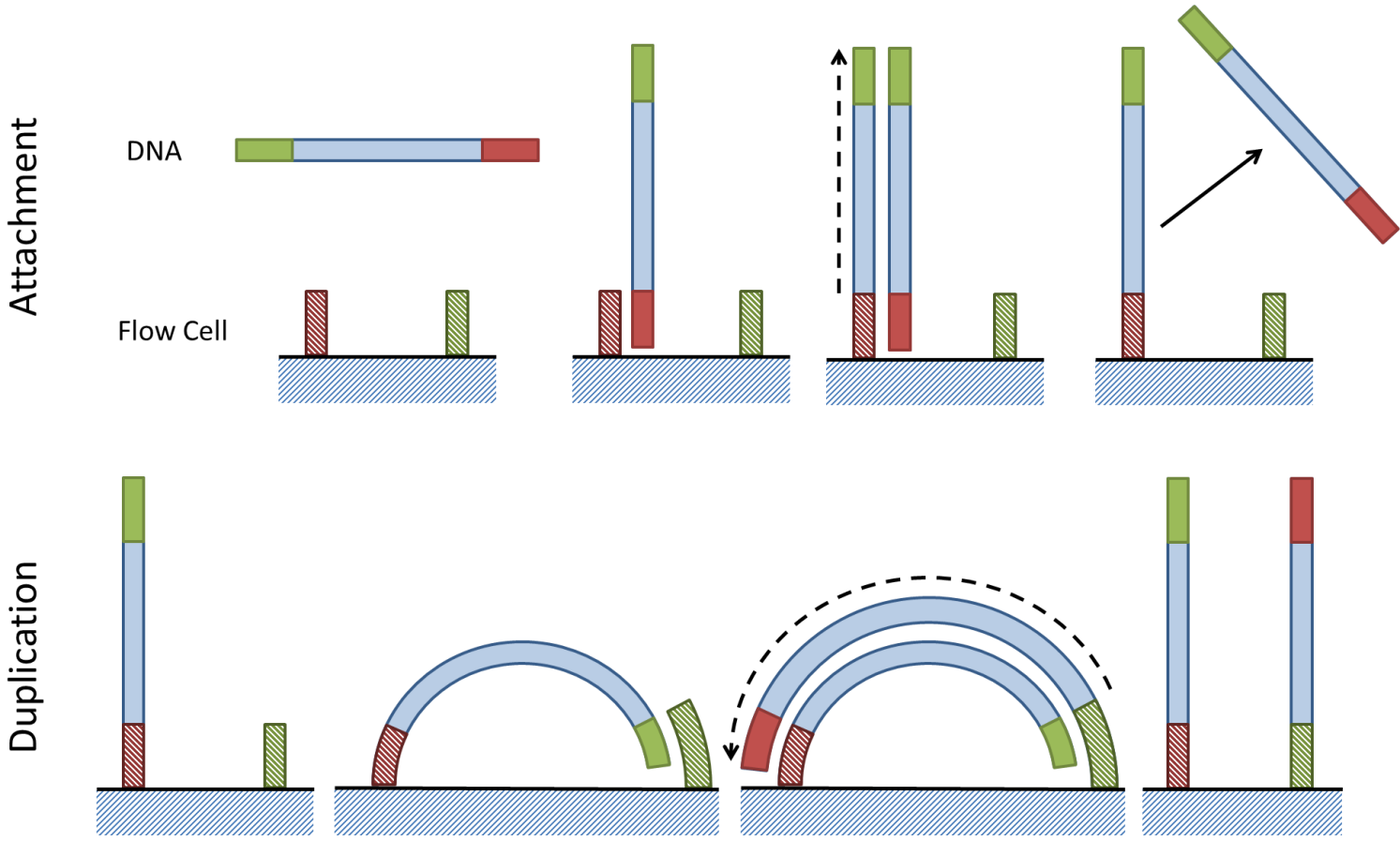


Figure 1.4: Diagram depicting bridge amplification on a flow cell. DNA (blue), flanked by sequencing primers (green and red), attaches to the flow cell by hybridizing to affixed olgio primers (striped green and red) that complement the ends of the sequencing primers. The oligio primer is extended (dashed arrow) to form a copy of the DNA fragment and sequencing primers after which the original fragment is discarded (solid arrow). The newly formed DNA fragment is then duplicated by having it form a "bridge" to the other oligio primer (striped green). Extension (dashed arrow) from the oligo primer (striped green) results in amplification. The process is repeated until DNA fragments form dense clusters. Once amplification is complete, the fragments can be sequenced. Figure adapted from Schlebusch and Illing (2012).

After amplification, clusters get sequenced from one of the two sequencing primers. Nucleotides are washed over the flow cell and appended after the primer (Bentley et al, 2008). These nucleotides are reversibly terminating, meaning that only one nucleotide can bind at a time. Each of the four nucleotides has a fluorescent dye associated with it. This fluorescent signal is amplified by the fact

that each of the fragments in a cluster bound the same nucleotide. The dye is then chemically removed, which allows for a new nucleotide to bind to its 3' end and the sequencing to continue. In this way, 1 nucleotide is sequenced per cluster per wash and the number of washes determines the read length. When the second read of a pair needs to be sequenced, the newly formed fragment is washed away and the process starts all over again using the second sequencing primer.

Long Read Sequencing

Many of the difficulties associated with the short read length of Illumina sequencing can be overcome using modern long read sequencing, currently exemplified by Pacific Bioscience (Eid et al, 2009) and Oxford Nanopore (Clarke et al, 2009) sequencing. These technologies can generate long reads (>10kbp) without any amplification step. This is significant for genomic sequencing, especially plant genomes (Li et al, 2017), as it allows reads to sequence through and span repetitive elements, even if those regions had abnormal GC contents that would make them difficult to amplify (during Illumina's bridge amplification, for example).

Unfortunately, sequencing with these technologies costs more than Illumina sequencing (Paajanen et al, 2017) and has a higher sequencing error rate than Illumina sequencing. For this reason, it is often advantageous to use these long read technologies in conjunction with Illumina sequencing. In this way, it is possible to use the long read technologies to scaffold the assembly, while the Illumina technology provides high coverage and accuracy.

Data Processing

Due to the large amounts of data generated in a NGS project, even basic analyses can be non-trivial. Generally speaking, the first priority is to assess the quality of the data and eliminate any sequencing errors. The data comes in Fastq format, with every base having a phred-like $\log_{10}$-based quality score associated with it (in ASCII code), based off the confidence the sequencing platform had when the initial call was made (Table 1.1). Generally, the quality of the first few bases is lower than average, as the sequencing platform calibrates itself. The quality then increases rapidly before gradually decreasing as you move towards the end of a read (Figure 1.5). These quality scores are used to trim low quality nucleotides from a read. Since the quality generally decreases as you move towards the

end of the read, the majority of quality control involves simply trimming the end of the read once the quality starts to drop below a certain threshold.

Table 1.1: Sequencing quality scores and their equivalent chance of error. A common range of quality scores and their corresponding ASCII notation is displayed. The chance of error is displayed as a fraction.

Quality Score	ASCII code	Chance of Error
10	+	10^{-1}
20	5	10^{-2}
30	?	10^{-3}
40	I	10^{-4}

Trimming low quality sequences can significantly improve the viability of data. But even with strict trimming of low quality bases, the average genome sequencing project can expect to have millions of sequencing errors go untrimmed[2]. An effective method for removing these remaining errors is to convert the reads into shorter segments known as k-mers. A k-mer is simply a sequence of k consecutive nucleotides. Reads can be broken down into all the associated k-mers by taking the first k nucleotides and then systematically shifting the selection by 1 nucleotide (Figure 1.6). A sequencing error in a read will alter the k-mers which are created using that nucleotide.

[2] A 1 in 1000 chance of error for 100 billion bases is still 100 million errors.

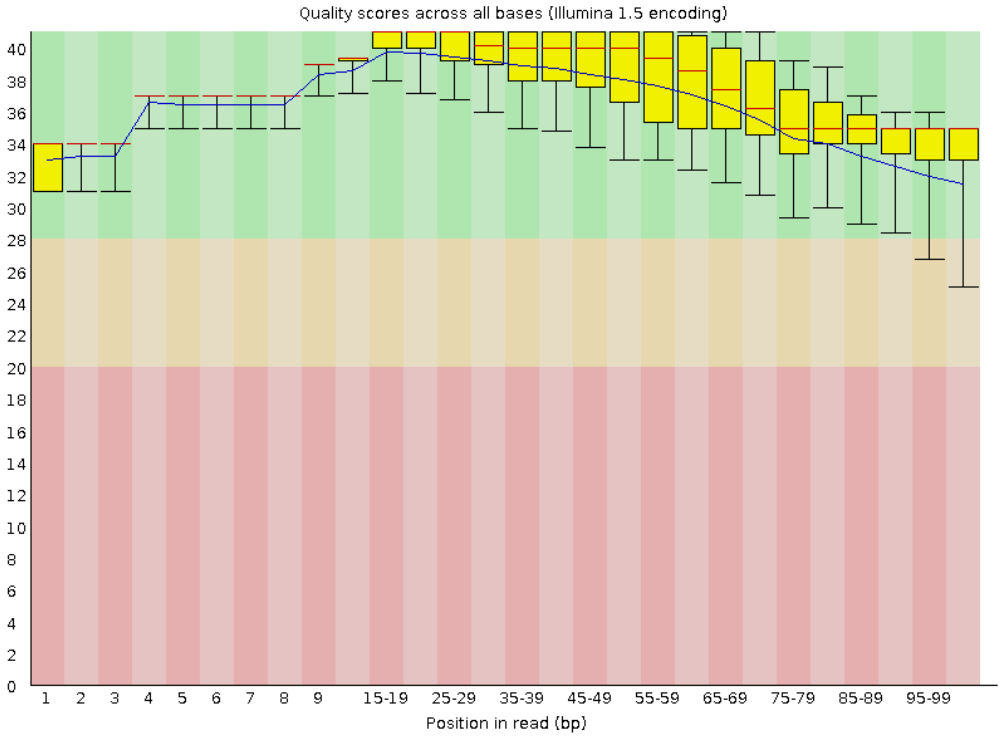


Figure 1.5: Example of sequencing quality scores across a set of reads. A box plots displays the variation in quality at varying base pair positions along the reads. Figure generated using Fastqc (Andrews, 2010).

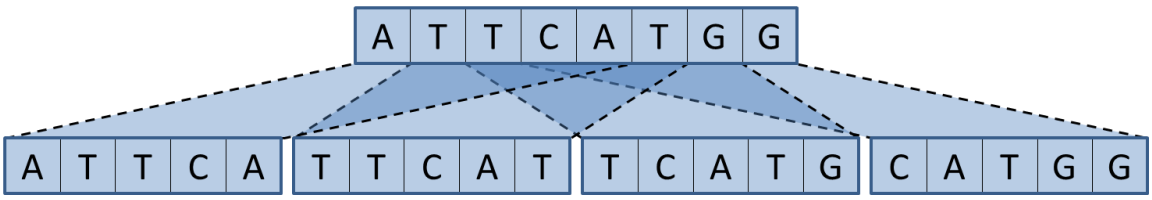


Figure 1.6: Breaking a sequence down into its constituent k-mers. The 8bp sequence above can be broken down into the 4 k-mers shown below, where k is equal to 5.

Due to the random nature of the genome fragmentation and sequencing, the average base pair will be sequenced a lot more than once in order to guarantee as many bases are sequenced as possible. The higher the sequencing coverage, the less of the genome is missed by chance and the more overlap there is between reads (important for genome assembly). This means that any k-mer from the genome should be sequenced many times and have a high frequency (Figure 1.7). In comparison,

a sequencing error should rarely occur at the same position and therefore k-mers created from sequencing errors should be unique or have a low frequency (Li et al, 2010; Pevzner et al, 2001)[3]. So the frequency of a k-mer can predict whether it is a real k-mer from the genome or a sequencing error, and the read can be adjusted accordingly.

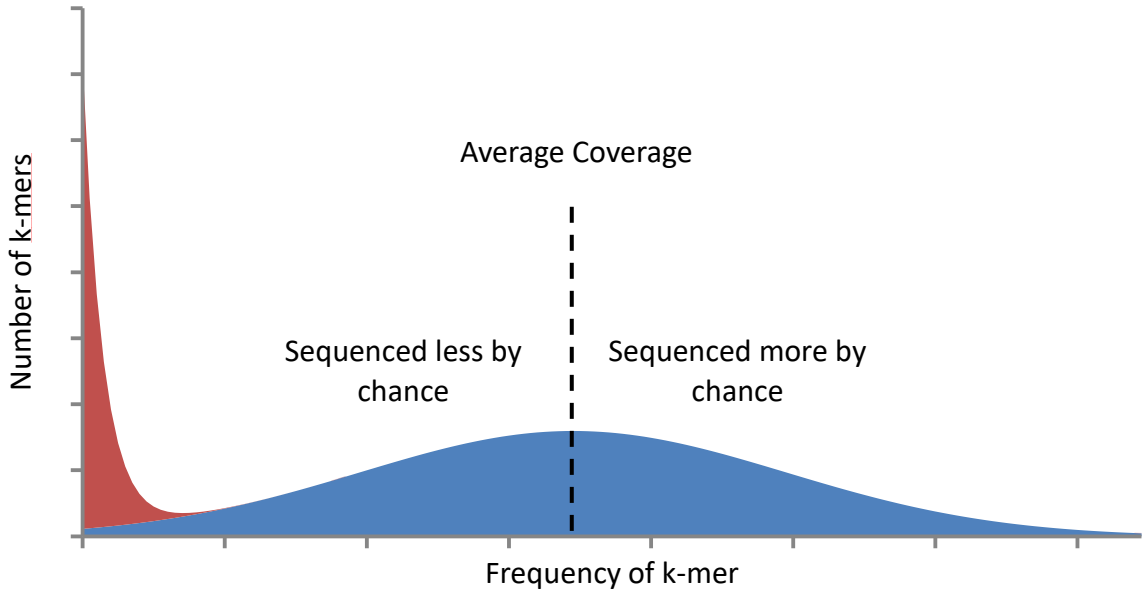


Figure 1.7: Expected distribution of k-mers at different frequencies resulting from whole genome sequencing for an inbred homozygous genome. Real k-mers within the genome (in blue) are sequenced multiple times with variation around the mean centred at the average coverage at which the genome was sequenced. Any k-mers created with a sequencing error in it (in red) are found at a low frequency, but usually in a high numbers, due to the number of sequencing errors found in the data.

Genome Assembly

Most current genome assembly programs are heavily reliant on k-mers and use a variation of the *de Bruijn* graph algorithm for genome assembly[4] (Gnerre et al, 2011; Kajitani et al, 2014; Luo et al, 2012; Simpson et al, 2009; Weisenfeld et al, 2017). This method scales well with the large amount of data generated by Illumina sequencing. And while there are disadvantages to the method, these

[3] Assuming 4^k is a lot larger than the genome size, where k is the k-mer size.
[4] Notable exceptions are SGA (Simpson and Durbin, 2011) and programs made for new long read technologies, such as PacBio (Chin et al, 2016).

aren't very noticeable with the short read data provided by the Illumina platform (Schlebusch and Illing, 2012).

De Bruijn graph algorithms work by directionally linking k-mers into a graph by joining any two k-mers which overlap by k-1 nucleotides, where k is the length of the k-mer (Figure 1.8). But in order for this to be effective, k-mers should be unique within the genome (if possible). This sets up a dichotomy, where a larger value for k results in more unique k-mers, but fewer sequences with the required k-1 nucleotide overlap (Miller et al, 2010). Once an effective value for k is found and the reads are plotted as k-mers, a consensus sequence can be read off without having to compute any overlaps between reads.

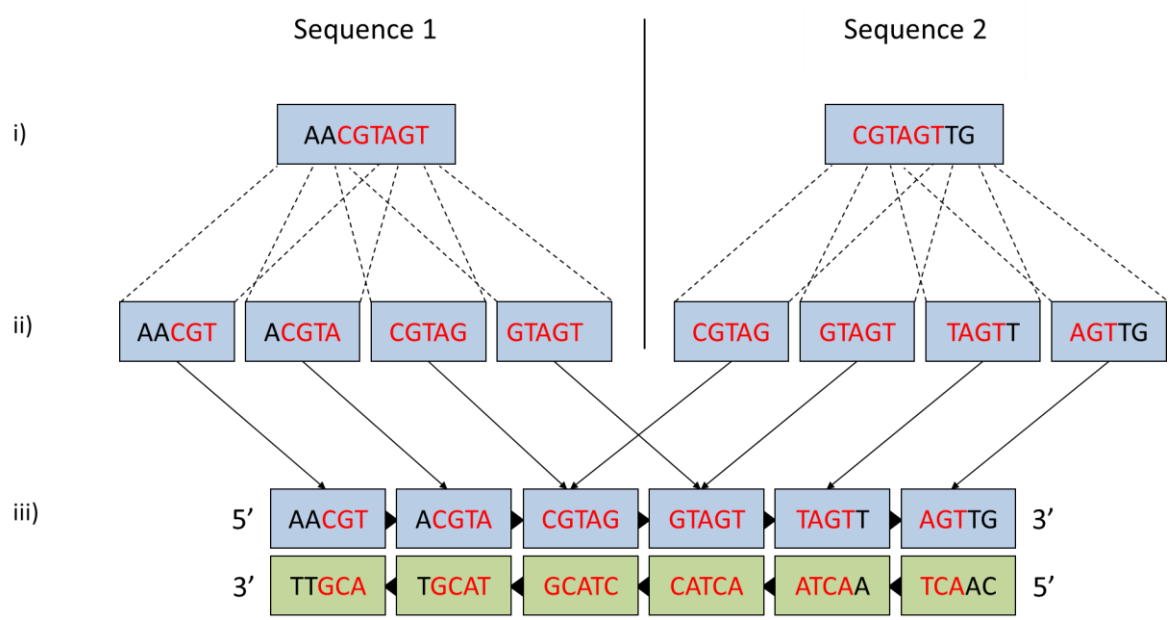


Figure 1.8: Two sequences with an overlapping region get joined by common k-mers. Sequence 1 and Sequence 2 are 8 nucleotides long, with an overlap of 6 nucleotides (i). These sequences can be broken down into 4 k-mers of 5bp each, two of which are common between the two sequences (ii). K-mers are mapped to a *de Bruijn* graph, where each k-mer is double stranded and linked to the adjacent k-mer (which overlaps by k-1 nucleotides) in a directional manner (iii). Figure adapted from Schlebusch and Illing (2012).

In a simple genome, such as a virus, this could be all that is required. But in a complex genome, filled with repeat elements and common sequences, the process is more complicated (Miller et al, 2010). Instead of a neat string of k-mers which can be read, the final product is a complicated tangle of sequences, with multiple nodes leading into common k-mers, paths diverging around heterozygous sites and sequencing errors adding large numbers of false k-mers into the equation (Figure 1.9)[5].

This complicated network of connections (Figure 1.9) can partially be resolved through path finding mechanisms and coverage calculations, but this approach is limited. Breaking a read down into k-mers loses valuable spatial information. In its original form, a sequence is not only linked to adjacent k-mers, but also to k-mers on the other end of the read (this can be seen in the reads in Figure 1.9i, compared to when it gets collapsed in the *de Bruijn* graph). In addition, the reads have insert length information associated with them which isn't portrayed in the *de Bruijn* graph. All of this information needs to be retrieved from the original reads (Figure 1.10). In this way, the knots and complex structures can be resolved and contiguous sequences (known as contigs) created. Contigs can then be ordered and orientated using the read pairs. This includes being able to approximate the distance between two contigs using the insert length of the library. This process is known as scaffolding, and the resultant sequences, which are a string of contigs separated by varying numbers of 'N's, are known as scaffolds.

Finally, reads can be used to fill in the missing nucleotides between contigs within a scaffold. Any read pair which includes a read that is predicted to map to a gap can potentially be used to fill in missing sequence. This allows for contigs to be extended by reads which presumably didn't have a large enough overlap to be connected in the *de Bruijn* network.

[5] Depending on where the error occurs within a read, a single sequencing error can add as many as 'k' new nodes to the graph. This means that if sequencing errors aren't removed, there will be billions of new k-mers to plot.

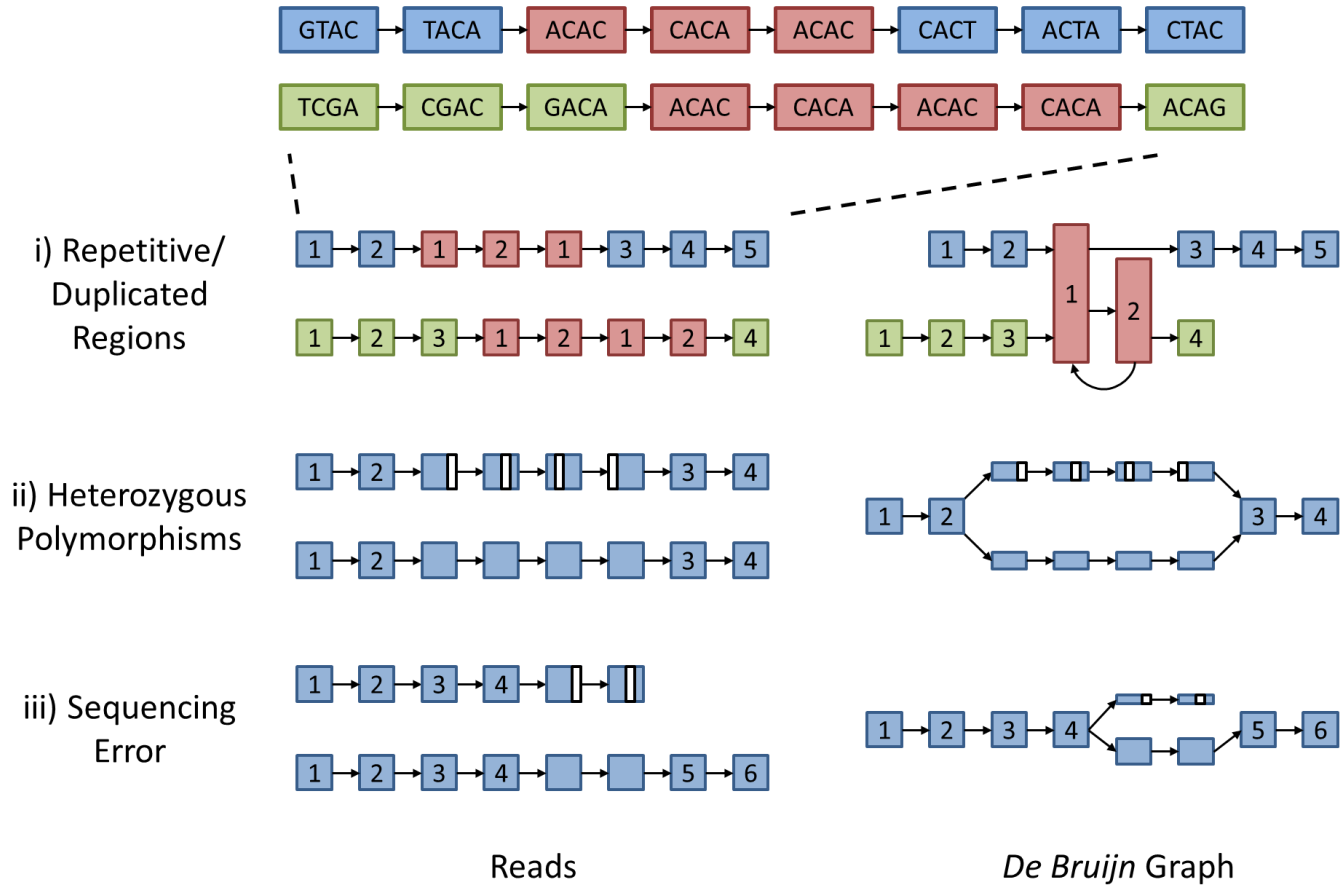


Figure 1.9: How complex genetic sequences appear when mapped onto a *de Bruijn* graph. Each block represents a single k-mer. The k-mers on the left represent k-mers broken down from reads. The k-mers on the right represent the k-mers mapped onto the *de Bruijn* graph. The relative coverage of a sequence in the *de Bruijn* graph has been shown by the height of the block. i) Repetitive or duplicated regions (represented in red) create ambiguous, high coverage k-mers in the *de Bruijn* graph. It is now unclear how many red blocks are associated with the blue and green sequences as well as whether the 2nd blue block is connected to the 3rd blue block or the 4th green block. ii) Heterozygous polymorphisms create two paths, each with half of the expected coverage. The polymorphism (in white) can be seen changing position in the alternative k-mers. iii) A sequencing error, which will have a low final coverage, normally occurs at the end of the read, and as such the error (in white) doesn't move through all the positions in the k-mers. As such, the two paths normally don't combine together again.

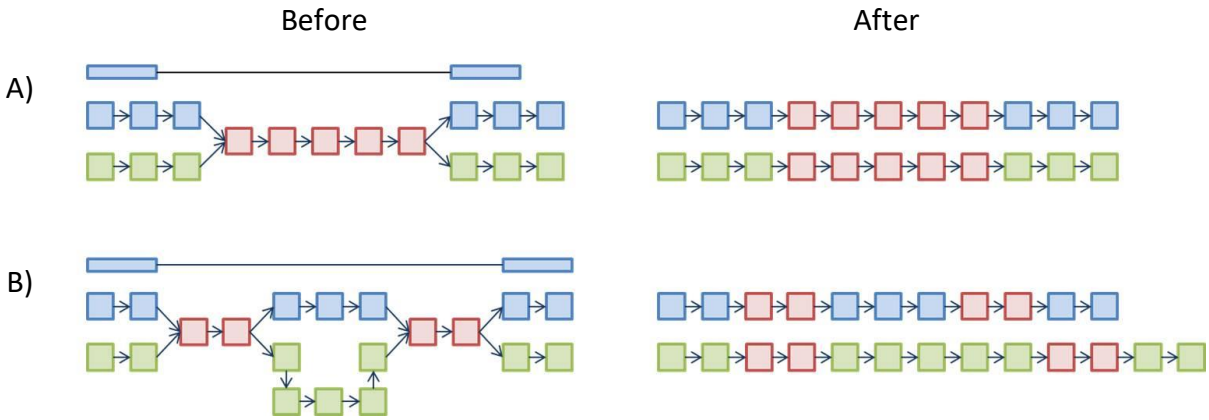


Figure 1.10: Using paired reads to resolve ambiguous *De Bruijn* connections. A) Reads which span common regions can resolve the ambiguity in a connection. B) Alternatively, the insert size itself can be used to resolve uncertain regions, where the length of one of the paths is prohibitive. Figure adapted from Schlebusch and Illing (2012).

Assessing Genome Quality

Due to the random nature of modern sequencing, "whole genome sequencing" very rarely actually results in the whole genome being sequenced. And the assembly is normally at least partially fragmented from ambiguous regions. So since the completion of a genome isn't a realistic scenario, it is necessary to find ways to evaluate incomplete assemblies and decide if it is "good enough" for the desired purpose. Measurements which seem intuitive, such as the average size of scaffolds and contigs within an assembly, are not very effective[6]. There are some other statistics which add to the general understanding of an assembly, such as the percentage of 'N's in the scaffolds, the cumulative length of the scaffolds and contigs, the lengths of the longest scaffolds and contigs, etc. But the most important statistics have proven to be the N50 and NG50 of an assembly (Earl et al, 2011).

The N50 is the length of the scaffold found half way through an assembly, if scaffolds are arranged from longest to shortest (Figure 1.11). This value effectively says that 50% of nucleotides in a genome assembly will be found in scaffolds of this size or larger. This statistic is a lot more

[6] This can been seen by imagining two assemblies which are identical except the second assembly has 100 extra contigs which are 100bp long. While these contigs are not very helpful, they are extra information which the first assembly doesn't contain. Despite this, the second assembly will have a worse average and median contig length.

representative of the assembly than something like an average, and is used instead in a similar manner. But the N50 value can be misleading if the size of the assembly is very different to the expected genome size. In these cases, it is better to use the size of the scaffold half way through the expected genome size, instead of the assembled size. This is known as the NG50 value.

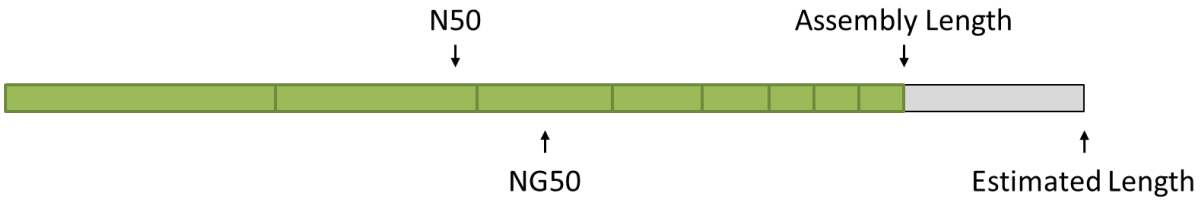


Figure 1.11: Positions of N50 and NG50 values within an assembly. Assembled scaffolds are portrayed in green. The unassambled region of the genome is displayed in grey.

Optimising a Genome Assembly

The existence of genome assembly metrics, such as the N50 size (which has a clear directionality in terms of what is better and what is worse) combined with easy to manipulate variables, such as the k-mer size, means that genome assembly optimisation can be attempted. Useful variables to manipulate, other than the k-mer size, include the insert size estimates of read pairs, and the genome assembly program used. The exact nature of the optimisation will vary with each genome assembled and project specific information should be taken into account whenever possible. For example, projects with less sequencing coverage might try use more lenient quality trimming steps in order to maximise the amount of sequencing that is used in the assembly process (with a higher risk of sequencing errors being incorporated into the assembly).

While metrics like the N50 size are good to use for optimisation, it is important to make sure other metrics, which aren't as useful for optimisation alone, are still monitored. For example, it is possible to increase your N50 size by increasing the size estimate of your insert lengths for the read libraries, but by doing so, a corresponding change would be observed in the percentage of N's assembled.

Benefits of Assembled Genomes

An assembled genome represents a wealth of possible information. The problem is, it normally also represents a lot more information that isn't of interest. It is the proverbial haystack that a researcher must search through. It is therefore a good idea to have specific goals in mind, with methods for accomplishing those goals, before undergoing a genome sequencing project. Common goals include:

- RNA-seq: Using modern NGS technology, RNA transcripts can be identified and quantified. While an annotated genome isn't necessary for RNA-seq analysis, it helps by allowing reads to be mapped to genes instead of transcripts, which may exclude variants. But in order for this to be effective, the genome assembly needs to be of high enough quality that genes are not split across multiple scaffolds.
- ChIP-seq: Chromatin Immunoprecipitation sequencing (ChIP-seq) allows for structural information of DNA, like bound transcription factors or histone modification, to be investigated. But in order to use this data effectively, a high quality genome assembly is essential. ChIP-seq data is normally mapped to non-coding regions, such as promoters and enhancer elements. Enhancer elements, which can be 100 kbps away, need to be linked to their associated gene. This means that scaffolds need to be very large, such that genes on either side of presumptive enhancers are known.
- Comparison of gene sequence: For the purposes of comparing the genetic sequences of two or more genes, using an annotated genome, rather than a transcriptome assembly, has the advantage that genes don't need to be actively expressed. There are several disadvantages however, namely the extra cost, difficulty in assembly and problems with *ab initio* gene prediction.
- Analysing Repetitive elements: The repetitive elements within a genome can vary a lot, not only between orders, but even between families and species. Depending on the context, these differences can be of interest, revealing genomic restructuring and local duplication activity. However, these are the regions that the *de Bruijn* graph method, using Illumina sequencing, struggles to assemble. This means that there is a risk that assembly error and/or bias could play a major role in the analysis.
- Promoter analysis: A successful promoter analysis obviously requires a genome in order to be completed. What may be overlooked is that some form of RNA-seq data is probably also required in order to find the first exon of transcripts, which can often be difficult to find using *ab initio* techniques.

- Population Genetics: A simple population genetics study does not require a high quality assembly. Reads just need to map uniquely with high confidence. However, analyses such as estimating population size changes over time and recombination rates benefit from higher quality assemblies. Unlike other genome sequencing endeavours, a large number of individuals often need to be sequenced.

In the following chapters, many of these themes will be explored to varying degrees. In Chapter 2, the *Miniopterus natalensis* genome was sequenced to allow for RNA-seq and ChIP-seq analysis and facilitated the comparison of gene sequences between *M. natalensis* and selected terrestrial mammals. Chapter 3 uses Olive and Yellow baboon individuals in a population genetics analysis, searching for gene flow between the two species. Three succulent plants from the sub family Ruschioideae are used in Chapter 4 to try and identify reasons for their rapid speciation. This analysis looks at differences in repetitive elements to try and account for observed differences in gene copy number. And finally, the desiccation tolerant plant *Xerophyta humilis* was sequenced for the purpose of RNA-seq and promoter analysis, however as will be discussed, problems caused by abundant somatic mutations partially hampered these efforts.

Chapter 2: *Miniopterus natalensis* – The Natal Long Fingered Bat

Comparative biology is a branch of biology that aims to use differences between species to deduce information on a particular trait. The field therefore tends to focus on non-model organisms with unusual morphologies. The focus on non-model organisms means there is a lot of potential for Next Generation Sequencing to contribute to the field of study. This is especially relevant considering that exploratory projects that aim to link a morphological trait to a genetic component lend themselves to large scale sequencing projects. This is an area in which NGS technology is well suited, allowing for a large proportion of an organisms genes and genome to be tested.

An example of a mammal that displays potential for study using this framework is the bat (order Chiroptera). Bats are interesting in this regard for multiple reasons. They have previously been studied for their ability to echolocate (Parker et al, 2013), their disproportionately long life spans (Seim et al, 2013) and their highly adapted forelimbs (Mason et al, 2015; Hockman et al, 2008). As a result of this sort of study (and general interest), the Chiroptera have had 8 high coverage genomes sequenced and assembled (Figure 2.1). These studies have been buoyed by the fact that bat genomes (2.4 Gbp) are, on average, smaller than other mammal genomes (primates and rodents are 3.6 Gbp; Gregory, 2019). This decrease in genome size is partially explained by a disproportionate loss of DNA combined with lack of DNA gain and less transposable elements (Kapusta et al, 2017).

In order to study the genetic reason for the divergence in morphology of the forelimb compared to the hindlimb, it is helpful to have access to the contrasting gene expression levels at the time of divergence. But in order get this information, limb tissue is required from developing embryos. This requires a breeding population of bats that is accessible and which won't be significantly impacted by the loss of several individuals. For these reason, the Natal Long Fingered Bat, *Miniopterus natalensis*, was chosen.

M. natalensis is a Vesper bat found throughout Africa (Figure 2.2A), and is considered a low risk of becoming a conservation concern (Monadjem et al, 2017). Every year, a migratory population of *M. natalensis* roost in the Guano maternity cave (34° 25' S / 20° 20' E) in the De Hoop Nature Reserve (McDonald et al, 1990). Pregnant bats can be collected from this site without making a big difference to the integrity of the population, making the site ideal.

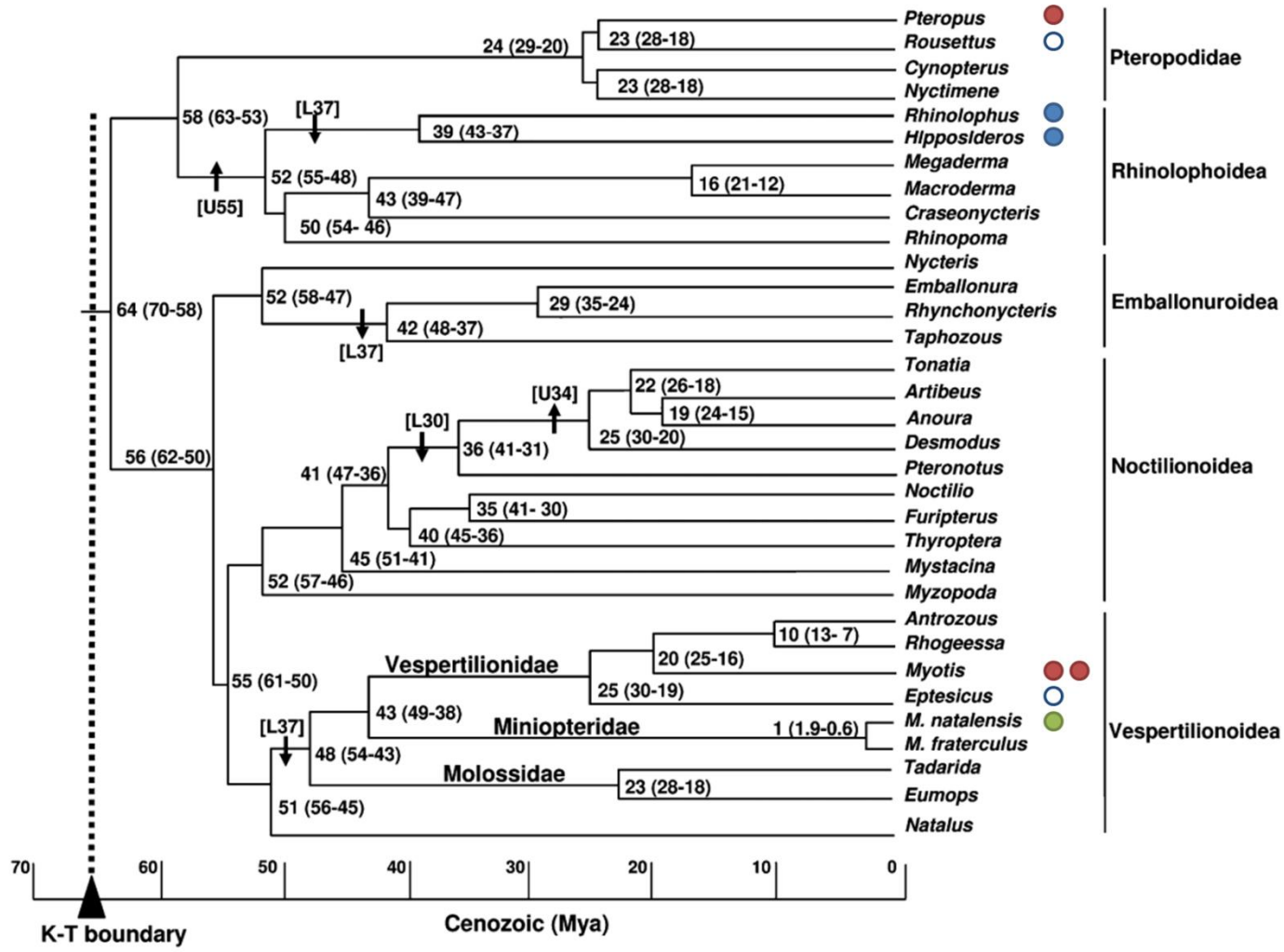


Figure 2.1: Chiroptera phylogeny showing sequenced high coverage genomes. The number of species with a sequenced genome in a given branch is displayed by the number of coloured dots displayed next to the name. The green dot displays the genome sequenced as part of this project (*Miniopterus natalensis*, Eckalbar et al, 2016). The blue dots represent genomes that were sequenced after the *M. natalensis* publication (*Rhinolophus sinicus* and *Hipposideros armiger*, Dong et al, 2017), while the red dots show genomes that were sequenced before (*Pteropus alecto* and *Myotis davidii*, Zhang et al, 2013; *Myotis brandtii*, Seim et al, 2013). Hollow dots represent unpublished genomes (*Rousettus aegyptiacus*, Accession: LOCP00000000.2; *Eptesicus fuscus*, Accession: ALEH00000000.1). Phylogeny adapted from Miller-Butterworth et al (2007).

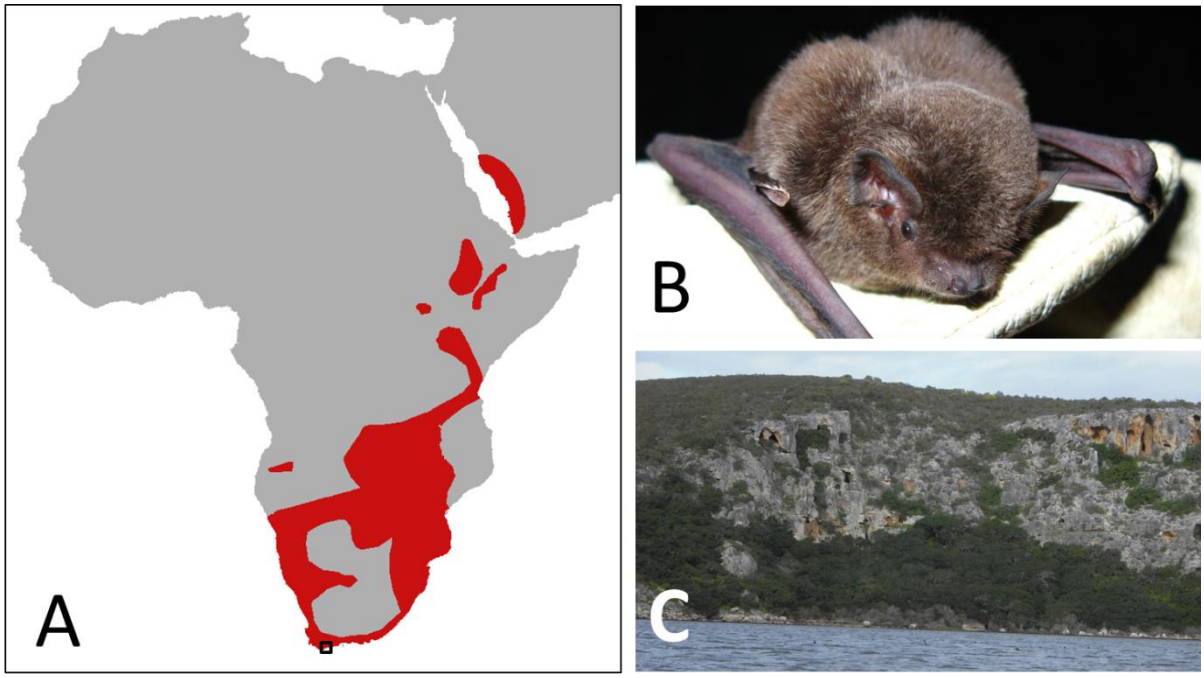


Figure 2.2: An introduction to the bat *Miniopterus natalensis*. A) The distribution of *M. natalensis* is shown in red (Monadjem et al, 2017). The position of the De Hoop Nature Reserve is shown by the small black box at the tip of Africa. B) The Natal long fingered bat, *M. natalensis*. C) Picture of the area around the roosting cave in the De Hoop Nature Reserve where the bats were captured.

The Project

In order to investigate the genetic mechanisms enabling flight in bats, a multifaceted approach was undertaken in collaboration with Prof Nadav Ahituv from the University of California, San Francisco (UCSF). This collaboration aimed to identify the key genetic determinants involved in the formation of the bat's wing by comparing RNA-seq and ChIP-seq data from the forelimb to that of the hindlimb in developing *M. natalensis* embryos. In addition, whole genome sequencing and assembly was performed in order to effectively utilise the data. As a member of the collaboration, I was mainly focused on the creation and annotation of the required reference genome, as well as lncRNA characterisation and dN/dS analysis. The work was published in Nature Genetics (Eckalbar et al, 2016).

Genome Annotation

Genome annotation is the process of identifying the positions of genetic features within the genome[7]. For the purposes of this project, it involved finding the co-ordinates of genes and repetitive elements, but could also refer to annotating other features, such as enhancers, SNPs, etc.

There are several different strategies that can be used to annotate the gene locations, including *ab initio* prediction, aligning assembled transcripts to the genome and looking for known genes using sequence conservation. Individually, each of these methods has flaws, but can be combined to make a more comprehensive annotation. Maker2 (Holt and Yandell, 2011) is very helpful in this regard, as it collates input from multiple types of programs, such as Blast (Altschul et al, 1990), Augustus (Stanke et al, 2004), and Exonerate (Slater and Birney, 2005), and determines the best supported co-ordinates for the genes.

For annotation of repetitive elements, programs like Repeat Masker (Smit et al, 2013) can be used to search for both known and novel repetitive elements.

Identification of lncRNA

Long Non-Coding RNAs (lncRNA) can perform a variety of functions in the cell (Mercer et al, 2009) and are thought to play important roles in development (Kung et al, 2013). This includes the regulation of the developmentally important Hox genes (Rinn et al, 2007; Wang et al, 2011). Particularly, any lncRNA that is unique and conserved within the Chiroptera and expressed within the limbs would be of interest as it may have a role in flight. However, identifying lncRNAs can be challenging. For example, methods that are commonly used to analyse a protein coding gene, such as identifying conservation in amino acid sequences, will not be effective. In addition, sequences of lncRNAs are often poorly conserved across taxa (Johnsson et al, 2014). Unfortunately, lncRNA have little in the way of defining features, as a group, other than their size and the fact that they don't get translated. Hence, identification of putative non-coding transcripts focuses on these factors rather than functional identification using tools such as the Coding Potential Calculator (Kong et al, 2007).

[7] Not to be confused with annotation of genes or transcripts, which would involve naming the genes, allocating GO terms, etc.

dN/dS Analyses

Another way to identify genes that may be involved in flight is to look for genes that have been under selection in the bat lineage. A method to determine if selection has been acting on a gene is to compare the rate that non-synonymous mutations (dN) to the synonymous mutation rate (dS). This provides insight into the strength of selective pressure that has affected a gene. Genes that have relatively little selective pressure on them will have dN/dS ratios near 1 (Kimura, 1977; Yang and Bielawski, 2000; Goldman and Yang, 1994; Muse and Gaut, 1994). If a gene had undergone positive selection instead, you would expect more non-synonymous mutations, but a similar number of synonymous mutations, and therefore a higher dN/dS ratio. And if the gene had been under purifying selection, there would be fewer non-synonymous mutations and therefore a low dN/dS ratio.

A dN/dS analysis, looking for genes under positive selection in bats, has already been conducted by Seim et al (2013) using *Myotis brandtii*. This analysis used a focused approach for finding genes under selection, excluding genes that either weren't present in all the aligned species or that were from error prone gene families. This analysis, while admittedly not focused on the evolution of flight, did not highlight any genes of interest for limb development. Therefore, in order for another dN/dS analysis on bats to be relevant and novel, the methodology will need to differ from that employed by Seim et al (2013).

Methods

DNA isolation and sequencing

DNA was extracted from the muscle tissue of an adult *M. natalensis* male, captured at the de Hoop Nature Reserve (Cape Nature Permit AAA007-00041-0056; UCT Science Faculty Animal Research Committee ethics approval 2012/V39/NI). The extraction used a phenol-chloroform extraction (Strauss, 1993), followed by RNAse A treatment and purification by column (Qiagen Genomic Tip 100/G, Ref 10243). The extracted DNA was sent to the University of Colorado, where three short insert libraries (175bp, 300bp and 600bp) and three long insert libraries (2 kbp, 5-6 kbp and 8-10 kbp) were generated. The libraries were then sequenced on a HiSeq2500 at the University of California Davis, with a second round being ordered after initial assembly results suggested a higher coverage was required (Table 2.1 for sequencing plans).

Table 2.1: The amount of sequencing planned for each library. The sequencing was conducted in two phases. A lane of sequencing is up to 200 million pairs of reads.

Number of Lanes		Insert Size
1st Round of Sequencing	2nd Round of Sequencing	
2	0	175bp
2	0	300bp
0	1	600bp
0.33	0.33	2-4 kbp
0.33	0.33	5-6 kbp
0.33	0.33	8-10 kbp

Read Processing

Sequencing quality was checked using Fastqc (version 0.10.1) (Andrews, 2010) before using Trimmomatic (version 0.32) (Bolger et al, 2014) to remove bases with a quality score of 17 or lower[8]. Trimming was done from the 5' and 3' ends of the 175bp and 300bp libraries. Only the 3' end of the other libraries was trimmed to allow for more effective duplicate removal from these libraries. Reads that were shorter than 60bp after trimming were removed from the dataset. Then the pairs from the 175bp library were merged together using Flash (version 1.2.6) (Magoč and Salzberg, 2011) if there was an overlap of 10 or more nucleotides between them[9]. Read pairs which weren't merged were treated as having an insert size of 200bp where appropriate in downstream processing.

A k-mer frequency plot (with k = 27) was created using the trimmed reads from the 175bp and 300bp libraries (these libraries had the least duplicates) and KmerFreq_HA (version 2.01) from the SOAPdenovo package (Luo et al, 2012). These k-mer frequencies were then used to error correct all of the read libraries, with any k-mer with a frequency of 3 or less being flagged as untrustworthy by Corrector_HA (version 2.01) (also part of the SOAPdenovo package). If possible, these k-mers are

[8] Equal to a 1 in 50 chance of a sequencing error

[9] Quality trimming these reads before merging means that less of them overlap and merge with Flash. The problem is that the quality trimming software assumes a steady decline in quality in a read, and this assumption doesn't hold if two reads have been merged end to end.

changed by 1bp to a more common k-mer from the set unless the sequencing confidence score was 40 or higher. A maximum of two corrections were performed on each read unless it was one of the merged 175bp reads, in which case four corrections were allowed. Further erroneous k-mers result in the read being trimmed (with a minimum read length of 60bp).

After error correction, duplicate reads were removed from the relevant libraries (600bp and larger) using FastUniq (version 1.1) (Xu et al, 2012). This order of processing, with duplicate removal occurring last, was found to be the most effective way of removing duplicates (duplicates can still be identified after a sequencing error). But as a result, these libraries could not be trimmed from the 5' side and could not be used in the k-mer frequency calculations. Due to the large percentage of duplicates found in these libraries, this emphasis seemed appropriate. And in this case, the downside was not very serious. Read quality on the 5' side was generally good (unless the whole read was bad) and there was enough coverage in the other two libraries to create differentiation between the real k-mers and the sequencing errors in the k-mer frequency plot.

Genome Assembly

The processed reads were assembled using SOAPdenovo2 (version 2.04, k=49) (Luo et al, 2012). The assembly process was primarily optimised using the N50 scaffold statistics, while statistics such as the genome size, the percentage of unknown nucleotides (N's) and the N50 contig size were monitored and taken into consideration. Parameters that were adjusted included the k-mer size of contig creation and scaffold creation, insert size between read pairs, the order in which the libraries were used in the scaffolding process, and the order of the read processing steps (for example, were duplicates removed before or after error correction, was the 175bp library was merged before or after quality trimming, etc).

Once the assembly with the largest N50 scaffold size was identified, the GapCloser program (version 1.12) (from the SOAPdenovo package) was used to reduce the percentage of N's and improve the assembly[10]. Finally, Cegma (version 2.4) (Parra et al, 2007), a program that identifies conserved genes that should be present in a eukaryote genome assembly, was used as a final quality check for the genome and to make sure the assembly was coherent.

[10] It would have been preferable to gap close all the assemblies after gap closing and then check their statistics, but due to the run time of the GapCloser program, this was not an option.

The genome assembly, as well as the raw sequencing reads are available on NCBI (accession code PRJNA283550).

Genome annotation

In order to annotate the genome, transcripts were used from a variety of sources. First, a *de novo* transcriptome assembly (Eckalbar et al, 2016) was aligned to the genome using Blastn (version 2.2.29) (Altschul et al, 1990). This draft was assembled to maximise the number of transcripts (at the expense of redundancy), and contained 6.1 million transcripts. To expand coverage in the assembly, which had been generated from RNA gathered from limb tissue, 960 thousand transcripts from *M. brandtii* were aligned with slightly relaxed Blastn settings[11]. These transcripts were generated with RNA from liver, kidney and brain tissue (Seim et al, 2013). Finally, 78 thousand mouse proteins were aligned using Tblastn. Exonerate (version 2.2.0) (Slater and Birney, 2005) was used to improve the exon/inton boundaries implied by the Blast alignments.

Ab initio gene prediction was conducted with Snap (version 2009-02-03) (Johnson et al, 2008) and Augustus (version 2.5) (Stanke et al, 2004). Snap was optimised with earlier runs of Maker, using the RNA-seq data, while Augustus used the predefined "Human" optimisation settings. Repetitive elements were soft-masked using RepeatMasker (version 3.1.6) and the mammalian repeat database in order to facilitate downstream analysis (Smit et al, 2013).

Bat LncRNA identification

All of the annotated transcripts were compared to the UniProt database using Blastx (Altschul et al, 1990). Genes that had no hit were aligned to the lncRNAdb v2 (Quek et al, 2015) and the GENCODE v7 lncRNA gene annotation database (Harrow et al, 2012). In addition, all the transcripts were tested using the Coding Potential Calculator (Kong et al, 2007), which looks for homology to known sequences, open reading frames and a bias towards mutations in the 3rd codon position. The transcripts were also aligned to the genomes of mouse, human, dog, horse and cat as well as the

[11] 75% coverage, 80% identity and an e-value cut off of 5e-9

bats: *E. fuscus*, *M. brandtii*, *M. davidii* and *P. alecto*[12] using Blastn in order to check homology throughout the mammalian kingdom.

dN/dS analysis

Orthologues of the annotated *M. natalensis* genes were obtained from Ensembl's annotated genomes for the following species where available: *Homo sapiens*, *Mus musculus*, *Bos Taurus*, *Canis lupus*, *Felis catus*, *Equus ferus* and *Sus scrofa*. The coding regions were then aligned using Macse (version 1.01b) (Ranwez et al, 2011). The dN/dS value for *M. natalensis* was determined to each of the other available orthologues in a pairwise manner using Codeml, which is part of the Paml package (version 4.7) (Yang, 2007), for each gene that had 300bp or more aligned. A gene was excluded from the analysis if an orthologue was not available for at least 3 of the 7 aligned species. Genes were then analysed as part of a pathway. Pathways of interest were identified and defined using Ingenuity Pathway Analysis (QIAGEN Inc., https://www.qiagenbioinformatics.com/products/ingenuitypathway-analysis) and differences in gene expression in Eckalbar et al (2016). The identified pathways were the Eif2, Wnt PCP, Wnt B-catenin and Fgf pathways.

Results

Sequencing and processing

There were large differences observed in the sequencing quality of the libraries. The 175bp library was of much lower quality than the 300bp and 600bp libraries (Figure 2.3), while the longer insert libraries had high levels of duplicates. This resulted in a lower final coverage after quality trimming and error correction than expected (Figure 2.4). This low coverage contributed to the need for the additional sequencing of the 600bp and mate pair libraries. The second round of sequencing raised the coverage from 49.7x to 77.2x.

[12] *M. lucifugus* and *P. vampyrus* were not used due to their incomplete nature.

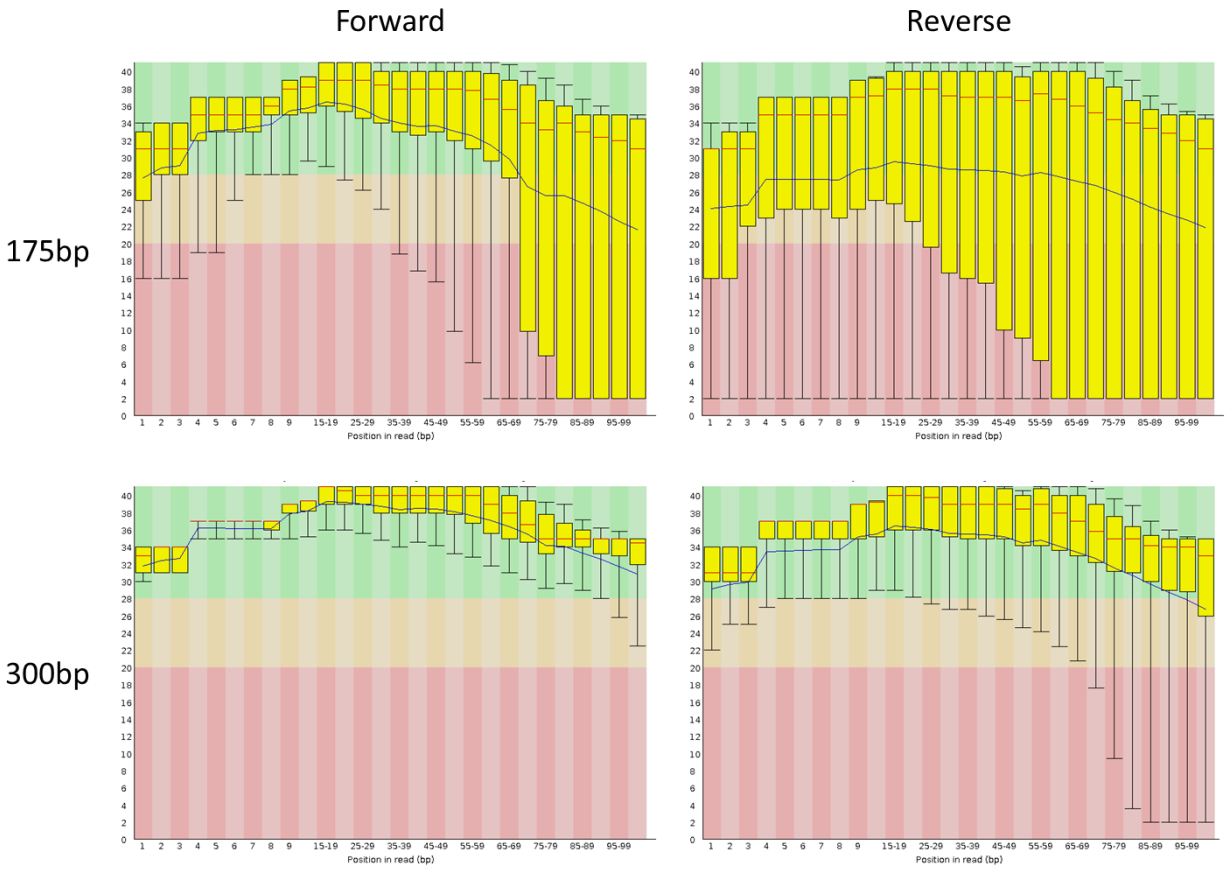


Figure 2.3: Change in the quality score along the raw reads from the 175bp and 300bp insert library. The red line represents the median value, the inner and outer quartiles are shown by the yellow boxes and the 10%/90% values are represented with the error bars.

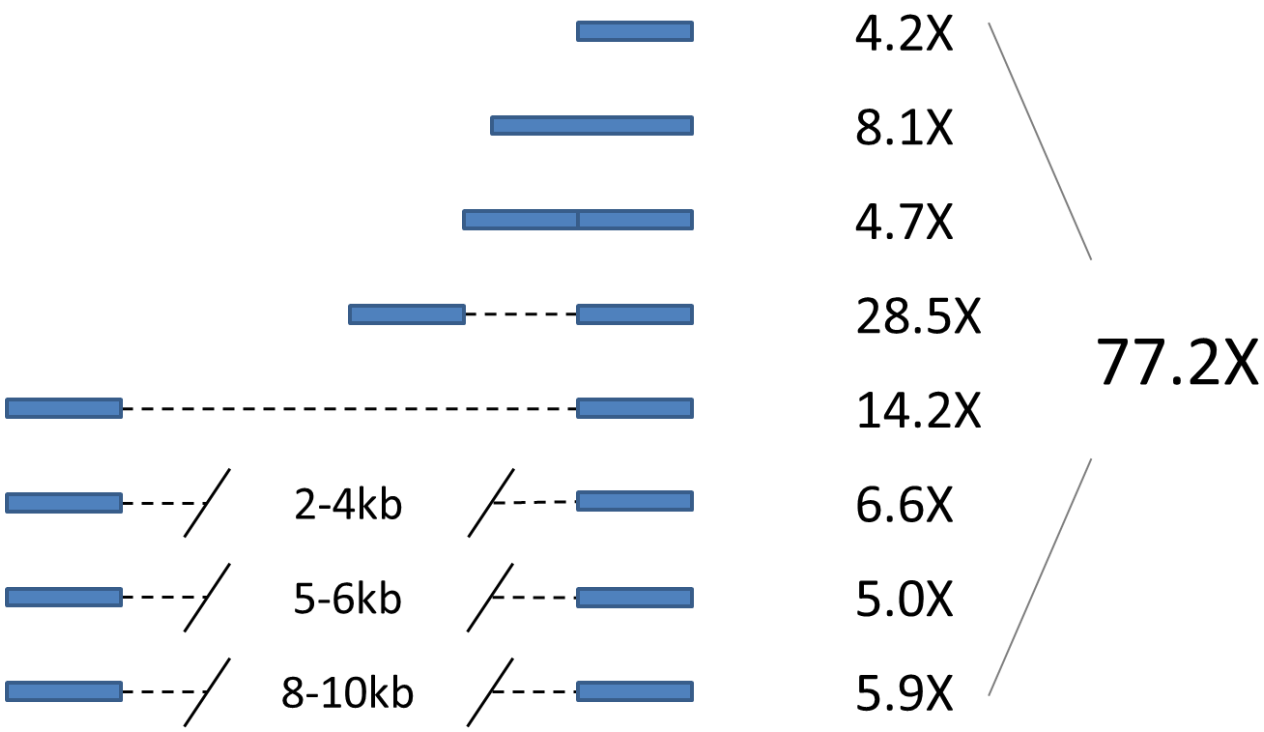


Figure 2.4: Relative coverage of different insert size libraries after processing. Single reads (with a coverage of 4.2x) are the result of one read of a pair being eliminated due to poor quality. Merged reads from the 175bp library had a coverage of 8.1x, while the unmerged reads had a coverage of 4.7x. In total, the coverage amounted to 77.2x.

Genome Assembly and Annotation

Once a number of parameters, including the k-mer value, average insert size, and read processing were optimised, a final high quality genome was assembled (Figure 2.5). Interestingly, the addition of new data, raising the coverage from 49.7x to 77.2x, did not see an immediate improvement in assembly quality, but eventually, through the optimisation process the 77.2x coverage assembly became substantially better than the 49.7x assembly. This shows the importance of good optimisation.

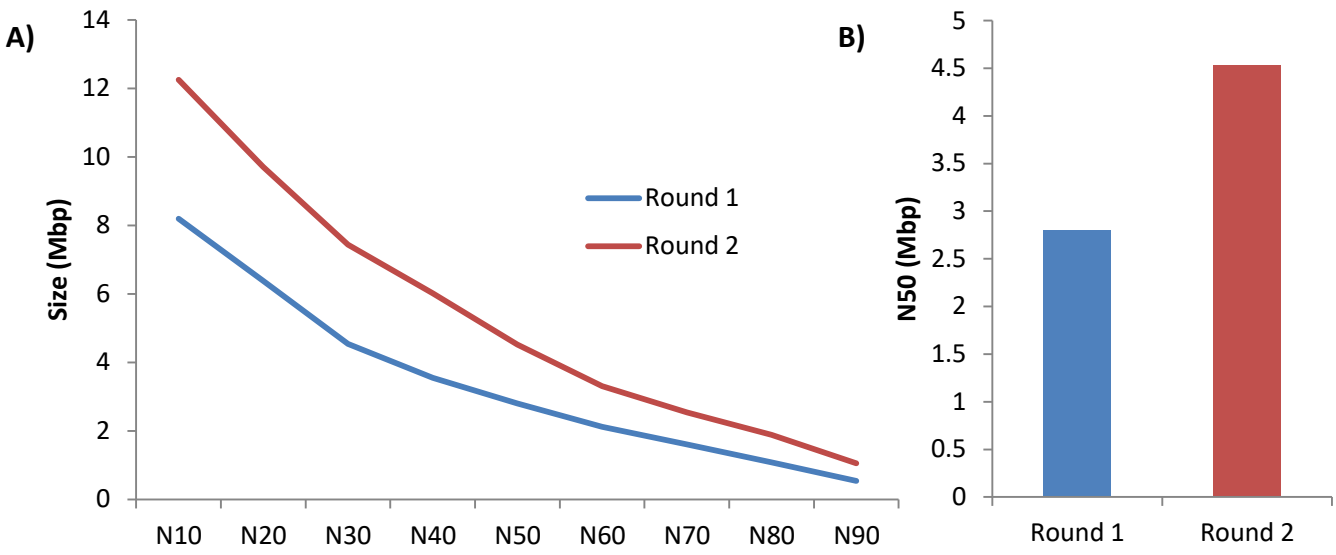


Figure 2.5: Difference in assembly quality before and after additional long range sequencing was added to the *M. natalensis* genome. A) Size of scaffolds across the genome assemblies, with the 1st round of sequencing in blue (49.7x coverage) and the added second round of sequencing in red (77.2x coverage). B) Highlight of the N50 statistic of the two assemblies. The additional sequencing, which was disproportionately longer libraries, increased the N50 size from 2.8Mbp to 4.5Mbp.

Gap closing reduced the N50 value from 4.5 Mbp to 4.2 Mbp and the total assembly length from 2 Gbp to 1.8 Gbp. While this makes the assembly look less impressive on paper, it is deceiving. All of the information originally contained by the assembly is still present after gap closing. The percentage of N's in the assembly dropped from 14.8% to 3.8%. This means that many N's have either been removed because the estimated gap size was too large (171 Mbp), or else converted into nucleotides by the addition of the extra information (53 Mbp). This accounts for the reduction in size in total length and the change in the N50 size.

This reduction in size could be a result of gaps which are over-estimated being easier to discover than gaps which are underestimated (Figure 2.6). It could also be a result of slightly longer insert size estimates being able to assemble easier. Or else it could be some combination of both of these options.

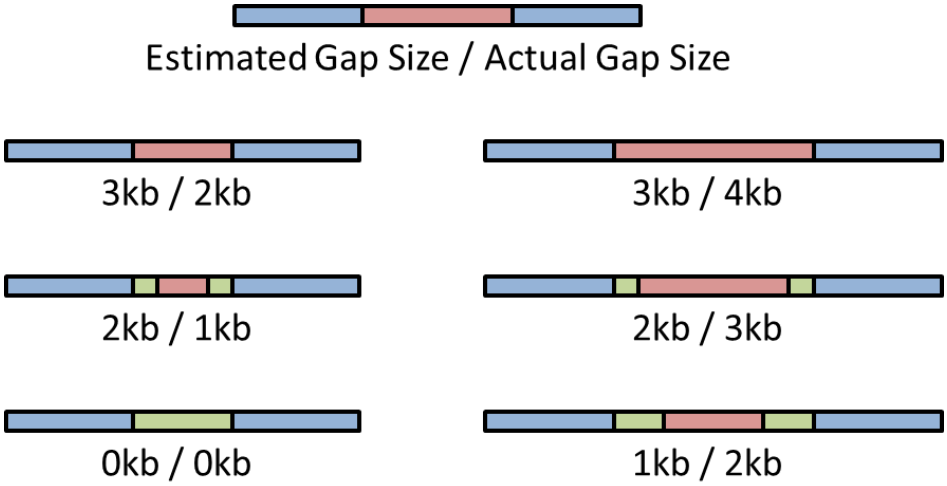


Figure 2.6: Gap closing poorly estimated gaps between contigs. If a gap is smaller than expected (left), the gap is more likely to be closed in the gap closing process and the overestimate discovered. This reduces the length of the scaffold. If a gap is larger than expected, the underestimate is not discovered and the scaffold length remains unchanged.

Finally, the assembly's quality was quantified using Cegma (Parra et al, 2007). The final genome statistics can be seen in Table 2.2. This assembly was then compared to the *M. davidii*, *M. brandtii* and *P. alecto* genome assemblies due to their similar sequencing coverage and use of Illumina sequencing. The results of this comparison are displayed in Figure 2.7.

Table 2.2: The final *M. natalensis* genome assembly statistics.

Total Scaffold Length	1.8 Gbp
Longest Scaffold	32.1 Mbp
Scaffold N50	4.2 Mbp
Scaffold NG50	3.6 Mbp
Contig N50	29.7 kbp
Contig NG50	25.6 kbp
%N	3.7%
CEGMA partial	96.0%
CEGMA complete	92.7%
Repetitive Elements	33%
Heterozygosity	0.13%
No. Genes	24 239

The comparison with *M. davidii*, *M. brandtii* and *P. alecto* showed that despite lower coverage in the mate pair reads, the *M. natalensis* assembly was comparable to the two Myotis assemblies. The *P. alecto* assembly appears to be of significantly higher quality however. This might demonstrate a phylogenetic difference between the two clades in the ease of assembly within the Chiroptera order, such as lack of transposable element activity, for example (Cantrell et al, 2008). This is supported by the fact that the *P. alecto* assembly was done by the same group that produced the *M. davidii* assembly.

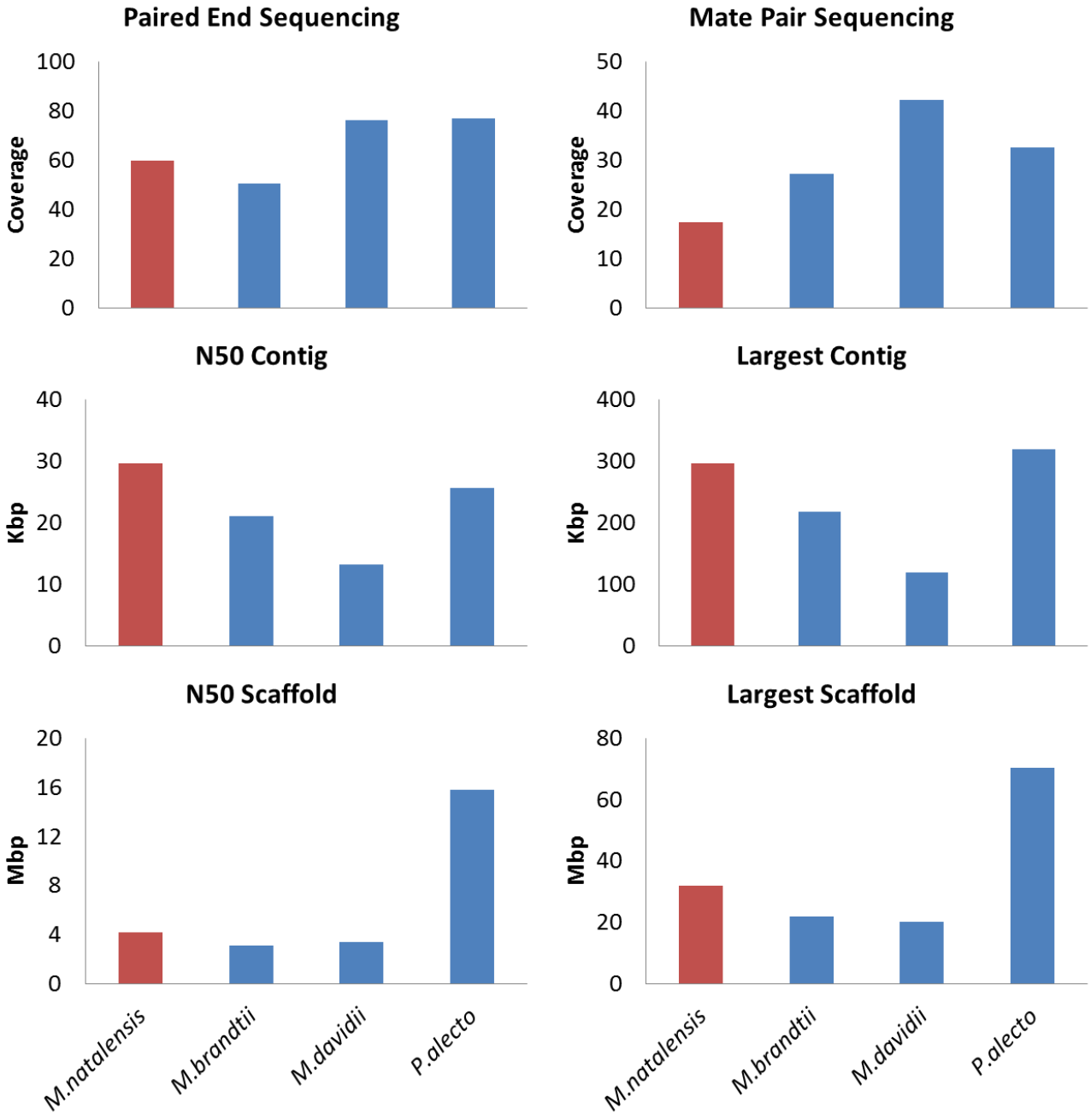


Figure 2.7: Comparison of the *M. natalensis* assembly to the *M. brandtii*, *M. davidii* and *P. alecto* genome assemblies.

The annotation process resulted in 24 239 genes being annotated, with a total of 984 766 exons. This is well within the expected range for a mammalian genome. RepeatMasker worked in a similarly promising fashion, masking 33% of the genome as repeat elements; which is consistent with other bat genomes (Figure 2.8).

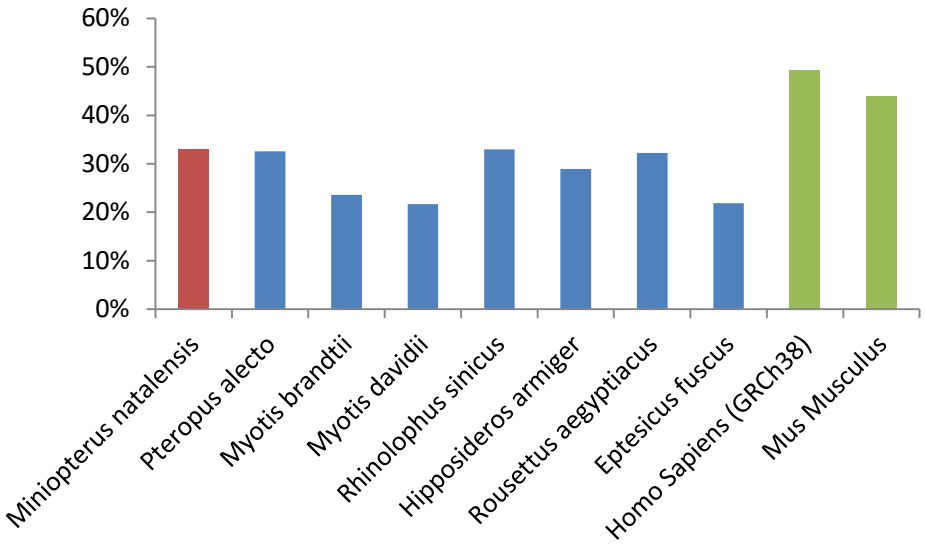


Figure 2.8: Comparison of repetitive element content in bat genomes. The percentage of repetitive elements for several bat genomes (blue) was obtained from the NCBI annotation reports and compared to the Repeat Masker results for M. natalensis (red). Human and mouse genomes were included for comparison (green).

Lnc RNA analysis

Of the annotated genes, 227 had no similarity to known protein sequences. These genes also had low Coding Potential Scores, as measured by CPC (Kong et al, 2007), suggesting the pipeline has successfully identified lncRNAs (Figure 2.9). Of the 227 putative lncRNAs, 12 matched known lncRNAs. In contrast, 34 genes appeared to be unique to *M. natalensis*, having no conservation to any of the tested genomes (Supplementary Table 2.1).

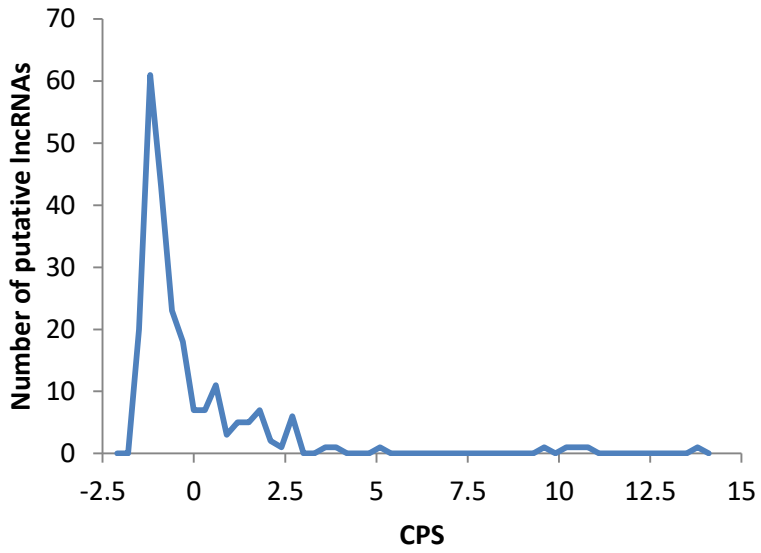


Figure 2.9: Distribution of Coding Potential Scores for identified putative lncRNAs. The higher the Coding Potential Score (CPS), the more likely a sequence is to be a protein coding gene, as opposed to a lncRNA.

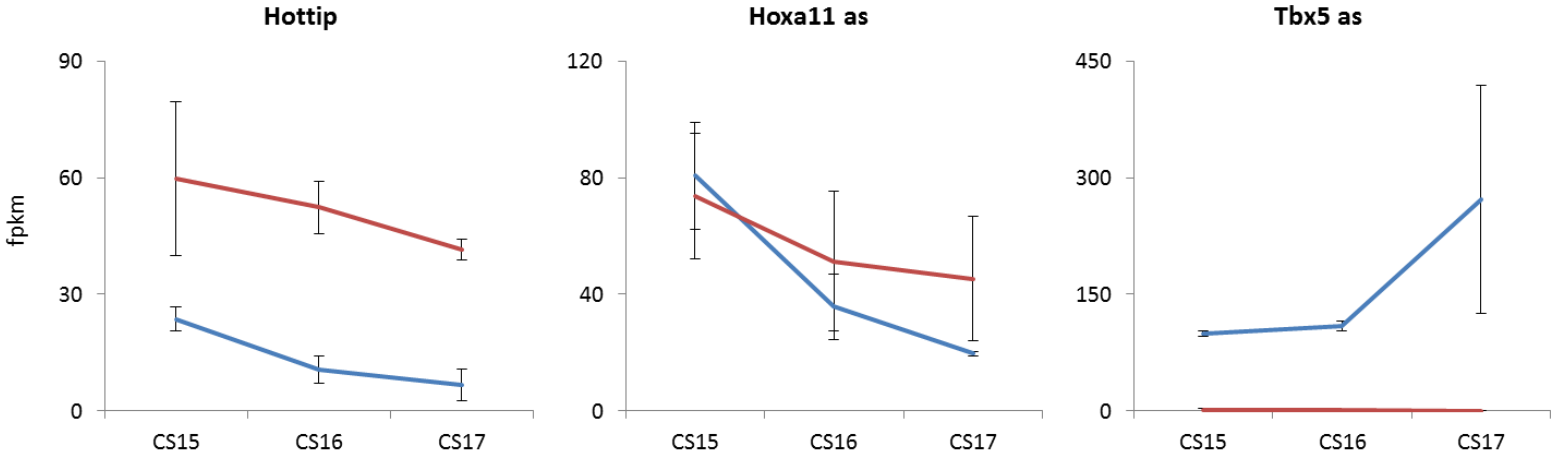


Figure 2.10: Expression of select lncRNAs in developing M. natalensis limbs. The fpkm (fragments per kilobase of transcript per million mapped reads) values for RNA-seq data from Eckalbar et al (2016). The forelimb (blue) and hindlimb (red) are shown for 3 stages of embryonic development, equivalent to mouse E12.0, E12.5 and E13.0 (Hockman et al, 2009).

Of notable interest, HoxA11as and Hottip were differentially expressed (Eckalbar et al, 2016) between forelimb and hindlimb (Figure 2.10). Additionally, of the 227 putative lncRNAs, 5 were found to be absent in the other mammalian genomes while being conserved within the Chiroptera genomes. Of these, 4 were differentially expressed (Figure 2.11), although none were near a gene of particular interest. One final gene of interest was found to be the antisense of Tbx5, a crucial gene in limb development (Figure 2.10). The results from this analysis are displayed in Supplementary Table 2.1.

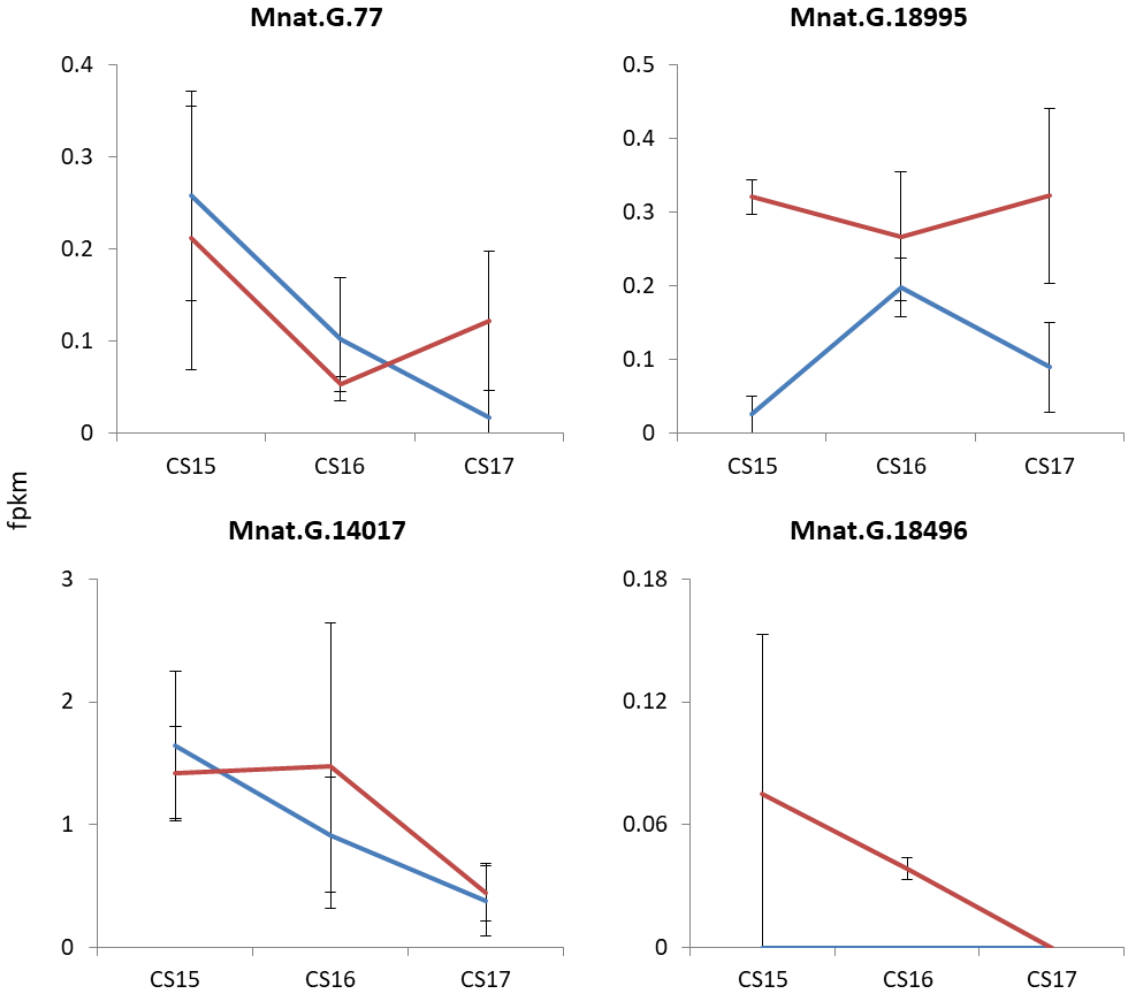


Figure 2.11: Expression of differentially expressed lncRNA unique to the Chiroptera order. The fpkm values for the forelimb (blue) and hindlimb (red) are shown for 3 stages of embryonic development. Gene names refer to the unique gene identification numbers.

dN/dS Analysis

In order to identify signals of positive selection, 1:1 homologues were aligned to the *M. natalensis* genes from 7 terrestrial mammal species. This analysis aimed to take a broad approach, analysing as many genes as possible, and therefore opted to not exclude genes that were missing transcripts from some of the species.

The disadvantage of this lenient strategy however, is that because each gene has different species in the alignment, each gene effectively requires its own phylogeny. This could potentially make the analysis quite complicated. But the variance in the divergence time between the species is not large. The relevant mammalian species all diverge from the Chiroptera order at approximately 80-90 million years ago (Figure 2.12). This means that the branch lengths of the different comparisons are about the same length (in years). By comparing the species in a pairwise manner instead of within the phylogeny, a small amount of error is introduced, but the analysis is greatly simplified which allows for otherwise problematic genes to be included.

By the end of the pipeline, there were 11 033 genes with at least 3 of the 7 species aligned to the relevant *M. natalensis* gene. The majority of these genes (75.5%) had all 7 of the selected homologues aligned to them. Among the 11 033 genes, 2 were obviously erroneous outliers, with dN/dS values of 16.8 and 67.0. All of the other genes were distributed between 0.0 and 1.3 (Figure 2.13), with an average 0.14.

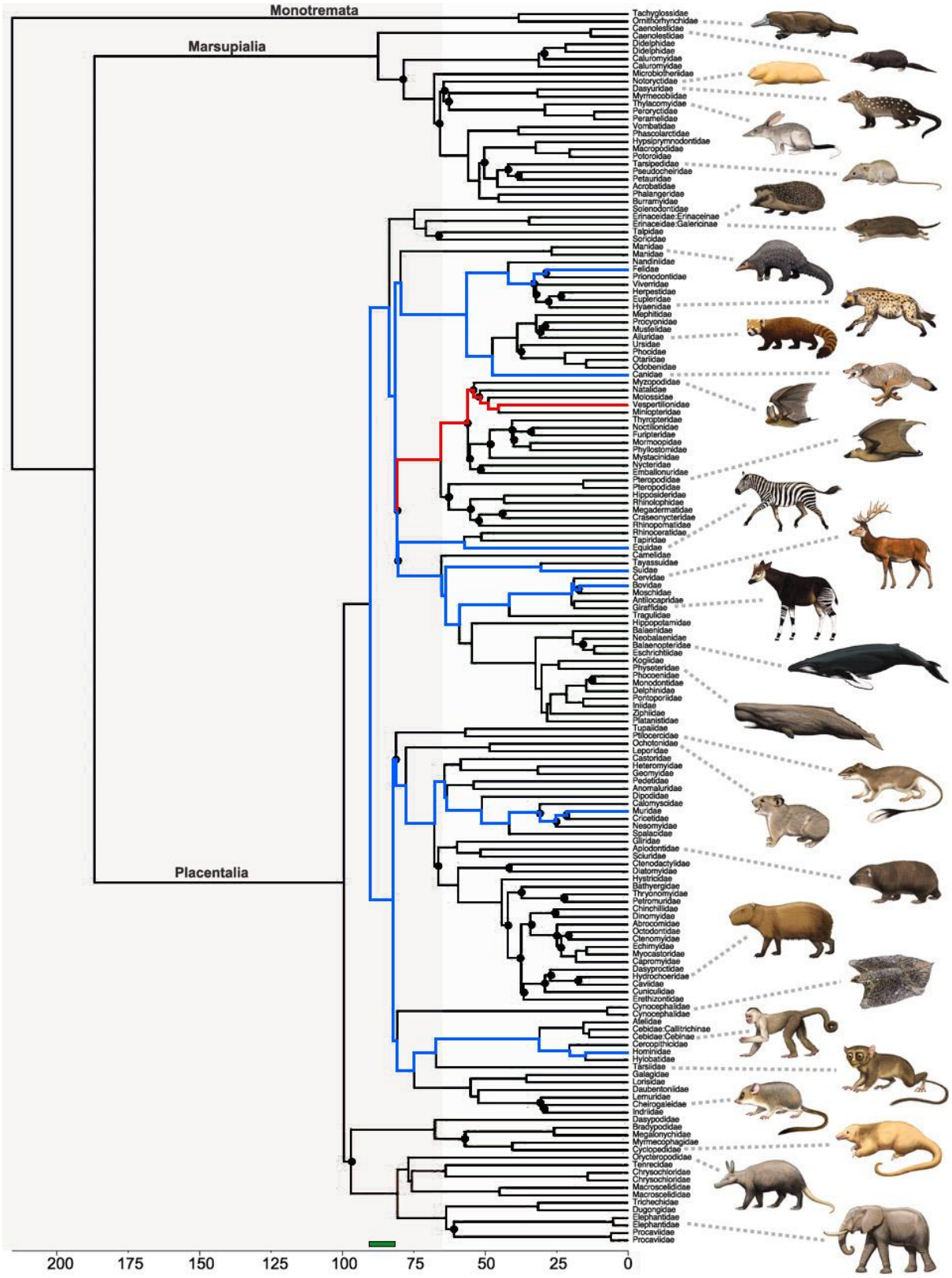


Figure 2.12: Mammalian phylogeny adapted from Meredith et al (2011). The blue lines represent branches leading to species used in the dN/dS analysis. The red line represents branch leading to *M. natalensis*. The x-axis denotes millions of years of divergence. The green bar above the x-axis shows the difference in the time of divergence between the most recent and most divergent taxa in the dN/dS analysis to *M. natalensis*.

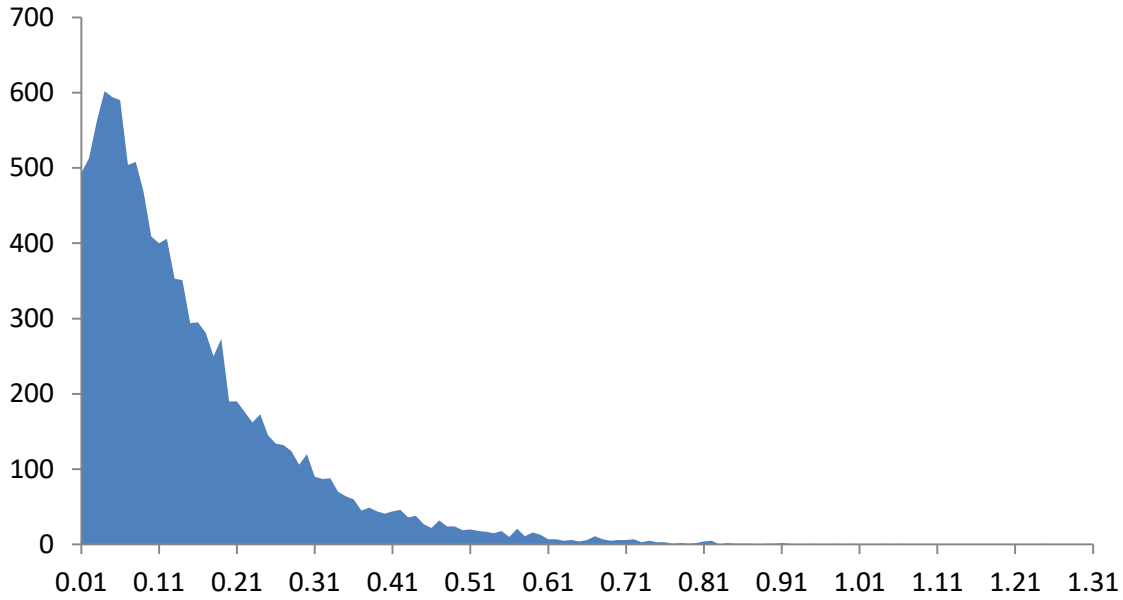


Figure 2.13: Distribution of dN/dS values for 11 031 genes. The distribution is heavily skewed to the right relative to the mean value (0.14).

Unfortunately, due to the nature of the analysis, any errors resulting from sequencing, homologue identification or misalignment are all likely to increase the dN/dS value (Table 2.3). This is a problem since these errors make genes appear more interesting, by generating a higher dN/dS value. Therefore the genes with the most extreme dN/dS values are likely to be populated by errors. Sometimes the gene might have an obvious error once investigated, but other genes might have subtler issues which are harder to detect by simple inspection. In order to minimise the error created at each stage of this pipeline, genes were analysed as part of a pathway. By viewing the dataset in this manner, the impact of any error is greatly mitigated.

Pathways of interest and their associated genes were identified using an Ingenuity Pathway Analysis (QIAGEN Inc., https://www.qiagenbioinformatics.com/products/ingenuitypathway-analysis) and the differential gene expression between developing forelimbs and hindlimbs in *M. natalensis* (see Eckalbar et al, 2016, for details). All of these pathways were composed of genes which had average dN/dS values lower than the global gene average (Figure 2.14). This does make sense, since these pathways are made up of important developmental genes, which have been shown to have highly conserved functions (Carroll, 2008).

Table 2.3: Sources of error in the dN/dS pipeline.

Pipeline	Source of Error
Sequencing	Incorrect Base Calls
Assembly	Structural Error
Annotation	Poor definition of exon boundaries can lead to frame shifted alignments
Homologue and Transcript Identification	Incorrectly identifying a paralogue instead of an orthologue for comparison will inflate the divergence time and the dNdS results. This can also have a smaller and subtler role in differences in splicing between identified Orthologues.
Alignment	Alignment error
Branch length approximation	The divergence between two species affects the expected number substitutions. Therefore, differences in divergence should be taken into account, which it was not.

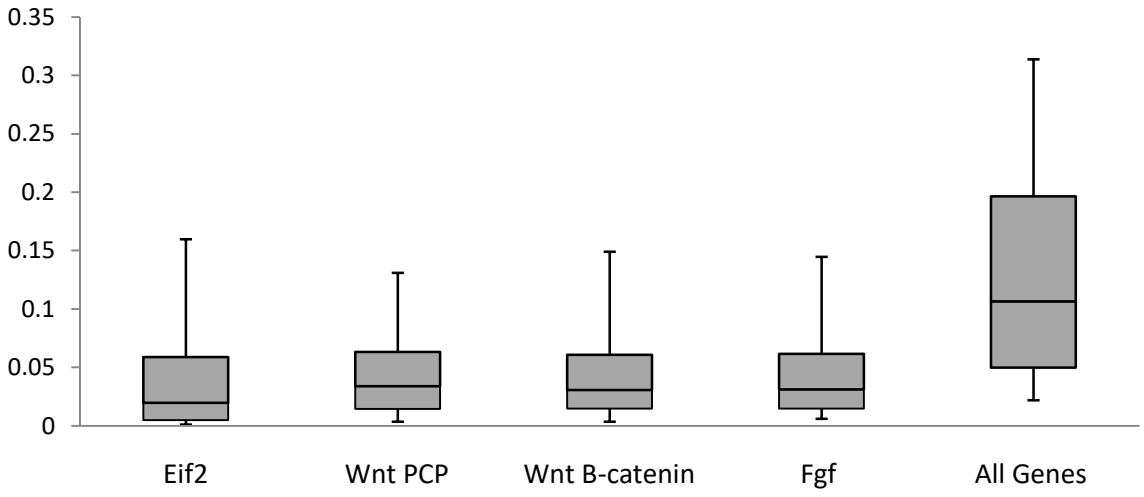


Figure 2.14: Box plot of the dN/dS values of upregulated pathways in *M. natalensis*. Each box plot represents the 90th percentile, the upper quartile, median, lower quartile and 10th percentile.

Discussion

The goal for this section was to identify the genetic mechanisms that have led to the evolution of the bat wing. In attempting to achieve this aim, a genomic resource was developed using *M. natalensis.* This resource took the form of an annotated genome assembly and formed the base for an extensive study that used RNAseq and ChIPseq to try and identify genes and pathways of interest (Eckalbar et al, 2016). To compliment these analyses, a Dn/Ds and lncRNA analysis was performed using the predicted annotated genes.

Assembly and Annotation of the Genome

Optimisation of the genome assembly was a time consuming, but ultimately successful process. Substantial gains in the quality of the genome assembly were achieved, resulting in a N50 size (4.2 Mbp) comparable to other published genomes, despite lower coverage of mate pair sequencing. Overall, this assembly process gave a firm grounding of what to expect and aim for with future assemblies.

The annotation process was less obvious to optimise. Unlike the assembly process, which has several key metrics to use for comparison and optimisation, it is not clear which statistics are better or worse between different annotation versions. This meant the annotation process was more a matter of providing as much good data as possible for Maker to use and, assuming the output numbers seem reasonable, trusting in Maker to be effective.

Applications of the Assembled Genome

The primary function for the annotated genome was to provide a base for alignment of RNA-seq and ChIP-seq data from the developing limbs (Eckalbar et al, 2016). Of the 24 239 annotated genes, 7 172 were found to be differentially expressed at some point in the developmental process (either between stages or between limbs). Of these, 2 952 genes were found to be differentially expressed between the forelimb and hindlimb. These included expected genes such as *Hoxd10* and *11,* and *Tbx4* and *5,* as well as several uncharacterised genes (see Eckalbar et al, 2016). The differentially expressed genes were further categorised into pathways and checked for consistency (i.e. are the

genes within a pathway consistently up-regulated or down-regulated). This method highlighted pathways such as Fgf and Wnt-PCP and was later used for the dN/dS analysis.

ChIP-seq was used to highlight open and closed chromatin using H3K27 acetylation and tri-methylation respectively. This highlighted 2 475 regions which showed differential enrichment in both markers between the forelimb and hindlimb. Regions highlighted by H3K27 acetylation were also used to try and find enhancers that have undergone positive selection in the bat lineage (termed BARs or Bat Accelerated Regions; see Eckalbar et al, 2016).

In addition to the RNA-seq and ChIP-seq data, which the genome was assembled to support, the genome has been independently used to study transposable elements in Vesper bats (Platt et al, 2016).

lncRNA Identification

The lncRNA identification pipeline appears to have worked, disproportionately identifying genes with a low potential for coding activity (Figure 2.9). In addition, several known lncRNAs were identified (Figure 2.10). Some of these long non-coding genes, such as Hottip (Wang et al, 2011) and Tbx5as are associated with genes known to be important in limb development (HoxA13 and Tbx5).

Unfortunately, the search for conserved lncRNAs within the Chiroptera order was less convincing. All of the differentially expressed lncRNAs conserved among the Chiroptera order were expressed at very low levels (Figure 2.11). This is not what you would expect if these genes were important in the process of bat wing development.

dN/dS Analysis

The general trend of the dN/dS analysis, for genes involved in the limb developmental pathways to have low dN/dS values, should have been expected. A dN/dS analysis can be used to identify coding sequences that have been under positive selection. It sounds like it should be a good way to find the genes responsible for the extreme phenotype observed in the bat forelimb. But there are several reasons why this might not be the case. Firstly, the Chiroptera lineage is an old one (Figure 2.1). This means that any positive selection involved in the initial adaptation of the bat wing has been followed by at least 60 million years of stabilising selection. The bat lineage is too old, and the trait of interest

has been too stable, for something like a simple dN/dS analysis to be too effective. This can be seen in the consistently low dN/dS values of the genes in the pathways of interest, compared to the rest of the genes (Figure 2.14). In hindsight, a branch specific test needed to be done, looking for selection on the phylogenetic branch before the radiation of the Chiroptera order. But this would have been limited by the lack of annotated bat genomes available and complicated by the variable pool of species used for each gene.

In addition, most of the genes involved in the formation of the forelimb are also involved in hindlimb development. This means that any protein coding changes which would drive a limb closer to the wing morphology would need to be cancelled out in the hindlimb. This is not a parsimonious path of adaptation. This problem is compounded by the fact that many key transcription factors in limb development are also involved in the development of other important structures, such as the nervous system. As a result of this tendency for important transcription factors to be multi-purpose, it is more common for morphological change to result from a change in promoter sequences rather than the protein sequence. In this way, a gene can gain specific functions without compromising its ancestral role (Carroll, 2008).

Theoretical limitations of a dN/dS analysis aside, the analysis also proved to be highly susceptible to an accumulation of errors. Any problems with a gene's alignment, resulting from sequencing errors, alternative splicing, gene paralogues, gaps in the sequencing, etc, result in high dN/dS values. This casts doubt on other high dN/dS values, even if they are not identifiably wrong. This is a strong argument against the strategy employed, attempting to keep the dataset as large as possible, as it required lenient filtering of genes. That said, viewing genes as being part of a pathway, rather than an individual gene, proved effective in dissipating error and made the method a lot more robust. This can be seen in the consistency of the dN/dS values of the various pathways (Figure 2.14).

Everything considered, a dN/dS analysis was probably not the right approach for finding genetic causes of wing development in bats. A better approach would be to look for positive selection in the non-coding regions, where it is less restricted by alternative protein functions. A method along these lines using H3K27Ac ChIP-seq peaks to identify potential enhancers was employed in Eckalbar et al (2016).

Otherwise, another method employed by Seim et al (2013), which looked for conserved amino acids in bats which differed from conserved amino acids in other mammals would probably also be better. This accounts for the genes in question potentially having low dN/dS values and also allows for a much smaller, local alignment. This means whole transcripts would not need to be aligned to whole

transcripts, significantly reducing some of the problems encountered. The area of the local environment can be restricted to otherwise highly conserved regions, increasing the faith in the alignments as a whole. And finally, the short sequence could be extracted directly from the genomes of the mammals relatively easily, meaning an annotated genome wouldn't be necessary, which would vastly improve the sample size of species used.

Conclusion

The assembly of the *M. natalensis* genome has formed the base for a multi-faceted analysis. This aimed to identify the genetic mechanisms responsible for the evolution of flight in bats. And while a concrete answer isn't available yet, the ground work for making *M. natalensis* a staple of future studies has been laid. The combination of a well annotated, high quality genome, with RNA-seq and ChIP-seq from the developing limb tissue make it the most comprehensive bat resource available to date.

Chapter 3: *Papio cynocephalus* – Primate Hybridisation

Papio cynocephalus, or the Yellow Baboon, are from the Cercopithecidae family, which diverges just before the emergence of the Hominoidea clade in Primates. Common throughout East Africa, *P. cynocephalus*' range partially overlaps with *Papio anubis*, the Olive Baboon (Figure 3.1). Within these overlapping regions, hybridisation between the two species has been observed (Alberts and Altman, 2001). This is interesting, not only because of the potential for horizontal gene transfer and the selective pressures acting on the populations, but also because of our own history of hybridisation.

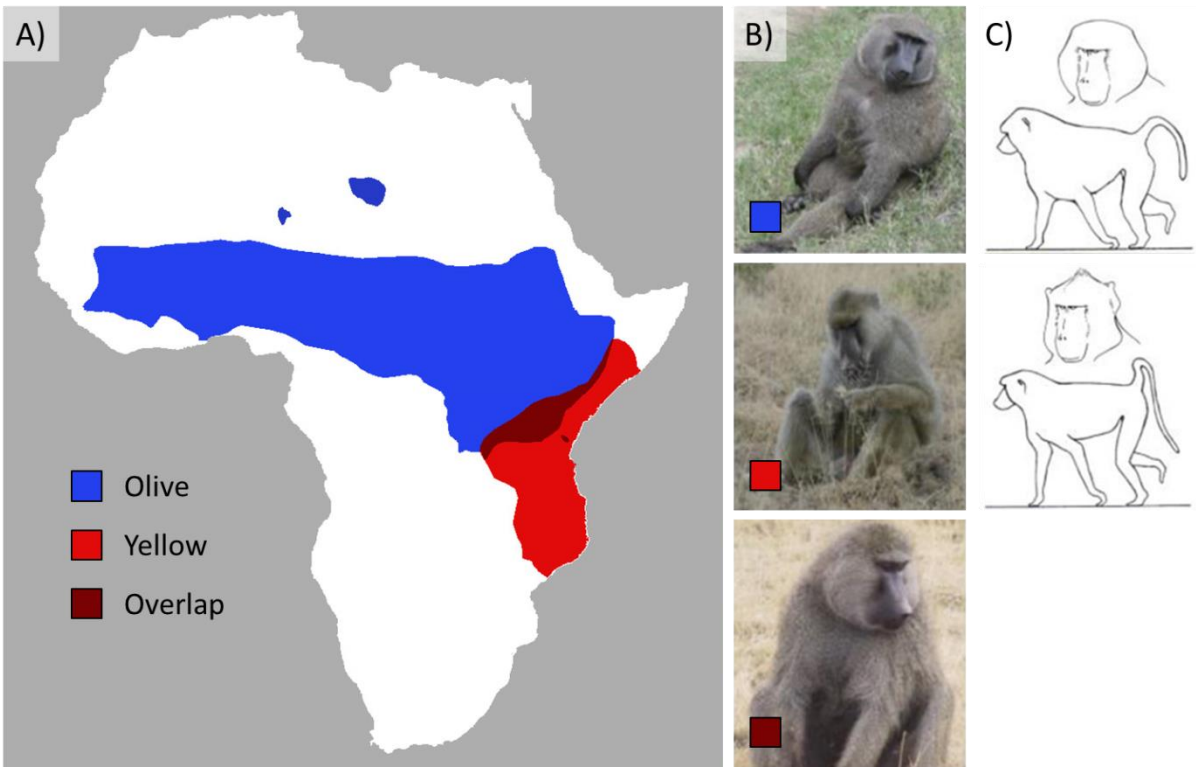


Figure 3.1: An introduction to *P. cynocephalus* and *P. anubis*. A) The distributions of *P. anubis* (the Olive baboon) in blue and *P. cynocephalus* (the Yellow baboon) in red (Kingdon et al, 2008; Kingdon et al, 2016). B) Photos of *P. anubis, P. cynocephalus* and a hybrid individual from an area of overlap (Wall et al, 2016). C) Diagrams highlighting morphological differences between *P. anubis* and *P. cynocephalus* (Wall et al, 2016).

Hybridisation in Primates

There is evidence that hybridisation in primates is surprisingly common (Arnold and Meyer, 2006; Zinner et al, 2011). This includes the notable example of past hybridisation in our own lineage (Sankararaman et al. 2012; Wall et al. 2009). When these hybridisation events result in fertile offspring, it allows for admixture between the two parental populations. This can be a source for new genetic diversity within a population and can be identified after the event if unadmixed samples are also available for sequencing.

The Amboseli Population

The Amboseli Baboon Research Project in Kenya has been observing *P. cynocephalus* troops in Amboseli National Park since 1971 (Alberts and Altman, 2012). In 1983 a *P. anubis* population entered the area, allowing for potential admixture between the two populations. Hybrid individuals are not just fertile (Alberts and Altman, 2001), but potentially even display selective advantages over their *P. cynocephalus* kin. These advantages include reaching sexual maturity quicker and being more appealing to potential mates (Charpentier et al, 2008; Tung et al, 2012). This suggests that gene transfer is actively occurring between the two populations.

Gene flow between the two species could have long term consequences for these populations. But getting an accurate picture of what is occurring at the genetic level is tricky. The animals are wild, so invasive methods for acquiring DNA, such as drawing blood are not feasible. And less invasive techniques tend to result in lower amounts of DNA and less coverage for each individual.

The Project

The aim of this project is to investigate the extent of the admixture in the Amboseli population using modern NGS tools. In order to accomplish this, a collaboration was formed, with the purpose of: acquiring and sequencing genetic samples from multiple individuals from inside and outside the hybrid zone at Amboseli park, including a high depth sample, assemble the high depth sample into a genome assembly that can be used for effective mapping of reads from all the individuals, identifying genetic markers that inform the species of an individual using the populations from

outside the hybrid zone and estimating admixture of individuals inside the hybrid zone using said genetic markers.

Methods

In order to investigate the admixture in the Amboseli population, a genome assembly needs to be completed. Reads need to be aligned and SNPs identified. And a method needs to be developed that effectively filters for informative SNPs that predict ancestry.

This work was done in collaboration with other research groups. As such, there were several aspects of the work that I was not responsible. This included: obtaining samples, DNA extractions, read mapping and SNP calling. These methods are described in Snyder-Mackler et al (2016) and Wall et al (2016).

Read Processing and Genome Assembly

The genome assembly used Illumina reads from a *P. cynocephalus* individual from the Southwest National Primate Research Center in San Antonio, Texas. It was sequenced to moderate depth (46.8x) using seven different insert length libraries (175bp, 400bp, 3 000bp, 4 300bp, 5 800bp, 10 000bp). Read quality was assessed using Fastqc (version 0.10.1) (Andrews, 2010). Bases of quality 17 or less were trimmed from the reads using Trimmomatic (version 0.32) (Bolger et al, 2014). The two paired end libraries (insert size 175bp and 400bp) were used to establish the known 27bp k-mers of the genome using KmerFreq_HA (version 2.01) (SOAPdenovo package). All libraries were then error corrected using these k-mer frequencies and Corrector_HA before having duplicate reads removed with FastUniq (version 1.1) (Xu et al, 2012).

Genome assembly was done using SOAPdenovo2 (version 2.04, k=45) (Luo et al, 2012). Optimisation of the genome assembly used the k-mer size for contig building and scaffolding but primarily involved changing the estimated size and order of use of the various read libraries. The assembly with the best N50 value was improved using GapCloser (version 1.12) (SOAPdenovo package) before having its quality assessed with Cegma (version 2.4) (Parra et al, 2007). Scaffolds of size 500bp or less were excluded to make downstream analysis faster and easier. The genome assembly is available at: https://abrp-genomics.biology.duke.edu/index.php?title=Other-downloads/Pcyn1.0

Sequencing of Amboseli Baboons

In total, 23 baboons from the Amboseli park were sequenced. One of these individuals was sequenced to moderate coverage (19.6x), while the other 22 baboons were sequenced to low coverage (2.1x on average). These individuals included 9 suspected hybrid individuals and 11 suspected unadmixed *P. cynocephalus* individuals (including the higher coverage individual). The raw data used is available on NCBI (accession code PRJNA308870).

Species Training Sets

In order to quantify ancestry in the Amboseli population, it is necessary to determine what is typical of *P. cynocephalus* and *P. anubis.* Low coverage (1.1x on average, after duplicates were removed) sequencing from 9 *P. cynocephalus* baboons from Mikumi National Park in Tanzania was generated in conjunction with the sequencing data that was used for the genome assembly. These baboons are believed to be unadmixed and were used as the training set to identify SNPs that would predict *P. cynocephalus* ancestry. These data were compared to low coverage (2.1x on average, after duplicates were removed) sequencing from 13 *P. anubis* baboons (6 from the Washington National Primate Research Center and 7 from the Maasai Mara National Reserve in Kenya). All sequencing data is available on NCBI (accession code PRJNA308870). Finally, moderate coverage (21x) sequencing from a *P. anubis* baboon was also used (available from NCBI SRR927653-SRR927659).

Determining Genotype Frequency

Each individual has a "Phred-scaled genotype likelihood" (PL) score for each SNP, generated by GATK (McKenna et al, 2010), which predicts the relative likelihood of each possible genotype, as a ratio (Figure 3.2). These scores can be used to calculate the expected allele frequency for each individual by first calculating the probabilities of each possible genotype.

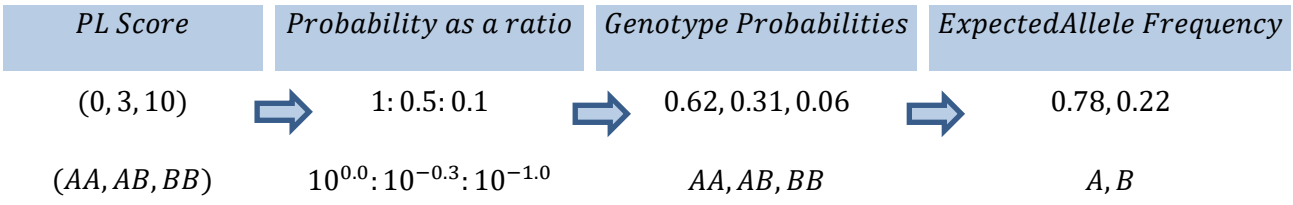


Figure 3.2: Transformation of PL Scores to Expected Allele Frequencies. A ratio of probabilities can be calculated from PL scores, which can in turn be turned into an expected allele frequency.

Identifying Species Specific SNPs

The identified SNP variants (Wall et al, 2016) were filtered to only include biallelic SNPs on scaffolds of 1000bp or more. The average allele frequency for each SNP was compared between the *P. cynocephalus* baboons from outside of the Amboseli park with the sequenced *P. anubis* baboons. SNPs were excluded if they didn't have at least 3 individuals from each species with sequencing data for that position. If a variant position had a difference in allele frequency of 0.8 or more between the two species, it was classified as being predictive of species ancestry, and used in the subsequent classification of Amboseli baboons.

Individuals used in identifying predictive SNPs were then systematically excluded from the analysis. Allele frequencies were re-calculated and predictive SNPs re-identified. These SNPs were then used to test the consistency of the Amboseli baboon classification as well as estimate the ancestry of the excluded individual.

Classification of Amboseli Baboons

For each of the Amboseli individuals, the genotype frequencies at the previously identified predictive sites were classified as either being closer to *P. cynocephalus*, *P. anubis* or a heterozygous 50:50 allele frequency using a difference of squares test. It is worth noting that, due to the low coverage, mischaracterisation of any particular SNP will be a common occurrence. For example, a heterozygous site sequenced at 2x coverage has a good chance of having 2 copies of the same allele, just by chance. And if a heterozygous site only has 1x coverage, instead of being identified as

heterozygous, it will have a 50% chance of appearing as a homozygous site of either allele in the dataset. The confidence for any one nucleotide position in the dataset will therefore be low and will need to be used collectively in order to compensate for this. This also means that unadmixed individuals won't appear to be 100% *P. cynocephalus* or *P. anubis*. But the estimates of the baboons' ancestry from outside the Amboseli park should give a better idea of what percentages to expect from unadmixed individuals.

Disproportionate SNP Frequency in Amboseli

The Amboseli baboon SNPs were tested to see if any SNPs are consistently associated with either *P. cynocephalus* or *P. anubis*. This was done by classifying SNPs as either being closest to *P. cynocephalus*, *P. anubis* or heterozygous frequencies. At each site, the probability of that number of species specific SNPs occurring by chance was calculated. This was done by using the product of the relevant individuals' expected genetic makeup (from the previous section) and accounting for the different combinations these positive results could have occurred in (i.e. differences in the order of individuals with no sequence coverage at that site). Finally, a Bonferroni correction was applied to solve the issue of multiple testing. Therefore the final formula for the adjusted probability (Padj) was:

$$P_{adj} = \prod_{i=0}^{k} P(anubis/cynocephalus)_i \times \frac{n!}{(n-k)!\,k!} \times No.of\ SNPs$$

Where 'k' is equal to the number of individuals with a *P. anubis*/*P. cynocephalus* classified SNP at that site and 'P(anubis/cynocephalus)' is the expected probability for a SNP from that individual being from the relevant species. The 'n' refers to the total number of Amboseli individuals. Therefore, '(n-k)' is the number of individuals with no sequence coverage at that site.

Results

Read Processing and Genome Assembly

The *P. cynocephalus* genome was sequenced with multiple libraries across varying size ranges. This was meant to be about 60x coverage, but as a result of the sequencing quality and PCR duplicates in

the sequencing, the coverage decreased significantly (Table 3.1). Combined these effects reduced the expected coverage of the genome by approximately 16x (from 63x to 47x).

After the reads were assembled with SOAPdenovo (Luo et al, 2012), there were over 33 thousand scaffolds that contained just under 220 thousand contigs. The total assembly length was 3.1 Gbp, with a N50 scaffold size of 887 kbp (Table 3.2 for other statistics). Furthermore, the assembly coherency appeared good, with Cegma (Parra et al, 2007) finding 85% of the 248 conserved Eukaryotic genes in their entirety, while 95% were found in at least a partial form.

Table 3.1: Summary of sequencing data used for genome assembly. Table adapted from Wall et al (2014). Single ended reads (SE) represent read pairs that had one read removed in quality control.

Insert Size Used (bp)	Raw Reads (10^8 pairs)	Processed Reads (10^8 pairs)	Proportion Unique	Coverage
0 (SE)	0	0.25		0.8
175	3.4	3.2	0.91	17.8
400	3.5	3.4	0.93	19
3000	0.45	0.31	0.83	1.6
4300	1.4	0.74	0.85	3.8
5800	0.19	0.14	0.19	0.2
10000	1.2	0.79	0.72	3.4
14000	0.32	0.23	0.17	0.2
			Grand Total	46.8

Table 3.2: Summary statistics of *P. cynocephalus* genome assembly. A 46.8x coverage of short reads, across multiple libraries, was used to assemble the genome.

Metric	Result
Assembly length	3.09 Gbp
Scaffold N50	887 kbp
Contig N50	28.9 kbp
Amount of 'N' nucleotides	6.57%
Partial CEGMA genes	95%
Complete CEGMA genes	85%

Identifying Species Specific SNPs

Identified SNPs were classified as being predictive of species if the difference in their estimated allele frequency between the baboons from outside of Amboseli National Park was 0.8 or more. These datasets were used to establish the baseline for what to expect for unadmixed individuals (Figure 3.3).

Classification of Amboseli Baboons

SNPs that were found to be predictive of species ancestry were used to classify the baboons from the Amboseli population. These results largely supported the *a priori* estimates for the individuals, with individuals that were estimated to be unadmixed having less *P. anubis* alleles than those that were predicted to be 3/4 *P. cynocephalus*, which in turn had less than the 1/2 *P. cynocephalus* individuals (Figure 3.4). The difference between the unadmixed Amboseli baboons and the 3/4 *P. cynocephalus* was significant (p=0.0016; Mann-Whitney U one tailed test). While there weren't enough samples to statistically test the prevalence of *P. anubis* SNPs in the 1/2 *P. cynocephalus* individuals, both had more *P. anubis* SNPs than the highest 3/4 *P. cynocephalus* individuals.

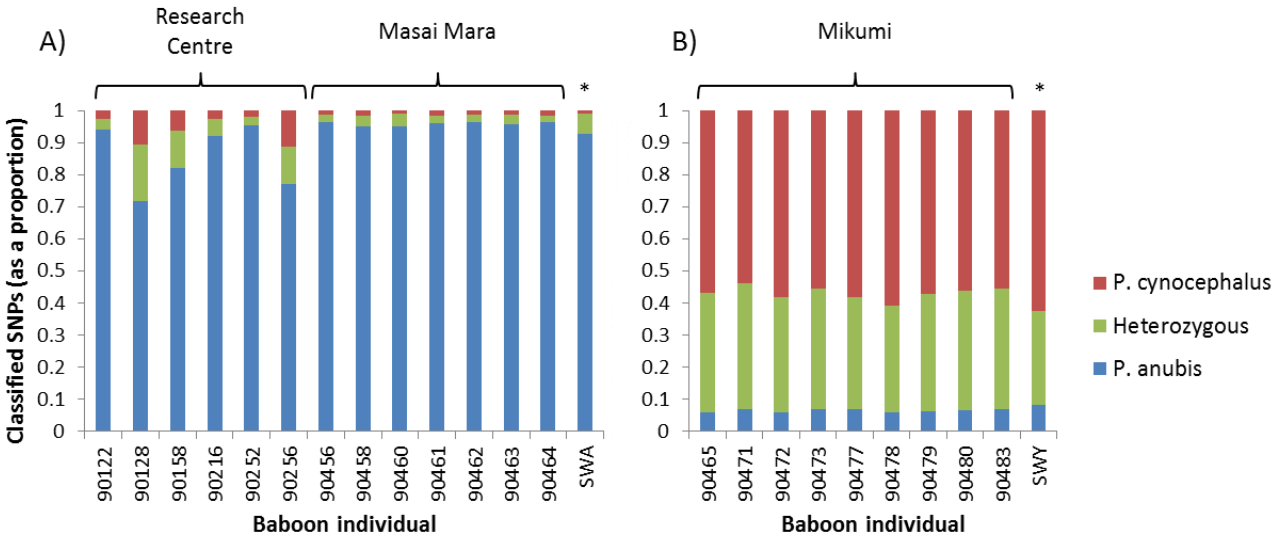


Figure 3.3: Classification of baboons used to define predictive SNPs. A) The two *P. anubis* groups SNPs classified along with the high coverage individual (marked with an *). B) The classification of SNPs of the *P. cynocephalus* baboons from the Mikumi National Park, along with the high coverage individual used for the genome assembly (marked with an *).

While the trend of the *P. anubis* SNPs in the Amboseli population follows what you would expect, with an increasing proportion of *P. anubis* SNPs as *P. anubis* ancestry increases. Deviating from what would be expected however, the unadmixed Amboseli individuals had significantly more *P. anubis* SNPs than the unadmixed *P. cynocephalus* control group ($p<0.001$; Mann-Whitney U two tailed test).

These data are consistent with the results in Wall et al (2016), despite the differences in methodology. The correlation between the estimated allele frequencies in Wall et al (2016) and here was high ($r^2=0.96$). And Wall et al (2016) concluded that there was evidence of *P. anubis* ancestry in the putatively unadmixed Amboseli individuals, which could explain the difference between the unadmixed Amboseli individuals and the Mikumi control group.

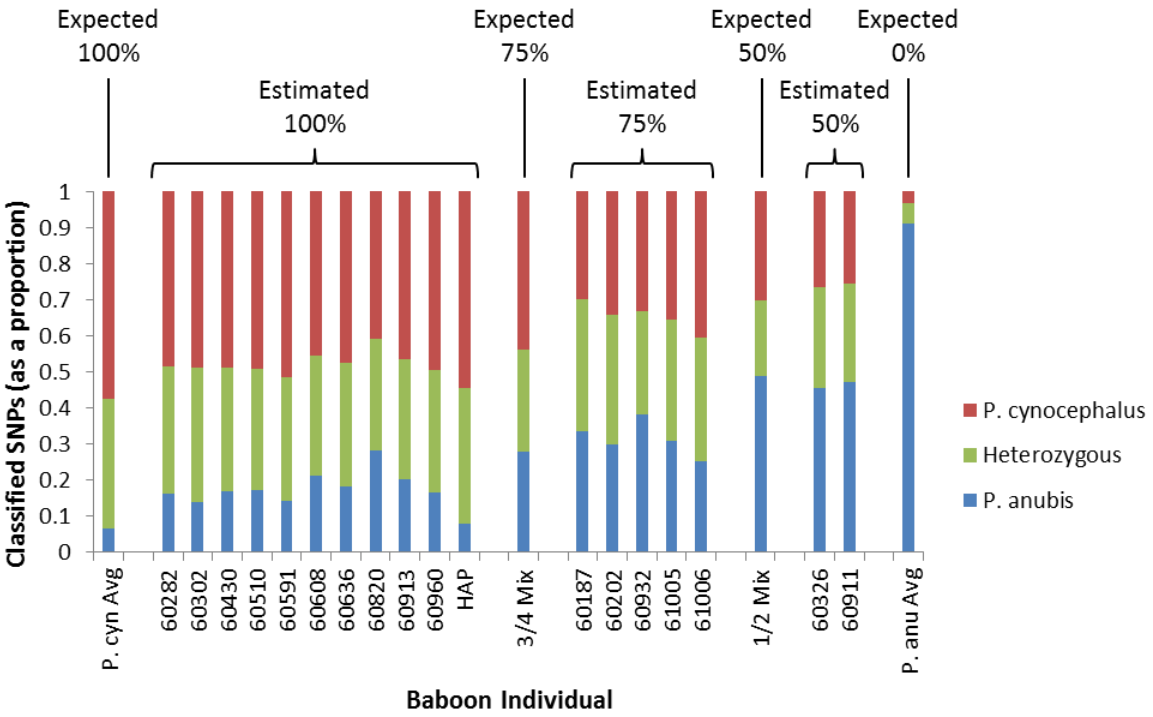


Figure 3.4: Comparison of the genetic analysis results from the Amboseli baboons to the ancestry estimates based off observation. The "expected" values are based off combining the averages of the *P. anubis* and *P. cynocephalus* individuals from outside the Amboseli park. The "estimated" values are based off observations from the field (see Wall et al, 2016).

As a result of the difference in *P. anubis* SNPs in unadmixed individuals in the Amboseli population compared to the other *P. cynocephalus* baboons, the Amboseli baboons with unknown ancestry were compared to the Amboseli baboons with known ancestry (Figure 3.5). Of the 5 individuals of unknown ancestry, 2 are consistent with the other "unadmixed" Amboseli individuals. Another two are closest to the 50% *P. anubis*/*P. cynocephalus* hybrids, although the difference between them is marked. The final individual has more *P. anubis* SNPs than any of the previously quantified Amboseli individuals, with 67.8% of the predictive SNPs being associated with *P. anubis* baboons. This is approximately half way between the average of the 50:50 hybrids and the unadmixed *P. anubis* individuals from outside of the Amboslei park. This suggests that the 67.8% value probably represents an individual that is 3/4 *P. anubis* and 1/4 *P. cynocephalus*.

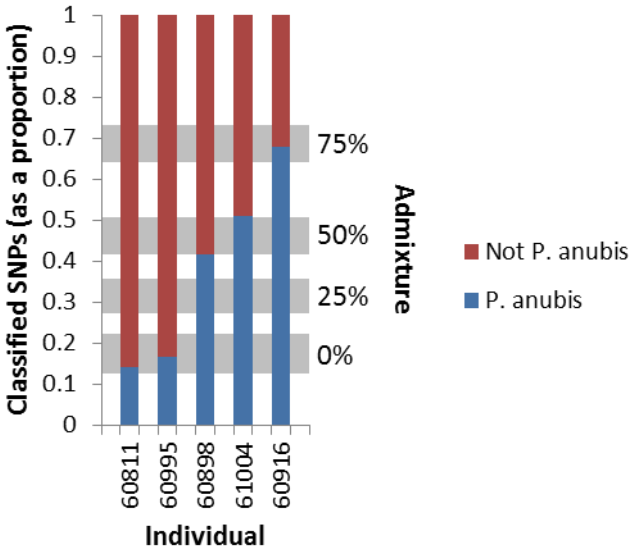


Figure 3.5: Estimating the ancestry of 5 Amboseli baboons of unknown descent. The proportion of *P. anubis* predictive SNPs has been compared to expected values (grey bars) for 0%, 25%, 50% and 75% *P. anubis* ancestry. The range displayed for 0% and 25% ancestry is the average +/- the standard deviation for previous Amboseli individuals with that ancestry. There were not enough individuals for a standard deviation estimate for the 50% and 75% ancestry range, so the displayed value is the average of the standard deviations of the 0% and 25% groups. The 50% ancestry range is centred on the average of the two known 50% indivduals, whereas the 75% range is centred half way between the 50% average and the average of the unadmixed *P. anubis* individuals.

Disproportionate SNP Frequency in Amboseli

Among the 23 Amboseli baboons, 54 SNPs were found to be disproportionately associated with the *P. anubis* populations while only 4 SNPs were found to be disproportionately associated with the *P. cynocephalus* population (Supplementary Table 3.1). This discrepancy is quite striking, especially considering that 13 of the 23 Amboseli baboons were supposed to be unadmixed individuals.

This suggests that prior to the recent observed hybridisation event, *P. anubis* alleles were transferred to the Amboseli population. These alleles have since increased in frequency to the point of fixation (or close to it), either through genetic drift or selection. If selection did play a role, it would be interesting to discover what traits are associated with these allelic changes, which were providing the selective advantage.

Discussion

In order to investigate the hybridisation and admixture between the *P. cynocephalus* and *P. anubis* populations in the Amboseli National Park in Kenya, the genome of *P. cynocephalus* was sequenced and assembled. This genome was used to compare low coverage sequencing data from multiple individuals from inside and outside the park.

Genome Assembly

The final genome assembly was of mediocre quality. A N50 scaffold size of 887 kbp was good enough to accomplish the goals for the project, but for the purposes of general use, a higher quality genome assembly would've been better. There are several potential ways that the genome assembly could be improved, including simply adding additional coverage. But any notable improvement would require additional financial investment and is beyond the scope of this project.

Identification of species predictive SNPs

Using the deviations in allele frequency between the two species of baboons, I was able to identify SNPs that were predictive of species. The SNPs consistently reclassified the baboons from outside the Amboseli Park into their respective species. The high levels of heterozygosity observed in the *P. cynocephalus* baboons were not expected however. This could be a result of the *P. cynocephalus* training data having a lower coverage on average, and fewer individuals in general. This means that there will be fewer individuals represented for any particular site, and sites that are not fixed will be selected more by chance due to the random sampling making them appear fixed. It could also be a result of previous admixture between the two species introducing *P. anubis* alleles into the *P. cynocephalus* population.

There are other problems with the analysis, and if it were to be repeated, there are several improvements that could be made. The main one would be the use of allele frequencies at a site for each individual. This was used to try and account for the uncertainty associated with the low coverage of the samples. But this method is flawed because it treated each individual as being independent. In reality, each additional homozygous individual should lower the estimated allele frequency in the population, which isn't necessarily the case for the current method. With better

allele frequency estimates across a population, more stringent site selection would be possible. Once these accommodations are made, this methodology is an option for future low coverage population genetic analyses, such as those that result from non-invasive, low yield dna extraction protocols (Snyder-Mackler et al, 2016).

Admixture in Amboseli Baboons

Using the species specific SNPs, the field observations describing the ancestry for the baboons in the Amboseli Park were largely confirmed, with predicted hybrid individuals having proportionately higher rates of admixture (Figure 3.4). Additionally, baboons with unknown ancestry were also able to be classified to plausible pedigrees using these species specific SNPs. Perhaps of more interest was the observation that the Amboseli individuals had a higher proportion of SNPs of *P. anubis* ancestry than you would expect if hybridisation was a recent phenomenon in the area. This result differs from what was thought from Amboseli park observation, which suggested that the admixture began in 1983 (Alberts and Altman, 2001). Instead, it appears that the Amboseli population has a history of admixture with the neighbouring *P. anubis* population. This result was supported by the finding that several regions of the Amboseli baboons' genomes appear to have the representative *P. anubis* allele present at a high frequency. This supports the findings in Wall et al (2016), despite differences in the methodology that were used in the final publication, and adds to the growing body of evidence showing just how common hybridisation in primates is (Arnold and Meyer, 2006; Zinner et al, 2011).

Conclusion

This chapter successfully confirms the hybridisation of the *P. cynocephalus* population in Amboseli National Park in Kenya with the recently arrived *P. anubis* population. This was done using low coverage sequencing obtained using non-invasive methods. Furthermore, there was evidence of past interactions between these two populations. And finally, this section resulted in a published genome assembly for a baboon species, whose clade was not previously represented in the literature.

Chapter 4: The Ruschioideae Expansion

The Greater Cape Floristic Region stretches along the West Coast of South Africa and into Namibia (Born et al, 2007). As the name suggests, it includes the biodiversity hotspot the "Cape Floristic Region", and has been expanded to include the surrounding winter rainfall region. The climate of this region changed approximately 10 million years ago, when it switched to a winter rainfall system and became colder and dryer (Diekmann et al, 2003; Krammer et al, 2006). This change placed new environmental pressures on the plants in the region and is thought to have precipitated widespread diversification through the creation of novel niches (Cowling et al, 2009).

The Aizoaceae are a family of leafy succulents of approximately 1800 species found within the Greater Cape Floristic Region (Richardson et al, 2001; Klak et al, 2017). Within this family is an extremely prolific tribe, known as the Ruschieae (Figure 4.1), with around 1500 species, nested in the subfamily Ruschioideae (Klak et al, 2013). The date of the proliferation for the Ruschieae has been contested (Arakaki et al, 2011), but is generally thought to have followed the change in climate in the region (Valente et al, 2014). Depending on the date of divergence used[13], the rapid radiation of the Ruschieae (Figure 4.1A) would be the highest known rate of speciation in land plants (Klak et al, 2004; Valente et al, 2014).

Consistent with other radiation events, like the Cichlids in Africa (Rabosky et al, 2013), this rapid speciation has been accompanied by large morphological changes. In the Ruschioideae, these changes often include features such as leaf shape, fruit morphology and overall size (Illing et al, 2009; Figure 4.1B for two examples). This rapid morphological change suggests that there are many available niches available and that genetic isolation is a common occurrence among the populations (Ihlenfeldt, 1994). This is thought to have resulted in the rapid ecological speciation (Rundle and Nosil, 2005).

[13] Valente et al (2014) estimated the radiation to begin 0.35-3.14mya, Klak et al (2014) estimated 3.8-8.7mya and Arakaki et al (2011) estimated 17mya.

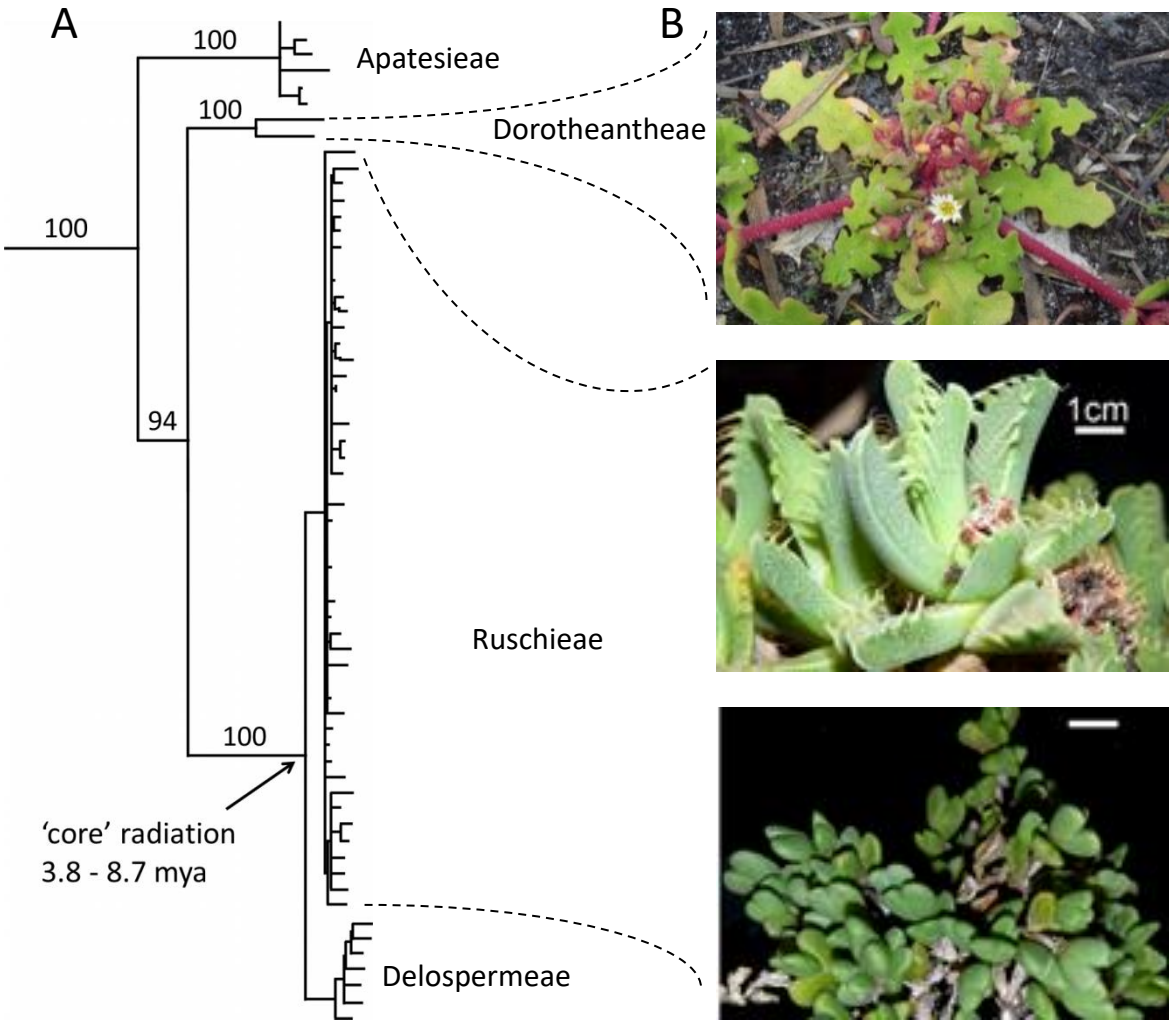


Figure 4.1: Rapid speciation and morphological change in the Ruschioideae. A) The rapid rate of speciation can be seen in the phylogeny of the Ruschioideae adapted from Klak et al. (2004). The 4 tribes of the Ruschioideae are labelled to the right of the phylogeny. B) This radiation is filled with diverse morphological adaptation. Pictured plants (from top to bottom) *Cleretum herrei*, *Faucaria felina* and *Polymita steenbokensis,* which were all used in this study.

Ecological selection pressures can go a long way to explaining the morphological diversity observed in the plants. But the Ruschioideae are not found in isolation, and other plant families which are also present in these environments do not show a similar profile of rapid morphological diversification. A proposed explanation for this is that the Ruschieae developed unique morphological traits that allowed for far more effective niche exploitation than the neighbouring species. These traits include the characteristic thick, succulent leaves which have replaced the more conventional thin, flat leaves seen in other tribes (for example, see *Cleretum herrei* from the tribe Dorotheantheae in Figure 4.1B).

But it also seems plausible that there is a genetic component that is contributing to the phenomenon of rapid speciation and morphological adaptation, perhaps by making the plants abnormally plastic or prone to speciation. This component could take several forms. It is possible there is a gene that is allowing for an abnormally high frequency of mutations, such as a change to the DNA repair mechanisms. A second option, for which there is some evidence (Illing et al, 2009), is that a recent genome duplication occurred before the radiation (Kellogg, 2016), allowing for gene duplicates to diverge in function while the original gene copies maintain their function (Panchy et al, 2016). This phenomenon has been shown to be associated with environmental change (Van de Peer et al, 2017), which is consistent with this particular case. A final option is that the genome has become susceptible to the occurrence of smaller, local duplications. These duplicates could then diverge in function in a similar manner as with gene duplication, where one copy retains the original function (Holland et al, 2017). This might occur as a result of Transposable Element activity, which can result in intervening material being copied (Cusack and Wolfe, 2007; Jiang et al, 2004), or else some other event (Freeling, 2009), such as unequal crossing over, resulting in tandem duplicates.

Whole Genome Duplication

Whole genome duplication is a surprisingly common occurrence among plants (Paterson et al, 2010; Wang et al, 2012). This can be a source of novel genetic sequence, with some genes diverging in function (Wang et al, 2011). Unfortunately, due to the large amount of duplicated sequence, whole genome duplication also causes large problems for genome assembly algorithms (Kyriakidou et al, 2018). This is especially true the more recent the genome duplication is, as the copies are not yet diverged enough for algorithms to properly disentangle.

Transposable Elements

Transposable elements are selfishly replicating sequences that move around within a genome (Bourque et al, 2018). These elements have large genomic restructuring potential (Springer et al, 2018). This restructuring can include the inadvertent copying of neighbouring DNA or the generation of large amounts of by product sequences such as Long Terminal Repeats (LTRs) or short and long interspersed nuclear elements (SINEs and LINEs).

The Project

This exploratory project sought to try and discover if there is a genetic component behind the rapid speciation and morphological change observed in the Ruschieae. In order to do this, two species from within the Ruschieae tribe with divergent morphological traits were chosen along with a Ruschioideae species from outside the Ruschieae tribe for sequencing and comparison. The aim of this project is therefore to use these genome sequences to identify commonalities within the Ruschieae genomes which are not present in the out-group genome and deduce whether or not these factors could plausibly play a role in the Ruschieae's rapid rate of speciation and morphological change.

Methods

Estimates of genome size and ploidy

The genome sizes of several Aizoaceae species (*Galenia africana*, *Tetragonia fruticosa*, *Mesembryanthemum crystallinum*, *Mesembryanthemum prasinum*, *Conicosia elongata*, *Dorotheanthus bellidiformis*, *Delosperma echinatum*, *Faucaria felina*, *Mossia intervallaris*, *Carruanthus ringens*, *Polymita steenbokensis*, *Scopelogena bruynsii*, *Cephalophyllum pillansii*, *Fenestraria rhopalophylla*, *Pleiospilos simulans*, and *Drosanthemum speciosum*) were roughly estimated by propidium iodide staining of plant nuclei following the protocol described in Doležel et al (2007). Aizoaceae samples were obtained from Cornelia Klak (Department of Biological Sciences, UCT), with the following exceptions: *M. crystallinum* seeds were obtained from John Cushman (University of Nevada, Reno) and *D. bellidiformis* seeds were bought from Starke Ayres (Reference:

909814BCAH). Additionally, plants with known genome sizes were used as standards: *Secale cereale*, *Pisum sativum* and *Solanum lycopersicum* seeds were obtained from Jaroslav Doležel (Palacký University, Czech Republic). The nuclei suspension fluorescent signal was quantified using a flow cytometer, BD Bioscience LSR II.

DNA Isolation

DNA was isolated from approximately 25cm^3 of fleshy leaf material from individual plants of *F. felina*, *P. steenbokensis* and *C. herrei*. *F. felina* and *P. steenbokensis* were obtained from Cornelia Klak, having been grown in the UCT glasshouses. *C. herrei* was collected from Silvermine (GPS 34° 5' 27.80"S 18° 25' 28.85"E). The isolation protocol used (protocol B from Lutz et al, 2011) made compensations for large amounts of DNase, which proved to be essential for obtaining high molecular weight DNA. It also included a nuclei extraction step to reduce chloroplast contamination and maximise genomic coverage in the sequencing results. The resultant DNA solution was further purified using a Qiagen Genomic-tip 100/G column according to manufacturer's instructions. The purified DNA was then sent to the Beijing Genomics Institute (BGI) to be sequenced.

DNA sequencing

Approximately 80x coverage was aimed for across each genome. *F. felina* and *P. steenbokensis* were sequenced with paired end 100bp reads, using two library insert lengths (500bp and 800bp), prepared by BGI. *C. herrei* was sequenced later with the same two libraries, but by the time of the sequencing, the technology had advanced and allowed for 125bp paired end reads. Following the initial assembly drafts of *F. felina* and *P. steenbokensis*, which were poorer than expected, additional sequencing using MiSeq was conducted from the Oregon Health & Science University sequencing unit. Although this does not improve the coverage by much, with the MiSeq platform allowing for much less throughput, it was hoped that the longer read lengths (300bp paired end reads, 500bp insert size) might improve the contig lengths in the final assembly.

Read Processing

The quality of the sequenced reads was assessed using Fastqc (version 0.10.1) (Andrews, 2010). Trimmomatic (version 0.32) (Bolger, 2014) was then used to trim bases with a quality score less than 17 from the end of the read and reads shorter than 60bp were discarded. A k-mer frequency plot was created from the reads using KmerFreq_HA (version 2.01) (from the SOAPdenovo package) with a k-mer size of 25 (Luo et al, 2012) and reads were error corrected with Corrector_HA (version 2.01) (Luo et al, 2012). The default settings were used for 100bp/125bp reads, while the longer MiSeq reads were allowed 4 corrections instead of the regular 2. Duplicate reads were then removed using FastUniq (version 1.1) (Xu et al, 2012).

Genome Assembly

The processed reads were assembled using Platanus (version 1.2.1) (Kajitani et al, 2014) after k-mer frequency plots suggested a high level of heterozygosity and early SOAPdenovo2 (version 2.04) (Luo et al, 2012) assemblies were poor. Assembly quality was predominantly assessed using the N50 scaffold size metric, while other metrics, such as the assembly length, proportion of N's and contig N50 were monitored for abnormal behaviour. Optimisation of assemblies primarily occurred after the scaffolding step and used varying k-mer sizes as well as methods like not using or not using the MiSeq sequences for scaffolding. The best two assemblies were advanced to the gap closing step as a precaution, but in all cases, the better assembly before gap closing remained superior after the final stage of processing.

Identification of Highly Represented Sequences

In order to identify highly represented sequences in each of the genomes, one million reads from the 500bp insert library from each plant were aligned to their respective genome using Bowtie2 (version 2.2.4) (Langmead and Salzburg, 2012). These reads were aligned such that there was no maximum number of alignment positions for each read, differing from the Bowtie2 default settings. Putative long terminal repeats were then manually assembled from the reads which aligned 1000 or more times in each respective plant. The one million reads were then aligned to each of the identified long terminal repeats, as well as the RepBase v19 database (Jurka et al, 2005) using Blastn

(version 2.2.29) (Altschul et al, 1990). In order to make the analysis more sensitive and account for the short read lengths, a Blastn wordsize of 10 was used with an e-value of 1e-6.

Checking for Recent Genome Duplication

In order to test whether a whole genome duplication preceded the divergence of the Ruschioideae, *Arabidopsis thaliana* genes were aligned to each of the three assembled genomes using Blastx (version 2.2.29) (Altschul et al, 1990). Thirty genes which had the first and last 50 amino acids align contiguously two or more times in each genome were selected for further analysis.

The selected portions of the genes were aligned to each other using ClustalW (version 1.2.0) (Thompson et al, 1994), and Maximum Likelihood Trees were created using Mega (version 5.05) (Tamura et al, 2011). These phylogenies were then assessed to see whether a consistent pattern of duplication could be observed across the different genes. This was done for 31 genes.

Measuring Relative Gene Duplication

The copy number of genes in each genome was estimated by aligning *Arabidopsis thaliana* genes to each of the genomes using Blastx (Altschul et al, 1990). From these results, the number of times each base pair of a gene was represented in each genome was calculated. The difference in gene representation between the two Ruschieae genomes and the *C. herrei* genome was tested for statistical significance using Mann-Whitney U tests. A Bonferoni correction was done to account for multiple testing, with an overall significance threshold of 0.05.

Tandem Duplications

In order to test whether the recent gene duplications within the Ruschieae were tandem duplications or not, *Arabidopsis thaliana* genes were aligned to the two Ruschieae and *C. herrei* genomes using Blastx (version 2.2.29) (Altschul et al, 1990). A gene was used if 45 of the last 50 amino acids aligned at least twice in one of the genomes, but only once in each of the other two genomes. Duplications were then counted as being tandem if two or more of the alignments of a gene within the same genome were on the same scaffold. In order to account for the differences in

genome assembly quality, the total number of gene duplicates in each plant was adjusted according to the size of the scaffold that the gene appeared on. This used the distances between identified tandem duplicates from *C. herrei* as the expected null distribution for all the plants. Each duplicate was therefore normalised according to the chance of identifying a hypothetical tandem duplicate, given the size of the scaffold.

Results

Estimates of genome size and ploidy

Flow cytometry analysis of propidium iodide stained nuclei showed that the Aizoaceae species had a wide range of genome sizes (0.46 Gbp to 3.9 Gbp). The estimation of genome size was complicated by extensive endopolyploidy (Figure 4.2A). This means that certain populations of cells in the leaf tissue have multiple duplicated genome copies in the nucleus. This creates multiple fluorescent peaks in the signal given off by propidium iodide stained nuclei while also reducing the number of cells with the usual 2n genome count. The low signal from the 2n cells combined with large amounts of noise made identifying 2n peaks difficult (Figure 4.2B). As a result of these difficulties, genome estimates should be viewed as an upper bound of the actual genome size. It is quite plausible that the 2n peak was obscured and a larger peak (4n, 8n, etc) was erroneously estimated.

These results included *M. crystallinum* at 460 Mbp, which had been previously measured to be 390 Mbp (De Rocher et al, 1990). In comparison, the three smallest Ruschieae genomes were *C. pillansii* (660 Mbp), *M. intervallaris* (820 Mbp) and *D. speciosum* (860 Mbp). Based on available plant material, *F. felina* and *P. steenbokensis* were chosen for sequencing from the Ruschieae tribe. Their respective genomes sizes were estimated to be 0.94 Gbp and 0.90 Gbp. In addition, the genome of a third plant, *Clereum herrei* from the tribe Dorotheantheae was also sequenced and assembled, and used as the outgroup in the comparison to the two Ruschieae genomes (*F. felina* and *P. steenbokensis*).

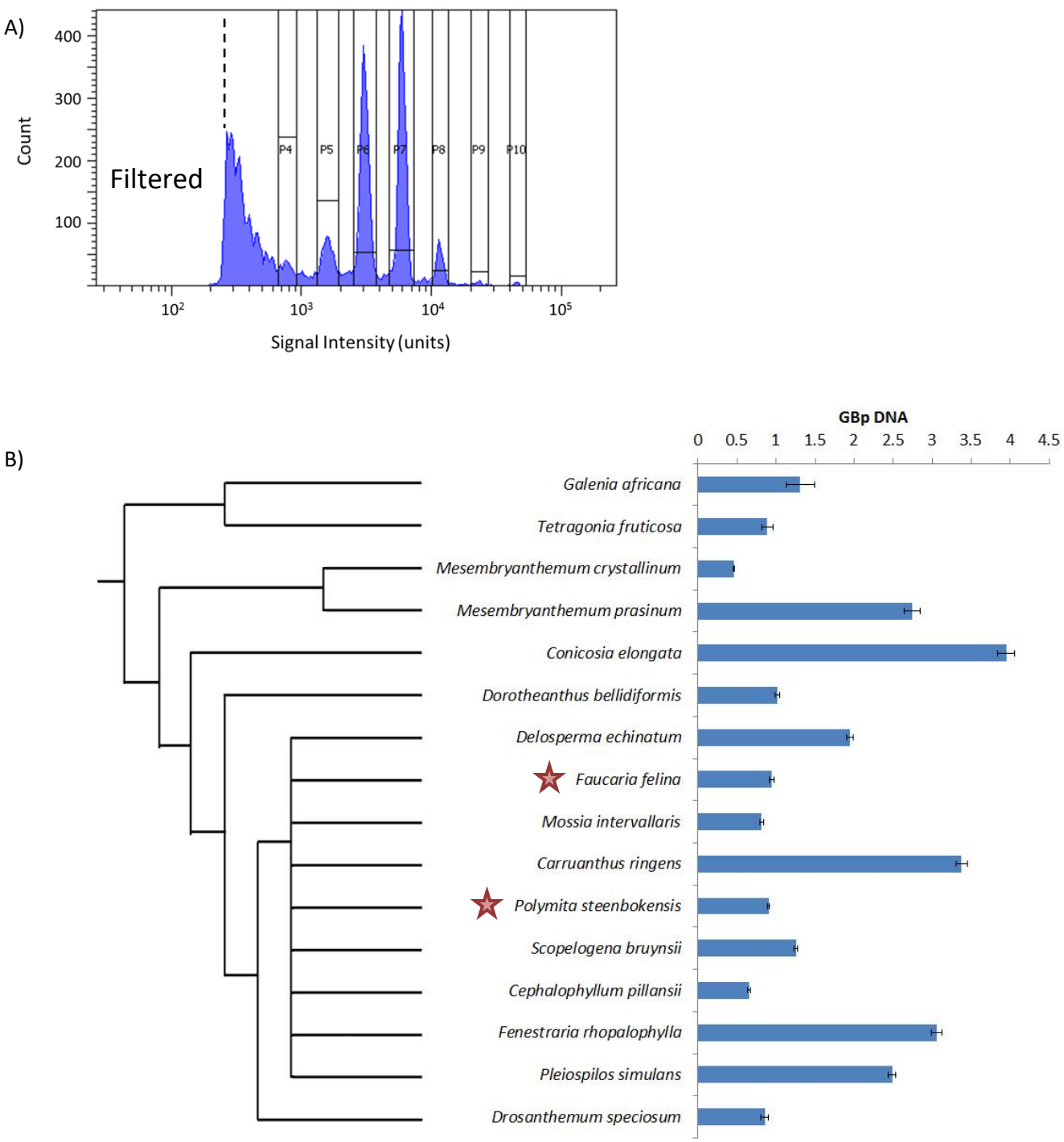


Figure 4.2: Estimates of genome size for 16 Ruschioideae species. A) A representative result of the cumulative fluorescent signal from the nuclei suspension of Aizoaceae plants (in this case *Mesembryanthemum crystallinum*). Each windowed peak represents a different nuclei size resulting from varying copy numbers of the genome. The peak with the lowest signal strength (left most window) represents the 2n genome size. The increasing amount of noise on the left hand shoulder makes it difficult to be confident the smallest visible peak is actually the smallest represented peak. Note that the noise stops increasing at the dotted line because it was filtered at this point at the

time of capture. B) Variation in genome size estimates obtained with propidium iodide staining. Estimates should be viewed as an upper bound of the genome estimate, as a larger peak (4n, 8n, etc) may have mistakenly been identified instead of the 2n peak. Error bars represent the standard error of the mean. Aizoaceae phylogeny adapted from parsimonious consensus trees from Klak et al (2003) and Klak et al (2007). The selected Ruschieae for genome sequencing, *F. felina* and *P. steenbokensis*, are highlighted with red stars.

DNA Isolation, Sequencing and Read Processing

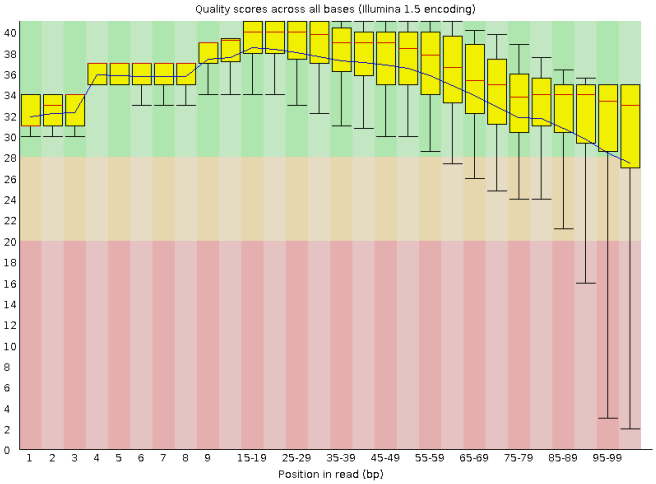
High molecular weight genomic DNA was successfully isolated from *F. felina*, *P. steenbokensis* and *C. herrei*, and sent for sequencing. The sequencing reads (Table 4.1)were of suitabally high quality for genome assembly (Figure 4.3).

Table 4.1: Resultant sequencing depth for each of the Ruschioideae. 'Genome Size' is estimated by KmerFreq_HA (Luo et al, 2012). Note that *C. herrei* reads are 125bp and MiSeq reads are 300bp. All other reads are 100bp.

		C. herrei	*F. felina*	*P. steenbokensis*
Genome Size (n)		320 Mbp	800 Mbp	960 Mbp
Number of read pairs	MiSeq	-	8.2 million	10.5 million
	500bp	54 million	185 million	161 million
	800bp	52 million	114 million	123 million
Coverage		82.6x	80.8x	65.7x

P. steenbokensis

Forward



Reverse

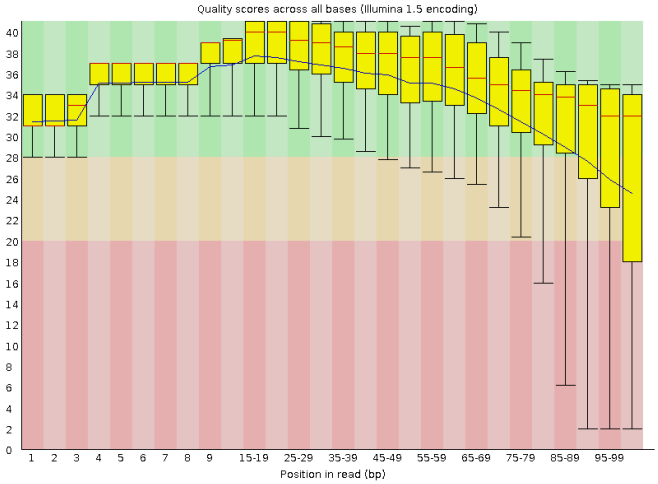


Figure 4.3: Representative read sequencing quality. The above example is from the 500bp insert library for *P. steenbokensis*. Images generated by FastQC (Andrews, 2010).

A lenient read trimming strategy was used in order to get the maximum possible coverage for the genome assembly process. This resulted in about 0-1.5% of reads being removed from each dataset due to poor quality, with fewer reads being deleted from the longer read sets. The opposite was true for the trimming statistics, with 12.6-51.1% of reads being trimmed, with higher percentages of reads being trimmed in the longer MiSeq datasets. This pattern is a result of how many nucleotides can be trimmed from longer reads before they are too short to use.

The k-mer frequencies of the trimmed reads (Figure 4.4A) were generated to use for error correction, where low frequency k-mers (<3x coverage) were deemed untrustworthy. These k-mer distributions showed several interesting features. The most notable of these features was the size of the heterozygous peak in *F. felina* and *P. steenbokensis*, which was absent in *C. herrei*.

The next feature of interest is the 4n peak in *C. herrei* (Figure 4.4A). This peak could represent an old genome duplication. If that is the case, it could have happened before or after the divergence of the Ruschieae. This is because the heterozygosity of the Ruschieae would hide any equivalent 4n peak, if

it was present[14]. If the 4n peak is a result of a genome duplication, it looks like most sequences have diverged at a 25bp resolution, since the bulk of the k-mers are in the 2n position.

The final feature of interest is the number of high frequency k-mers in the Ruschieae species (Figure 4.4B). There is no obvious whole genome duplication peak in these k-mers (although one would not necessarily be easy to see). There are however a lot of high frequency k-mers in general, suggesting that a lot of duplication of some kind has occurred in *F. felina* and *P. steenbokensis*, but not *C. herrei*.

Genome Assembly

The k-mer frequencies (Figure 4.4) have several features that suggest that the future Ruschieae genome assemblies will be difficult. The large number of heterozygous sites in the genome adds a lot of complexity to the assembly, with each heterozygous site adding 'k' new k-mers to the *de Bruijn* graph. The duplicated sequences will also cause problems, erroneously collapsing nodes in the *de Bruijn* graph. It can also be difficult to differentiate a highly heterozygous region from a duplicated region in the assembly process, so having high levels of both is not ideal.

As predicted, the genome assembly for the two Ruschieae species was challenging. The SOAPdenovo (Luo et al, 2012) assemblies failed dismally, despite many attempts to take into account the high heterozygosity. Efforts, such as increasing the aggressiveness with which heterozygous regions are merged (using the -M option of SOAPdenovo), were unsuccessful. SOAPdenovo N50 scaffold values consistently ranged in the 1-2 kbp range, with only the most minor of improvements being noticeable. As a result, Platanus (Kajitani et al, 2014), which was created with heterozygous plant genomes in mind was used instead. This proved more successful, although the assemblies are still in a highly fragmented state, with scaffold N50 sizes increasing to 5-6 kbp (10-15% of which is N's). The difference in the assembly quality of these genomes, compared to *C. herrei,* was notable, with an N50 scaffold size of 50 kbp, only 1.7% of which is N's (Table 4.2). This difference in assembly quality is presumably a result of differences in heterozygosity, genome size and repetitive elements. Finally, the assembly lengths of the 3 genomes closely match the predicted genome sizes from the k-mer frequencies (Table 4.1).

[14] In the absence of heterozygosity, duplications result in peaks at the frequencies of 2n, 4n, 6n (depending on the mode of duplication), 8n, etc. But with heterozygosity these peaks are diminished and intermediatory 3n, 5n and 7n peaks are created.

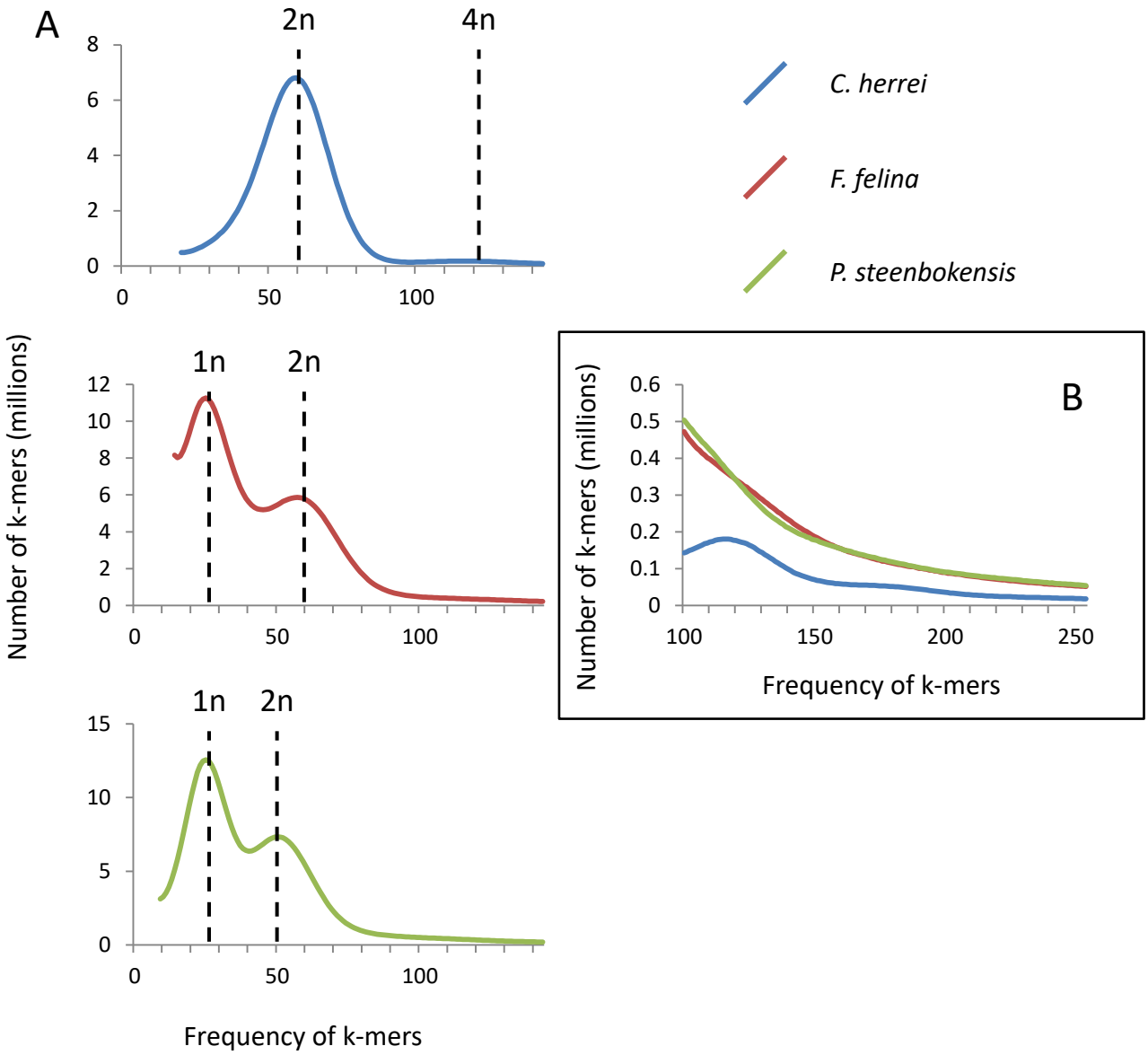


Figure 4.4: K-mer frequency plots for the three Ruschioideae species. K-mers (25bp) were derived from trimmed reads for each plant. The main heterozygous (1n) and homozygous peaks (2n) are shown in A). Interestingly, a 4n peak was observed in *C. herrei*, but not *F. felina* and *P. steenbokensis*. The high frequency k-mers of the three plants have been plotted on the same set of axes in B). This shows the 4n peak from *C. herrei* and an elevated high k-mer frequency levels in *F. felina* and *P. steenbokensis*.

Table 4.2: Genome assembly results of Ruschioideae genomes. *C. herrei* was assembled to a much better quality than *F. felina* and *P. steenbokensis*.

	C. herrei	*F. felina*	*P. steenbokensis*
Assembled Length	289 Mbp	796 Mbp	998 Mbp
Largest Scaffold	590 kbp	172 kbp	108 kbp
Scaffold N50	50.6 kbp	6.2 kbp	5.3 kbp
%N	1.7%	10.7%	15.4%

Identification of Highly Represented Sequences

In order to investigate the highly represented sequences identified in the k-mer frequency plots (Figure 4.4), one million reads from each plant were aligned to their assembled genomes without an alignment limit, in order to identify reads stemming from repetitive elements (Figure 4.5). This clearly showed a larger proportion of reads aligning more than once within the Ruschieae genomes (>70%) compared to the out group (<45%), *C. herrei*. This suggests that there has been a significant amount of duplication, in some form or another, since the divergence from *C. herrei*. This could also explain the large difference in the observed genome size of the Ruschieae (Figure 4.2 and Table 4.2). Curiously the maximum number of times a read aligned was very similar between the three genome assemblies. This could represent some sort of an upper limit for the genome assembly algorithm, such as the number of times a similar sequence can successfully be distinguished and assembled, for example. It seems unlikely that the most highly repeated sequences in the genome would A) be assembled accurately and B) would be the same level of representation in all three genomes. The exact numbers for these sequences therefore probably shouldn't be trusted. For this reason, it was decided to not exclusively focus on the reads that aligned the most, but rather look at the reads that aligned a lot in general (in this case 1000 times or more).

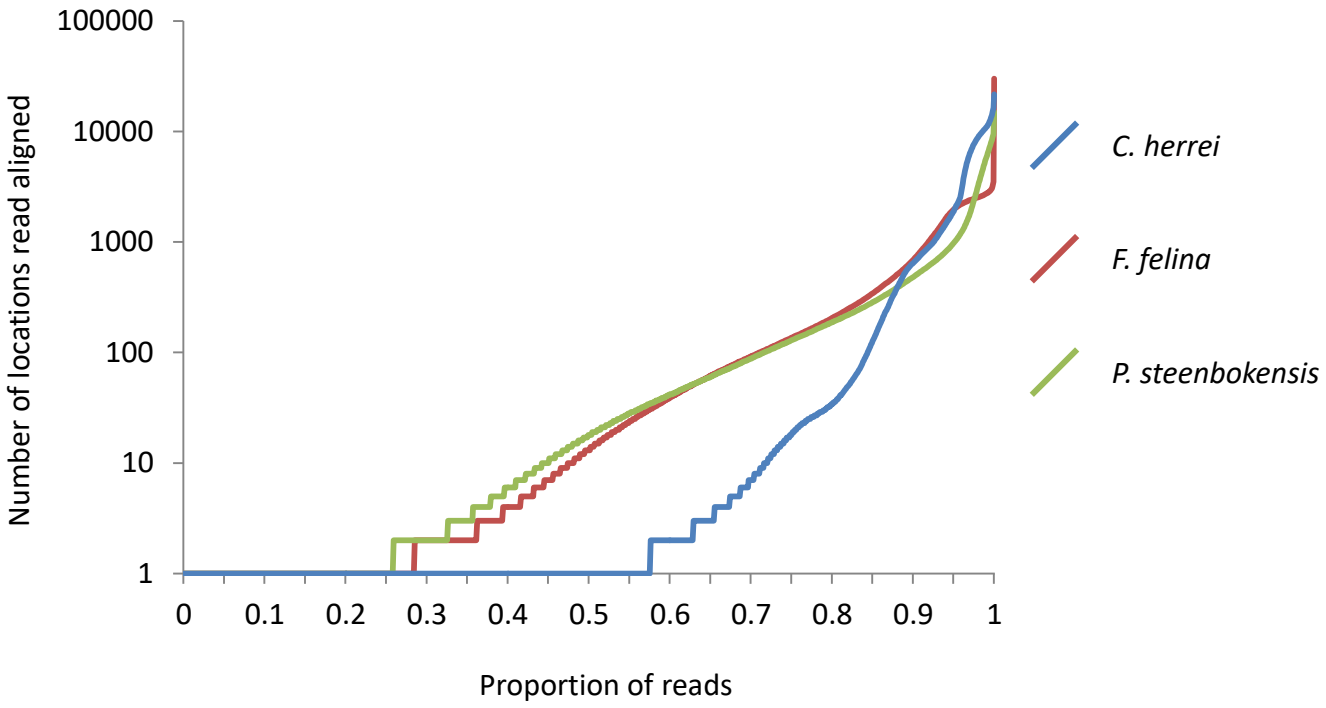


Figure 4.5: Alignment rates for reads from each of the Ruschioideae genomes. The two Ruschieae species follow a similar pattern with regard to their alignment rates (green and red lines), with reads starting to map more than once between the 25th and 30th percentile, compared to *C. herrei,* which starts between the 55th and 60th percentile (blue line). *C. herrei* appears to catch up to the two Ruschieae around the 85th percentile. Note that the y-axis is log transformed.

The sequences of reads that aligned 1000 or more times were investigated in order to identify common features. From these reads three common sequences were identified and assembled manually from the three plants. These sequences probably represent long terminal repeats, considering their size (200-500bp) and repetitive nature. The one million reads were then aligned to each of the identified long terminal repeats, as well as the RepBase v19 database (Jurka et al, 2005) using Blastn (Altschul et al, 1990). This was done using the unassembled reads to try and account for potential assembly bias within the genomic sequence. Of the three identified putative long terminal repeats, two were unique to the Ruschieae and one was unique to *C. herrei* (Table 4.3).

Table 4.3: Presence of identified long terminal repeats in whole genome sequencing reads. The proportion of identified long terminal repeats was determined using reads instead of genome assemblies to account for assembly biases.

LTR	*C. herrei*	*F. felina*	*P. steenbokensis*
Ruschieae 1	0.0%	5.32%	0.16%
Ruschieae 2	0.0%	1.74%	2.48%
Cleretum 3	5.12%	0.0%	0.0%
RepBase	5.86%	3.22%	3.06%
Total	10.98%	10.28%	5.70%

These results suggest that there has been recent transposable element activity (Table 4.3). Both of the long terminal repeats found in the Ruschieae plants were not present in *C. herrei* and the *C. herrei* long terminal repeat wasn't present in the Ruschieae genomes. Additionally, one of the long terminal repeats was found in substantially larger proportions in *F. felina* than in *P. steenbokensis*, meaning that the transposable activity continued after their divergence, during the species radiation of the Ruschieae.

Overall, taking the RepBase proportions into account (Table 4.3), there does not appear to be disproportionately more LTRs and highly repetitive elements in the Ruschieae compared to *C. herrei*. In fact, despite the comparable genome size between *F. felina* and *P. steenbokensis*, *P. steenbokensis* was the plant out of the three which had the lower proportion of repetitive elements. *C. herrei* in fact had the highest proportion of repetitive elements, although *F. felina* was not far behind.

Checking for recent genome duplication

Repeated whole genome duplication probably can't explain the thousands of speciation events of the Ruschieae. But one or two early whole genome duplications could potentially have created redundancy in the genes for natural selection to act upon. This, combined with early adaptive advantages (succulent leaves) and available niches could explain the species radiation. The

hypothesis is supported by the difference in genome size between the two sequenced Ruschieae and the divergent *C. herrei* (and *M. crystallinum*).

If one or more genome duplications did occur in the plants' evolutionary history, duplicated genes should share a consistent pattern of duplication in their phylogenies. There are several possible ways these duplications could arise in the plants, and these possibilities should result in different gene phylogenies. An ancestral gene/genome duplication should result in two separate clades, each mirroring the species phylogeny (Figure 4.6i). In comparison, a gene/genome duplication in the Ruschieae after the divergence from *C. herrei* would have the *C. herrei* genes (with an independent duplication) as an outgroup to the Ruschieae genes (Figure 4.6ii). The final possibility is that each plant has had the gene duplicated independently (Figure 4.6iii). In reality, each gene will have its own pattern or combinations of patterns in the phylogeny, as a result of that genes individual history. But if a whole genome duplication did occur, there should be a consistent pattern in all the genes.

In order to test these alternative scenarios of wholegenome duplication events, *Arabidopsis thaliana* genes were aligned to each genome using Blastx (Altschul et al, 1990). The results were then filtered for genes that had the first 50 amino acids and the last 50 amino acids align at least twice in each of the Ruschioideae genomes. Therefore only genes that had undergone some form of duplication in each plant were selected. The first and last 50 amino acids were used (instead of more of the gene) in order to partially account for the poor genome assembly of the Ruschieae genomes. This also allows verification of each genes phylogeny.

To characterise when the duplications occurred, 29 genes' phylogenies were created[15] using the first and last 50 amino acids (Supplementary Table 4.1). These were used to identify which duplication models (of the three considered scenarios in Figure 4.6) were supported. Note that it is possible for a gene to support more than one model if enough are found. A consistent duplication event in the genes occurring between the divergence of *C. herrei* and the Ruschieae could explain the rapid radiation of the Ruschioideae. If present, this may have been a causative factor in the radiation.

[15] One of the planned 30 genes failed to have the last 50 amino acids align using ClustalW.

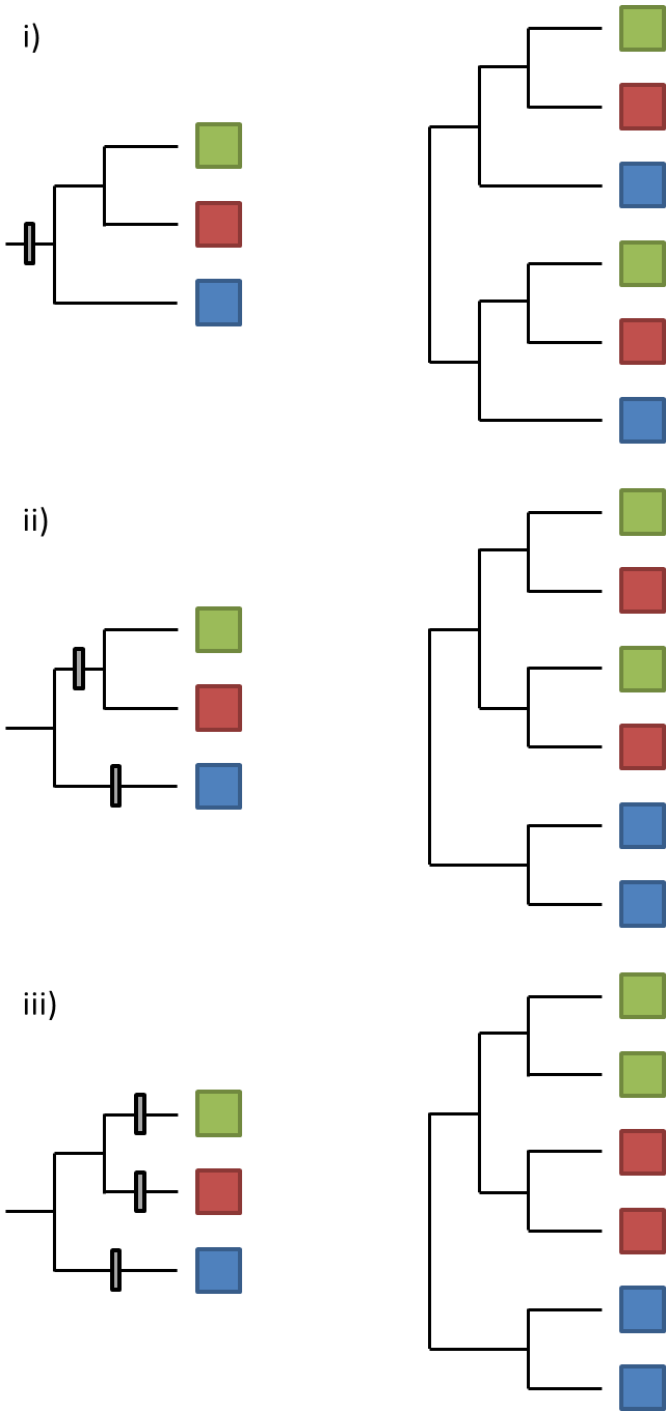


Figure 4.6: Three scenarios that result in duplicates in each plant. The diagrams on the left show the potential points of duplication (represented by long rectangles). The diagrams on the right show the resultant gene phylogeny.

A duplication event before the radiation of the Ruschieae was not consistently found ('Middle' in Table 4.4). This suggests that whole genome duplication is not responsible for the observed rapid speciation. Instead, most genes were found to have been duplicated early on in their evolutionary history, before the divergence of *C. herrei* (labelled 'Early' in Table 4.4). This shows that most genes that were duplicated in all three plants were duplicated before the plants diverged (not necessarily as part of a genome duplication). Finally, it is worth noting that the data does not support the possibility of genome duplications occurring independently within the Ruschioideae species (a 'Late' duplication in Table 4.4).

Table 4.4: Classification of gene duplication times in Ruschioideae. Homologues of *Arabidopsis thaliana* genes were aligned to the genome of *C. herrei*, *F. felina* and *P. steenbokensis*. Phylogenies were created from the first and last 50 amino acids of genes identified with duplicates. Variation in the Table represents disagreement between the phylogeny of the last 50 amino acids and the first 50 amino acids. An 'Early' duplication is a duplication before the divergence of the three plants. A 'Middle' duplication is after the divergence of *C. herrei*. And a 'Late' duplication is after the divergence of all three plants. One gene failed to align properly in the multiple sequence alignment and was removed from the analysis. It is possible for a gene to support multiple duplication hypotheses, and as such the numbers add up to more than the number of genes analysed.

Genes Analysed	29
Early	23-24
Middle	4-6
Late	4-5

Measuring relative gene duplication

If whole genome duplication is not responsible for the observed change in genome size and high frequency k-mers, it would suggest that some other type of duplication is occurring. If this

duplication process is playing a role in the radiation of the Ruschieae, it would need to be duplicating functional genetic sequences, even if it occurs at a relatively low efficiency.

Arabidopsis thaliana genes were aligned to each genome using Blastx (Altschul et al, 1990). The number of times each base pair of a gene was represented in each genome was calculated from these results. This gave a rough estimate of its representation within the genome but also controlled for variation across a gene in relative comparison. If a domain is common, it should generally be common in all the genomes. If the region is divergent and not conserved between *A. thaliana* and the Ruschioideae, its absence should be consistent across the Ruschioideae genomes. In this way, while the exact copy number of a gene might not be obvious, it should be possible to tell whether the copy number has increased or decreased in the two Ruschieae relative to *C. herrei*.

This method was partially chosen because it does not require the annotation of the genomes (which would be problematic given their fragmented state) and it allows for genes to be split up across multiple fragmented scaffolds. Missing sequences from the assembly will still have a negative impact on the overall result, and could cause some false positives, but this would be the case for any similar method used. The average representation for each of the *A. thaliana* genes in the three Ruschioideae genomes is shown in Figure 4.7.

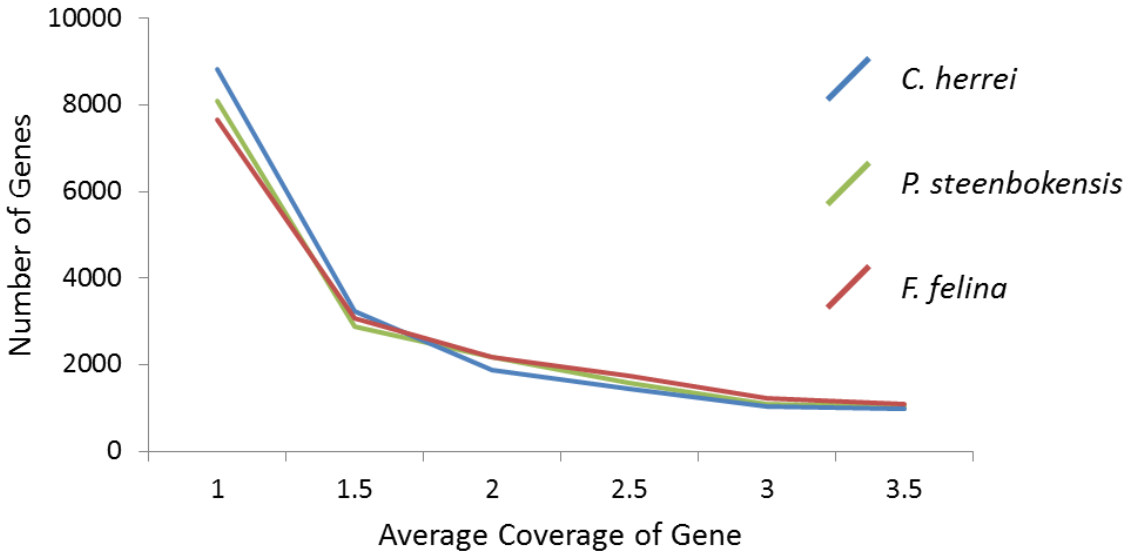


Figure 4.7: Representation of *A. thaliana* genes in three Ruschioideae genomes.

This comparison showed that there has been an increase in gene copy number, on average, in the two Ruschieae species compared to *C. herrei*. But the change in gene copy number was not as large as could've been expected, given the difference in genome size. Approximately one third of genes were found to have significant differences in representation between the Ruschieae genomes and the *C. herrei* genome (Figure 4.8). But of these significantly differentiated genes, only 56%-57% of these genes were higher in the Ruschieae. The corollary of this statistic is that 43%-44% of genes had significantly higher copy numbers in *C. herrei*, in a genome less than half the size. This discrepancy is presumably at least partially a result of the differences in assembly quality. If a gene has two homologues in each genome, but half of the one homologue is unassembled in the Ruschieae genome, the analysis is correct that the representation in the Ruschieae genome is less than in *C. herrei*. This highlights the difficulty of working with a poor quality assembly. It is difficult to determine what is missing. But even taking into account some under estimation in the Ruschieae numbers, this result does seem closer than might be expected. This could suggest that the manner by which genetic information has increased within the Ruschieae does tend to preferentially not duplicate genes. This could be a result of most duplications resulting in deleterious results for natural selection to act upon, or else it could be an intrinsic characteristic of the amplification method.

A better indication of the difference in duplication might be the number of genes that have undergone a twofold (or more) change in copy number. This statistic is not perfect, as the 'copy number' here is talking about the number of homologs in each plant. So if an *A. thaliana* gene already has 3 homologs in *C. herrei* and these genes expanded to 5 homologs in the *F. felina* genome, this expansion will not be counted. So in other words, gene families are going to be underrepresented. It also will not count a gene that was duplicated in the Ruschieae, but does not have both duplicates fully assembled. But at the very least the number of false positives found in *C. herrei* as a result of the poor assembly of *F. felina* and *P. steenbokensis* should be a lot less. Of the genes that had doubled (or more) in copy number, 82% had doubled in the Ruschieae when compared to *C. herrei*.

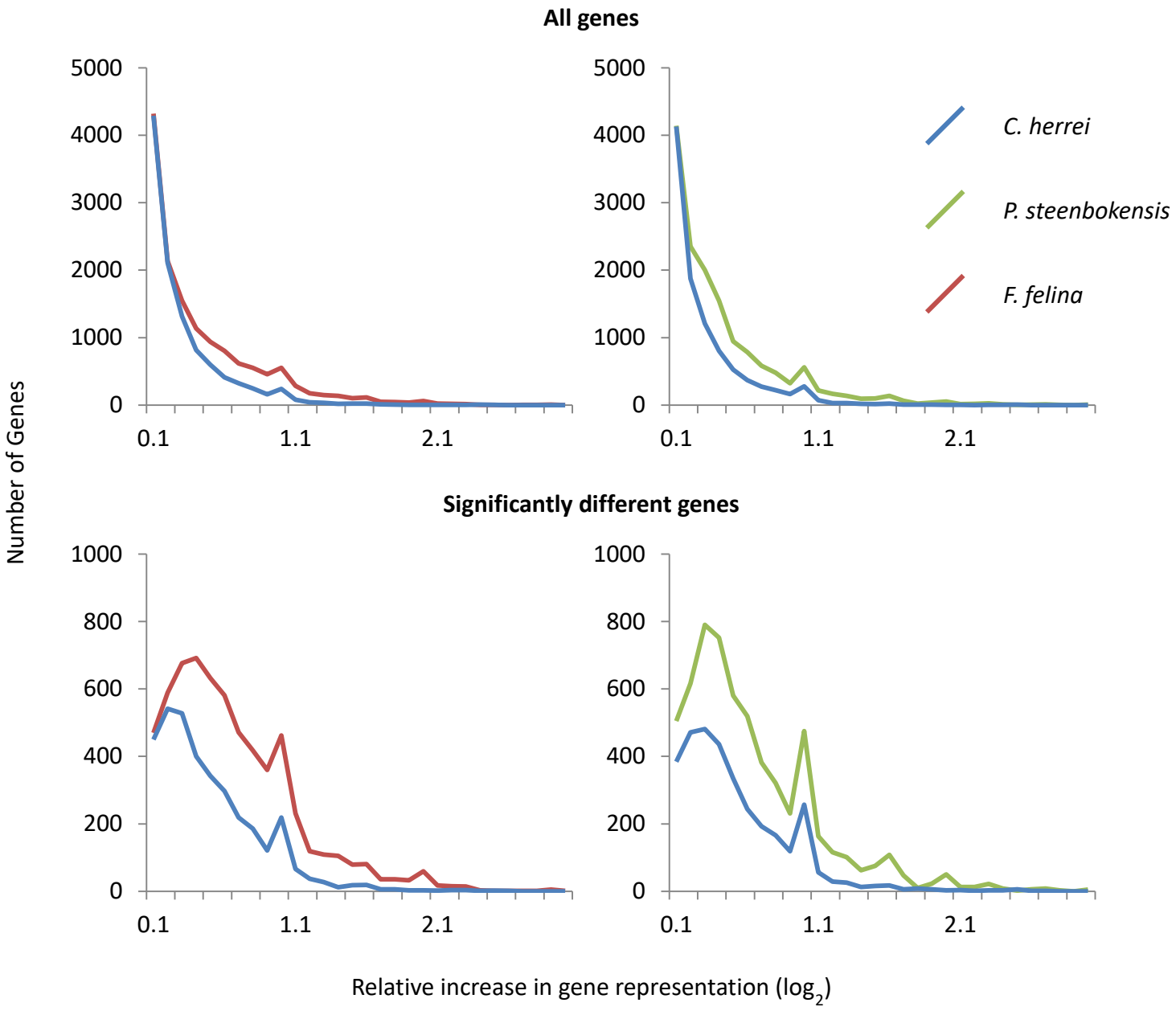


Figure 4.8: Differences in *A. thaliana* gene copy representation between *C. herrei* and the two Ruschieae, *F. felina* and *P. steenbokensis*. Figures comparing gene copy number in *C. herrei* (blue) to *F. felina* (red) and *P. steenbokensis* (green). The two graphs above represent all identified genes while the bottom two graphs only show genes found to be significantly different. The x-axis is on a log scale, with every unit representing a twofold change.

The dataset was not intended to give absolute copy numbers (Figure 4.7). There are several factors that can affect the representation of a gene within a genome, such as number of conserved/divergent features, assembly quality and evolutionary history. These factors mean that it is far better to view a gene in comparison to that gene in the other plants, to control for at least some of these factors (Figure 4.8). But a final thing to note from the absolute numbers of Figure 4.7 is that the modal coverage in each plant is around an average of 1. This means that most genes are in a single copy state. This supports the earlier data that whole genome duplications have not occurred in the Ruschioideae and is not the reason for the increased genome size (or rapid speciation).

Tandem Duplications

If the gene duplications are not the result of a genome duplication, discerning some characteristics of the duplication could be helpful in determining the mechanism of duplication. In this vein, an attempt was made to quantify tandem duplications in the recent duplications of the two Ruschieae plants, compared to *C. herrei*. However, measuring a statistic like this in a fragmented genome is wrought with technical difficulties. You need to be able to discern if a gene has a tandem duplicate, but most of the scaffolds are too short to potentially show it. It may seem rational, for example, to prioritise the longer scaffolds, since that is where you are most likely to be able to identify tandem duplicates. But long scaffolds are also disproportionately less likely to contain duplicated regions since they were, by definition, the easier regions to assemble.

Because this analysis aimed to discover the prevalence of tandem duplicates within the recent duplication events, genes were filtered such that duplicates were only found in one of the plants (i.e. the other two plant species only had one copy). This prevents ancestral duplications adding noise to the analysis. The prevalence of tandem duplicates was then measured as a percentage of these duplicated genes. Genes were counted as a tandem duplicate if 45 of the last 50 amino acids aligned two (or more) times to the same scaffold.

The distribution of the distances between all tandem duplicates from *C. herrei* was calculated and used as the expected distribution for the distances between duplicates in each of the plant species (Figure 4.9). Unfortunately, this distribution does have flaws. It will, for example, underestimate the number of long distance tandem duplicates, due to the fact that the *C. herrei* genome is still not very good, despite it being the best of the three Ruschioideae. But that shouldn't be much of a problem

given how much poorer condition the Ruschieae assemblies were in comparison. Each gene was then weighted according to how likely it was to actually see a tandem duplicate, if one existed, by measuring the distance to the ends of the scaffold from it. So a gene in the middle of a long scaffold was weighted as approximately 1.0. If there was a tandem duplicate, it should have been visible. A gene on the very end of a long scaffold would be 0.5, since there is a 50% chance that a tandem duplicate, if present, would have been on the unassembled side. And a gene with 1000bp on either side of it would have had a weighting of 0.16. The number of tandem duplicates was then compared to this normalised number, representing the number of opportunities where we may have expected to see a tandem duplicate.

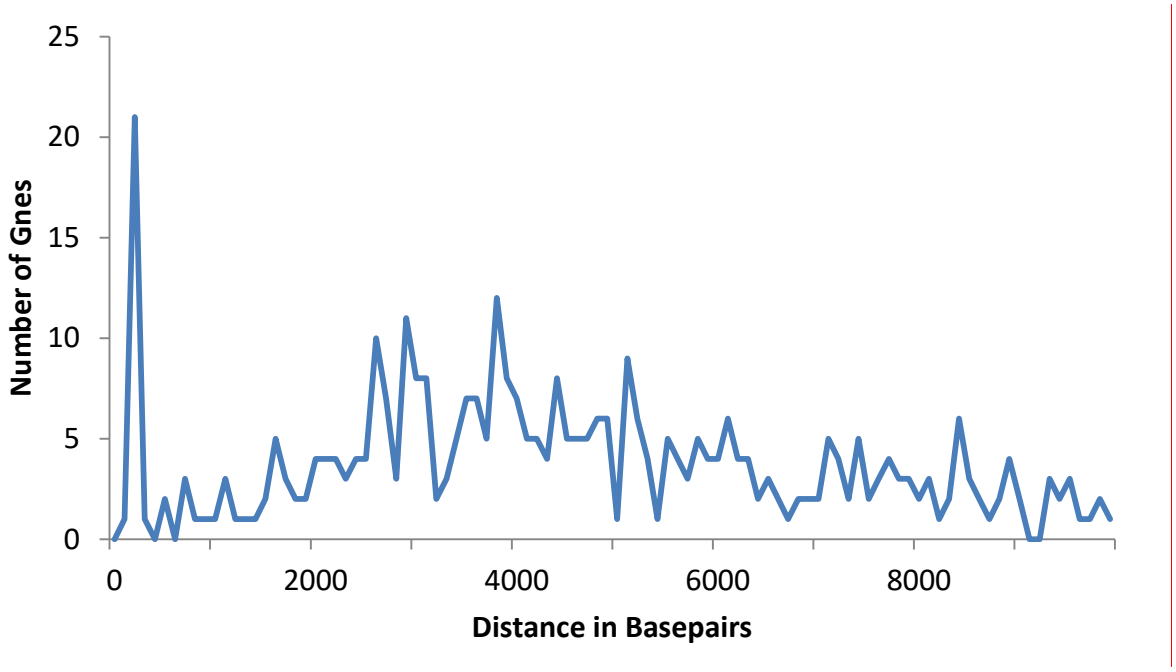


Figure 4.9: Distance between putative tandem duplicates within *C. herrei*. The y-axis represents the number of identified genes at a given distance on the x-axis.

Interestingly, even after taking into account the shorter scaffold size of the two Ruschieae genomes, less tandem duplicates were found in these plants than in *C. herrei*, even though the Ruschieae were found to have more unique duplications in general (Table 4.5).

Table 4.5: Observed frequency of tandem duplications in Ruschioideae species. The number of duplicated genes refers to the number of *A. thaliana* genes that were found to be uniquely duplicated in each plant. The normalised total refers to the effective number of opportunities available to identify tandem duplicates, given the size of the scaffolds.

	C. herrei	*F. felina*	*P. steenbokensis*
Uniquely Duplicated Genes from *A. thaliana*	144	234	197
Resultant Number of Genes	326	508	440
Average Copy Number	2.26	2.17	2.23
Identified Tandem Genes	65	6	4
Normalised Effective Number of Genes	279.7	131.5	117.4
Normalised Percentage	23.2%	4.6%	3.4%

Discussion

The Ruschioideae are an interesting plant system to work with. Their rapid speciation and morphological change suggests that something could be occurring at the genetic level. And the results presented here support that hypothesis, to a degree, despite the difficulties encountered.

In order to investigate the phenomenon of rapid speciation and morphological change, the genomes of three plants were sequenced with Illumina technology. Two of the plants that were sequenced (*F. felina* and *P. steenbokensis*) came from the expansive tribe, Ruschieae, which contains the vast number of species from within the core radiation of the Ruschioideae subfamily (Figure 4.1). Thesespecies were primarily chosen based off their different morphologies and availability of material, but genome size estimates were also taken into account.

Due to the exploratory nature of the project, only short insert sizes were used for the sequencing. Additional, long range sequencing can always be done at a later date if it is deemed necessary.

If the catalyst for the Ruschioideae radiation was a mutation to a key gene, it would probably be difficult to identify. This would probably be the worst case scenario, as far as the prospects for the analysis. Differentiating the mutations potentially responsible for the rapid speciation from the resultant random other mutations would be nearly impossible with the given sample size. And there is a precedent for rapid speciation in plants (Proteaceae) to be accompanied by a high frequency of mutations (Duchene and Bromham, 2013). Fortunately, previous studies taking candidate gene approaches found little genetic divergence between selected genes, making this possibility less likely (Illing et al, 2009; Klak et al, 2013). Instead, multiple instances of duplication were identified.

Whole genome duplications are a relatively common occurrence in the plant kingdom (Riesberg and Willis, 2007). This makes the possibility of a whole genome duplication in the clade fairly plausible. A whole genome duplication copying all the genes of a plant has a lot of potential. It creates redundancy within the genes and allows for one of the copies to remain unchanged while the other diverges in function. If this occurred in a plant with many niches available, the diverging gene functions could allow for a more rapid adaptation to the environment. If this is the underlying cause of the radiation, it should be possible to identify the features of the duplication in the Ruschieae, although it will make the genome assembly more difficult.

But duplications can be caused on a much smaller scale than the entire genome, while still being a large enough duplication to copy entire genes. This can have a similar effect to a whole genome duplication, in that it can create redundancy for natural selection to act upon, but can act in a more nuanced way, where genes that are detrimental when duplicated are not copied along with genes that are advantageous. These smaller duplications can be caused by a variety of factors, such as transposable elements or uneven crossing over during meiosis. Differentiating between these mechanisms might be difficult, but there are certain patterns that can be identified. For example, active transposable elements might occur in high copy numbers scattered throughout the Ruschieae genomes, while unequal crossing over results in tandem local duplications.

The first thing to note is how many of the future problems and results from the Ruschieae were visible in the k-mer frequency plots (Figure 4.4). These plots immediately showed that the genome assemblies would be tough, with the (relatively) large expected genome size, the high level of heterozygosity, and the proportion of duplicated sequences visible. *C. herrei*, in comparison, showed a visible 4n peak, predicting some level of duplicated sequences, but no heterozygosity to speak of and a smaller genome size in general.

These expected differences were borne out in the genome assemblies, with *C. herrei* assembling far better (Table 4.2), and with substantially less effort than the two Ruschieae genomes. To an extent, this difference in ease of assembly can be viewed as a result in and of itself. Something has happened at the genomic level that makes the Ruschieae substantially harder to work with. This is not just a few select nucleotide changes, as appeared likely from earlier work (Illing et al, 2009).

The poor state of the two Ruschieae genomes would go on to define the future methodology that was used in the study. Methods had to be cognizant of assembly bias, genes being dispersed across multiple scaffolds, and unassembled/missing sequences.

No Whole Genome Duplications

The first thing that needs to be established is that there were no whole genome duplications in any of the three plants' recent history. This is despite the large change in genome size between the Ruschieae and *C. herrei*. This result was hinted at in the k-mer frequency plots (Figure 4.4), which lacked a convincing 4n peak. But this was not conclusive since a genome duplication could have occurred and since had the duplicate sequences diverge. There was also a small 4n peak in *C. herrei*, indicating some form of duplication.

The lack of a recent whole genome duplication is shown more convincingly in Figure 4.7, which suggests that most genes are present in a single copy state. Additionally, Table 4.4 shows that when genes are duplicated, the point of duplication is inconsistent, with the majority (but not all) of the duplication events occurring before the divergence of *C. herrei*. All of this suggests that, despite the large difference in genome sizes, there was no whole genome duplication in any of the plants.

Duplicated Genetic Sequence

The story of the Ruschieae appears to be one of local genetic duplication. This was demonstrated in a multitude ways, from the expanded genome sizes (Table 4.1), the elevated presence of high frequency k-mers (Figure 4.4), the proportion of reads which mapped multiple times to the genome (Figure 4.5) and the proportion of duplicated genes (Figure 4.8).

The nature of the majority of the duplicated material is not known. The duplicated sequences that are present in the Ruschieae genomes in moderate copy numbers (Figure 4.5) are where the largest

differences were observed, in terms of frequency, when compared to *C. herrei*. But these regions were not able to be effectively investigated. These duplications could represent more repetitive elements, such as those which were identified, just in lower copy number or even just random dna. But the presence of duplicated sequences, particularly duplicated genes, is of importance in the Ruschieae, given the high rates of speciation and morphological change. This is a source of fresh genetic material for natural selection to act upon, potentially making the generation of novel morphological features possible without changing existing genes and gene networks.

Presence of Repetitive Elements

The repetitive element expansion, such as the long terminal repeats shown in Table 4.3, suggests that there has been transposable element activity in the recent past of all the Ruschioideae species. This is unfortunately rather ambiguous as far as a possible explanation for the Ruschieae genome changes and rapid speciation. Transposable elements have been known to effect gene copy number (Jiang et al, 2004) and function (Lisch, 2013). So this could explain the rapid morphological change and the observed duplication patterns (Figure 4.5 and Figure 4.8). But, contradicting this idea is the fact that *C. herrei* was found to have similar/higher levels of repetitive element representation (Table 4.3). It seems intuitive that if transposable elements were the cause of the duplication signature in the Ruschieae, a larger percentage of the genome would be the long terminal repeats that the transposable elements left behind. Therefore, the overall role of transposable elements remains unclear.

Lack of Tandem Duplications

Tandem duplications can arise from processes such as unequal crossing over. A disproportionately high occurrence of these duplicates could shed some insight into the mechanism of duplication within the Ruschieae. But the data shows that there are fewer recent tandem duplicates in the Ruschieae species than in *C. herrei* (Table 4.5). This could mean that the primary mechanism of duplication within the two plant clades differs, which would be very interesting. But it could also show that the intergenic distance has grown, with the increase of genome size in the Ruschieae, since their divergence from *C. herrei.* Unfortunately, given the quality of the genome assemblies, these results should be viewed as preliminary. As has been stated before, duplicated regions are

often difficult for genome assembly algorithms to accurately assemble and it is possible that there is some form of assembly bias present (for example Vukašinović et al, 2014).

Conclusion

The results from this section point to a lot of activity within the Ruschioideae genomes: variation in genome size, large scale duplication of sequences and recent transposable element activity. Unfortunately, the nature of this genomic activity has made it very difficult to get an accurate picture of the genome as a whole. This makes analysis and drawing concrete conclusions tough. Duplication is occurring within the genomes of the Ruschieae, but it is unclear what is causing it. It is also unclear whether the duplications are what caused the rapid speciation in the clade. It may be necessary to use long read sequencing technology in the future to improve the genome assemblies. But until then, it looks like the Ruschioideae are a very interesting, but very complicated genomic nut to crack.

Chapter 5: *Xerophyta humilis* - Ploidy Change and Somatic Mutation

The transition from water to land had many challenges for the early plant pioneers. One of the major new threats associated with terrestrial life was the risk of running out of water. Initially, this meant that a plant needed to be able to survive desiccation (i.e. it needed to be desiccation tolerant) (Fisher, 2008). But, with the evolution of more complex traits, such as tracheophytes, this strategy of desiccation tolerance was reduced to only occur in the seed.

This specialisation of tissue, with desiccation tolerance being limited to the seed, corresponded with the gymnosperm and angiosperm domination of the terrestrial world. But along with the domination of the terrestrial world, came exposure to a wide variety of possible niches for plants to occupy. Occasionally, these new niches facilitated the expansion of desiccation tolerance back into the vegetative tissue (Alpert, 2005).

The Velloziaceae are a family of monocotyledonous plants that has several vegetative desiccation tolerant members. These plants are capable of completely drying down and maintaining that state before rehydrating when water does become available. This includes many African species from the genus Xerophyta. An example of such a plant is *Xerophyta humilis* (Illing et al, 2005). It is a small plant commonly found in regions prone to variable rainfall in South Africa, Zimbabwe, Botswana and Namibia (Figure 5.1).

It has been argued that *X. humilis* leaves and roots have co-opted the genetic tools that allow embryos in plants seed to desiccate without dying (Illing et al, 2005). However, a key question is how these seed maturation genes are activated. One approach to investigating this question is to identify the desiccation transcriptome in drying seeds, seedlings and leaves. An assembled genome would aid in this endeavour. Additionally, an assembled genome would allow for analyses such as identification of promoters from genes of interest to check enrichment of motifs. The aim of this chapter was therefore to assemble the *X. humilis* genome and facilitate the research.

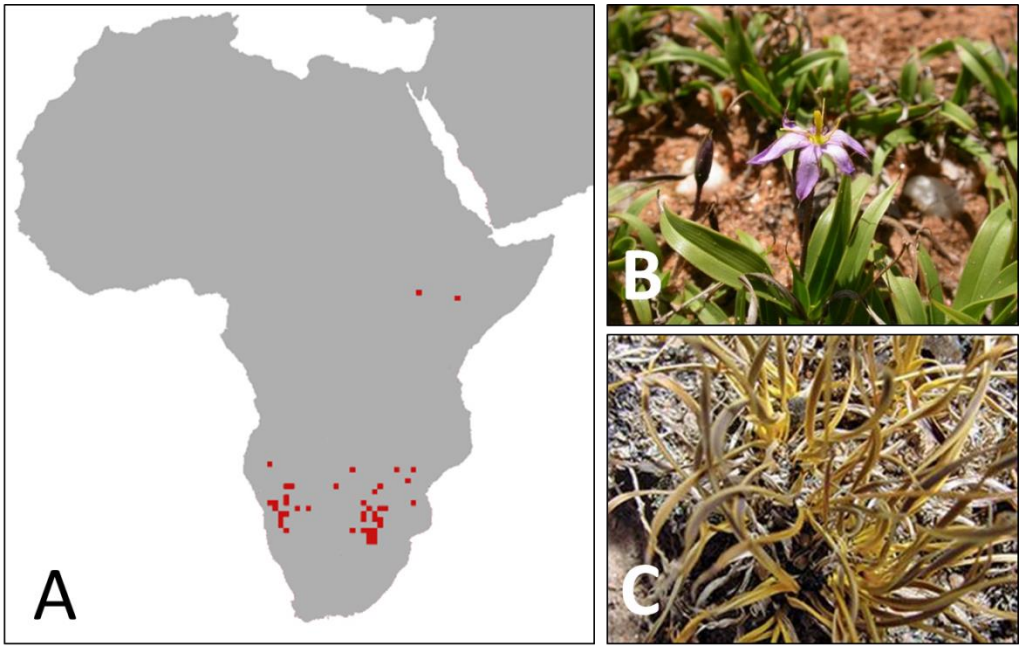


Figure 5.1: *Xerophyta humilis* is widespread in southern Africa. A) Recorded sites for *X. humilis* (Foden and Potter, 2005). B) *X. humilis* occurs in mats, flowering when water is abundant. C) When water is scarce, *X. humilis* can survive for long periods of time in a desiccated state.

Plant Genome Plasticity

As was discussed (and shown) in Chapter 4, plant genomes are abnormally susceptible to change. Genome duplications, transposable elements and small scale duplications are all common occurrences (Paterson et al, 2010; Wang et al, 2012; Springer et al, 2018). This potential for change, on both large and small scales, means that it is difficult to predict *a priori* what problems, if any, will be encountered when sequencing a genome *de novo*.

The genome size of *X. humilis* is estimated to be 532 Mbp (Hanson et al 2001). This is larger than *Cleretum herrei* (see Chapter 4) but, problems such as recent genome duplications and high heterozygosity notwithstanding, 532 Mbp should still have been manageable. However, the sequencing data, generated from leaves that had been through many cycles of desiccation, had unusual, unexpected problems. Novel analytical methods had to be devised to deal with what appeared to be exceptionally high levels of somatic mutations. Despite these challenges, the *X. humilis* genome was assembled, and this assembly was used to characterize the pattern of

mutation. The data generated suggests that repeated cycles of desiccation in these plants comes at a high cost of somatic mutation.

Methods

DNA isolation and sequencing

Leaves were harvested from several mature *X. humilis* individuals forming an intertwined ground cover mat of approximately 150cm^2 that were collected from the Barakalola Nature Reserve in the North West Province (North West Provincial Government Permit 062 NW-12; Cape Nature Permit AAA007-01733). DNA was isolated from the samples using a nuclei extraction step in order to reduce chloroplast contamination (Lutz et al, 2011), followed by RNAse treatment and purification by column (Qiagen Genomic-tip 100/G). DNA was then sent to the Beijing Genomics Institute (BGI) for library construction and to be sequenced. Sequencing was conducted in two rounds. The first round aimed to generate 26 Gbp of data (approximately 50x coverage) across 3 libraries (Table 5.1). Initial analysis suggested that the coverage was far too low, so an additional 90 Gbp of sequencing was ordered.

Table 5.1: The planned amount of sequencing for each library from BGI. The sequencing was conducted in two phases. A lane of sequencing is up to 200 million pairs of reads.

Insert Size	Gbp of Sequencing	
	1st Round of Sequencing	2nd Round of Sequencing
170bp	12	0
500bp	8	45
800bp	6	45

Read Processing for Genome Assembly

The sequencing quality of the reads was checked using Fastqc (version 0.10.1) (Andrews, 2010). Trimmomatic (version 0.32) (Bolger, 2014) was used to trim the bases with a quality less than 17 from either end of the read. Reads shorter than 60bp were discarded.

After trimming, a k-mer frequency plot was created from the reads using KmerFreq_HA (version 2.01) (Luo et al, 2012), first with the initial round of sequencing and then all the data combined. Corrector_HA (version 2.01) (Luo et al, 2012) was used to aggressively remove low frequency k-mers. In addition to the standard k-mer frequency cut off of 3, higher values were tested, namely 45, 50, 55, 60 and 90, based off the expected heterozygous coverage of 110[16].

After trimming, a k-mer frequency plot was created from the reads using KmerFreq_HA (Luo et al, 2012), first with the initial round of sequencing and then all the data combined. Unlike previous analyses, which had the expected k-mer frequency pattern of a unimodal, or bimodal peak with a shoulder on the left from sequencing error (Figure 1.7 and Figure 4.4 from previous chapters), neither of these k-mer frequency plots had a peak. As a result, multiple high values were used during optimisation for what k-mer frequency should undergo error correction using Corrector_HA (Luo et al, 2012). The tested values were 3, 45, 50, 55, 60 and 90 and were chosen based off the expected heterozygous coverage of 110. This aimed to reduce the low coverage k-mers in the dataset, which would make the assembly process a lot less complicated.

Genome Assembly

The processed datasets were assembled using SOAPdenovo2 (version 2.04) (Luo et al, 2012). Assembly quality was assessed based off the assembled genome size and the N50 scaffold size. Other than the error corrected dataset, the k value and bubble merging stringency were the main parameters that were altered. GapCloser (version 1.12) was then used on the best assembly to reduce ambiguous regions between contigs.

[16] Estimated by dividing the amount sequencing generated by the genome size

RNA Samples

RNA sequencing data from *X. humilis* was used from Lyall et al (in review). This RNA data was either generated from mature plants that had been harvested from the field or else young plants that had been grown from seeds in the lab (Table 5.2). The plants collected from the field had undergone many cycles of desiccation and rehydration, whereas the young laboratory grown *X. humilis* had never been desiccated.

Table 5.2: Summary of RNA sequence data used from Lyall et al (in review). Samples from the field are assumed to have undergone multiple cycles of desiccation and rehydration. Samples grown in the lab were never allowed to desiccate.

RNA sample	Million reads	Read Length (bp)	Desiccated	Replicates
Leaf (Field)	40	90	Yes	3
Seed (Field)	40	100	Yes	3
Leaf (Lab)	80	100	No	1
Root (Lab)	80	100	No	1

Checking Coherency of the Assembled *X. humilis* Genome

In order to check that the genome assembly method was successful in assembling something coherent, the draft transcriptome assembly, was aligned to the genome using Blastn (version 2.2.29) (Altschul et al, 1990). This transcriptome assembly was assembled independently (Lyall et al, in review), and included fungal contamination.

Identifying Codons

The predicted open reading frames of the assembled transcripts (Lyall et al, in review) were aligned using Tblastn (version 2.2.29) (Altschul et al, 1990). Codons were only identified and used within the

same exon as the start codon and within regions that had a genomic coverage between 60x and 400x.

Aligning Reads to Assembly

The method used to filter the raw data prior to assembly was highly unusual. To prevent propagating any error, these error corrected reads were not used for future alignment and analysis. Instead, the raw reads were quality trimmed with Trimmomatic (version 0.32) (Bolger et al, 2014), with a minimum quality score of 20 required, before alignment with either Bowtie2 (version 2.2.1) (Langmead et al, 2008) or Tophat2 (version 2.0.13) (Kim et al, 2013) on default settings for genome and RNA sequencing respectively.

Alternate Allele Frequency

The alternate allele frequency within the genome was calculated using the Samtools (version 1.3.1) (Li et al, 2009) mpileup function, with the flags 'D' and 'f'. This prevents any automated snp calling from Samtools. Snp calling was turned off in this manner to get a better understanding of the underlying features responsible for the k-mer frequency plot. In the interest of simplification, indels and sites which contained an 'N' as the reference were excluded.

Results

Sequencing

The sequencing from BGI worked well, with large amounts (Table 5.3) of high quality sequencing (Figure 5.2) being generated. The majority of reads came from the second round of sequencing and resulted in abundant coverage for assembly.

Table 5.3: Resultant coverage after 2 rounds of Illumina sequencing. Total coverage is estimated using a genome size of 532 Mbp.

Insert Size	Read Pairs After Processing (millions)		Coverage
	1st Round of Sequencing	2nd Round of Sequencing	
170bp	54.7	0	20.6x
500bp	47.1	305.2	132.4x
800bp	39.0	182.4	83.2x
Total Coverage			236.2x

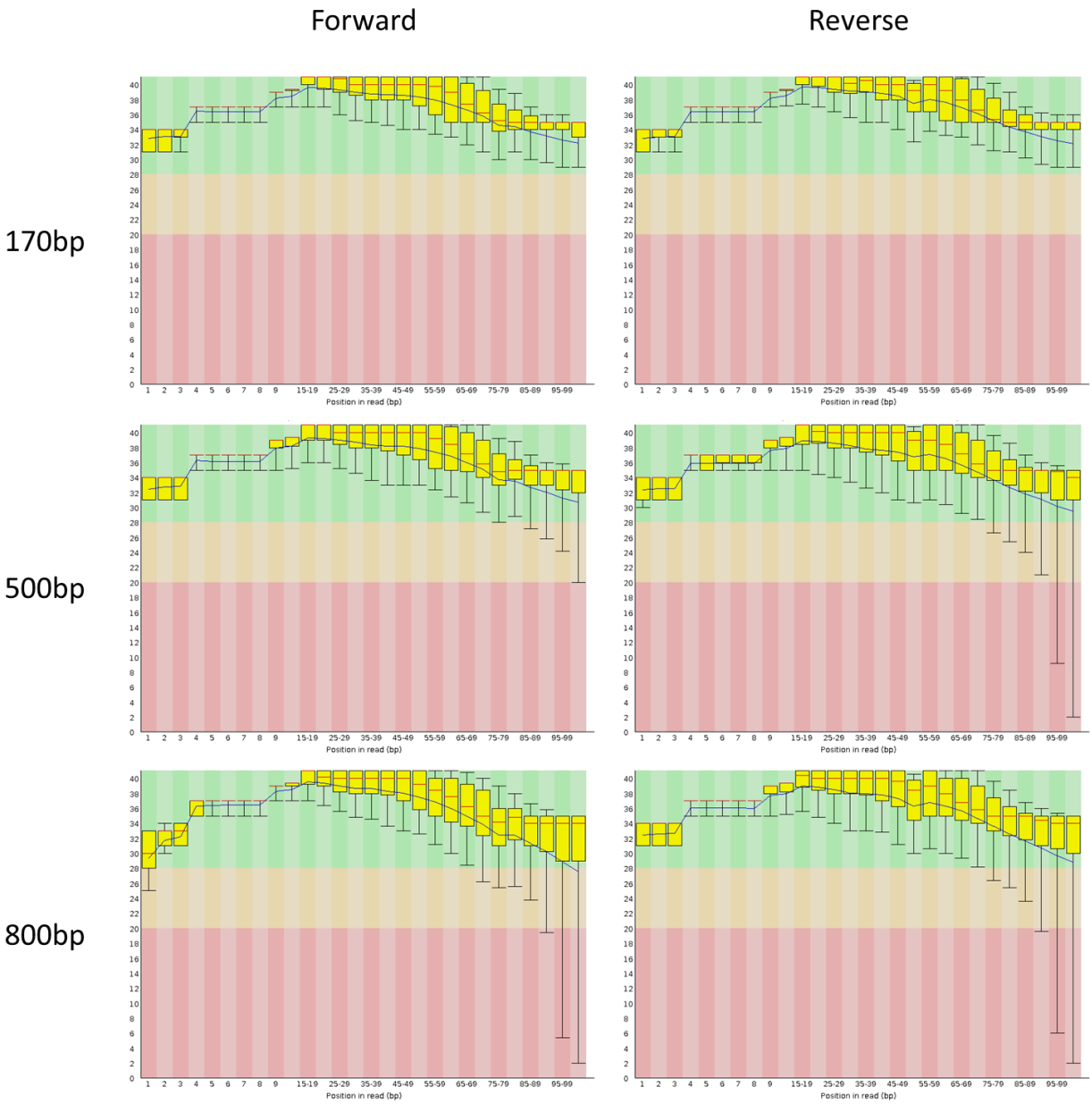


Figure 5.2: Quality scores for each data across the raw reads. The red line in the box plot represents the median value, the inner and outer quartiles are shown by the yellow boxes and the 10%/90% values are represented with the error bars.

Unusual K-mer Frequency Plot

Unlike previous analyses, which had the expected k-mer frequency pattern with a unimodal distribution, or bimodal peaks, with a shoulder on the left from sequencing error (Figure 1.7 and Figure 4.4 from previous chapters), the *X. humilis* data did not have any peak (Figure 5.3). This was true for the original round of sequencing and once the additional round of sequencing was added.

The abundance of low frequency k-mers could be caused by a large amount of somatic mutations and would interfere with the genome assembly. As a result, multiple high values were used during optimisation for what k-mer frequency should undergo error correction using Corrector_HA (Luo et al, 2012). This aimed to reduce the low coverage k-mers in the dataset, which would make the assembly process a lot less complicated.

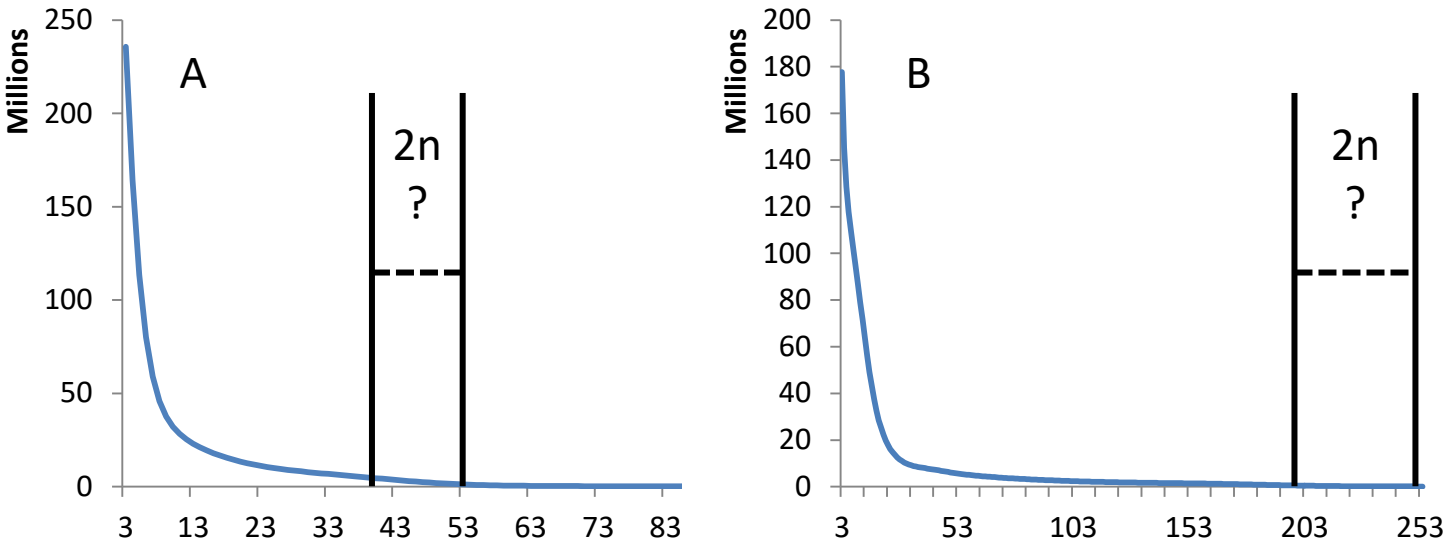


Figure 5.3: K-mer frequency plots created from *X. humilis* whole genome sequencing. A) The initial round of sequencing did not have a homozygous peak in the expected area. This could've suggested that the genome size estimate was wrong and the actual genome size was a lot larger than expected. B) After substantially more sequencing coverage, there is still no visible homozygous peak.

Genome Assembly and Read Error Correction

The genome assembly used the reads which had undergone error correction with k-mers of a frequency of 50 or lower. This error correction process removed a large number of the low frequency k-mers and eliminated a large percentage of the reads (Figure 5.4) in order to simplfy the *de Bruijn* graph for assembly.

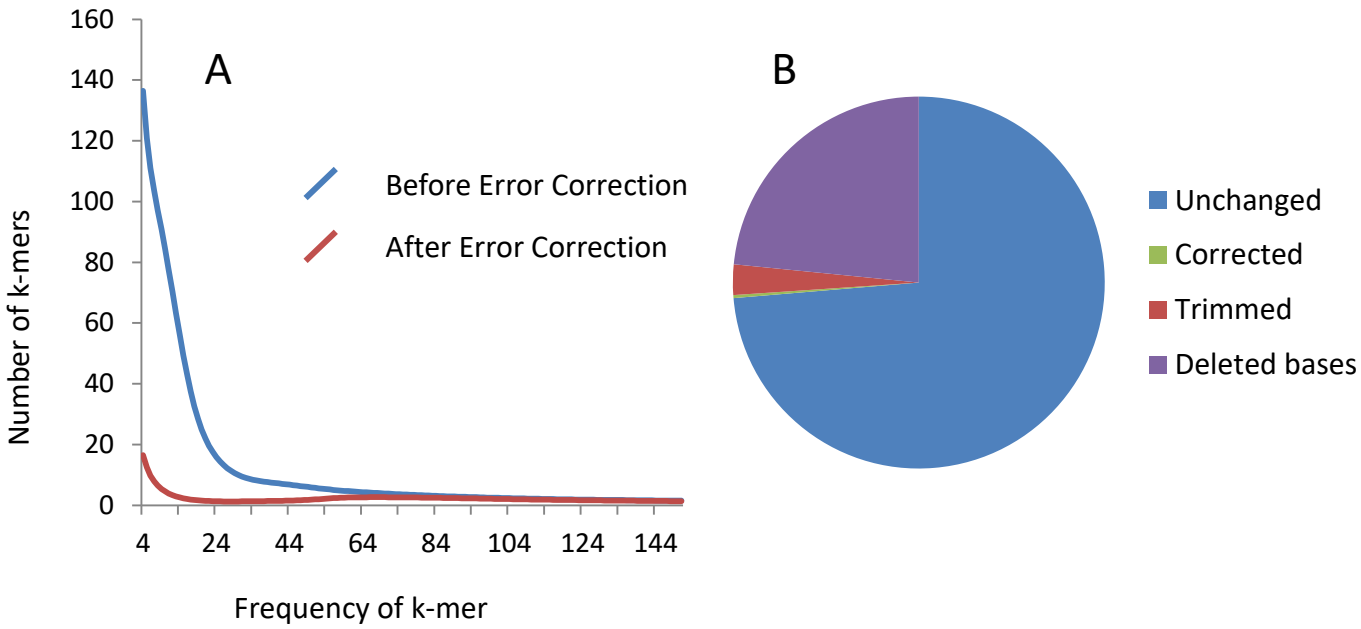


Figure 5.4: Results of error correcting k-mers with a frequency of 50 or less. A) Reduction in low frequency k-mers. B) Fate of bases after error correction. "Trimmed" bases were removed from the end of the read, whereas a "deleted" base was removed along with the rest of the read.

All of the genome assembly results were very poor. The best assembly (Table 5.4) used a k-mer size of 27, which is small, especially for such abundant coverage. This assembly, like all the assemblies, had a very low N50, and an exceptionally high percentage of N's. The gap closing algorithm had a substantial effect, both in shortening the N50 and reducing N's.

Table 5.4: Genome assembly statistics of *X. humilis* before and after gap closing. The initial assembly was of very poor quality. The highly fragmented and gap filled assembly was vastly changed by the gap closing program.

	Original	Gap Closed
Size:	445 Mbp	359 Mbp
Longest Scaffold:	108 kbp	80.9 kbp
Scaffold N50:	10 461 bp	7 490 bp
Ns:	52.15%	16.87%
Contig N50:	328 bp	1 742 bp

Checking Coherency of the Assembled *X. humilis* Genome

Transcripts assembled from RNA sequencing (Table 5.2) were aligned to the assembled genome using Blastn (Altschul et al, 1990). Of the transcripts that were predicted to come from a plant, 99.7% aligned, at least partially, to the genome. This is especially impressive considering how little of the genome was actually assembled. In comparison, only 2.7% of the Fungal transcripts aligned to the genome assembly.

Mapped Reads to Genome

The genome sequencing reads were aligned back to the genome assembly. Despite the unusual k-mer frequency plot (Figure 5.3), the coverage of the reads across the assembly did have distinct peaks (Figure 5.5). This suggests that, at the least, something coherent was assembled. Curiously, there are 3 peaks in the coverage. The major peak, at 202x, is consistent with a homozygous peak at the expected genome size. This peak is complimented by a peak at about 100x, which is consistent with a heterozygous haplotype. The third and final peak is at 40x. This could be the result of the sequencing coming from multiple individuals or some form of contamination. Alternatively, it could indicate a recent genome duplication, taking the place of the heterozygous haplotype peak (which would be expected at 50x). In this scenario, the 100x peak would actually represent the homozygous peak (of a 1TB genome) and the 202x peak represents a tetraploid sequence, where the duplicated DNA is too similar and has been merged into a single scaffold by the genome assembly process, giving it double the expected coverage.

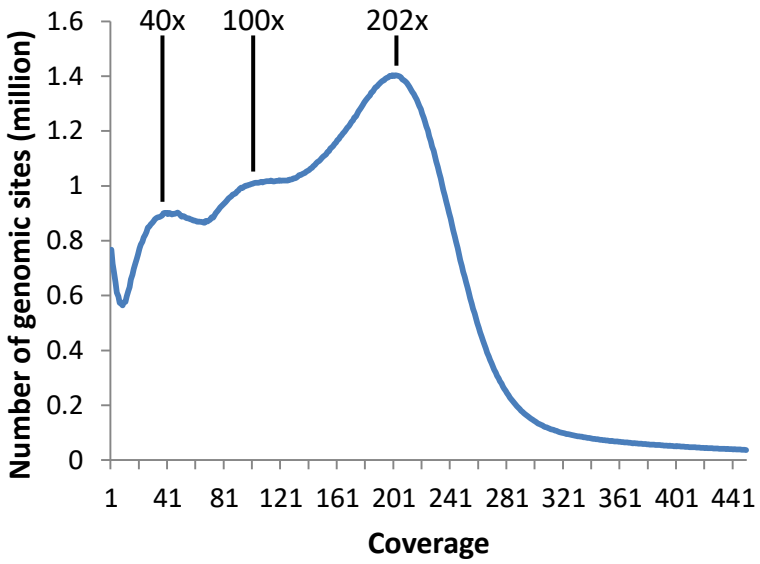


Figure 5.5: Coverage of mapped sites across the assembled genome. Three peaks are clearly visible, at 202x, 100x and 40x. Two peaks are consistent with a homozygous and heterozygous sequence. A third peak is unexpected and could suggest a tetraploid genome.

Alternate Allele Frequency & Identification of High Levels of Somatic Mutations

The alternate allele frequency of aligned reads was calculated for genomic sites with a coverage between 60 and 400 (Figure 5.6). This encompasses the genomic regions covered by the presumptive heterozygous and homozygous peaks, but excludes the peak that is of uncertain origin at 40x coverage (Figure 5.5). Under normal circumstances, the alternate allele frequencies should result in a peak at the 50th percentile, with half of the reads from heterozygous sites having the reference allele and half having the alternate allele. If the genome is tetraploid instead of diploid, there should be peaks present at 25%, 50% and 75%. But instead of either of these options, there are no peaks, and the graph is more reminiscent of the k-mer frequency plot (Figure 5.3). This pattern was reproduced using the RNA-seq data from dehydrating leaves and seed pods from plants grown in the field (Figure 5.7). These low frequency alternate alleles are at too high a coverage to be sequencing error and could represent a large amount of somatic mutations.

In comparison, the pattern of low frequency alternate alleles was not reproduced in the RNA-seq data from the plants grown in the lab, which had never desiccated. Instead, this data showed peaks

at the 25%, 50% and 75% regions, suggesting that the genome assembly does in fact represent a merged tetraploid genome.

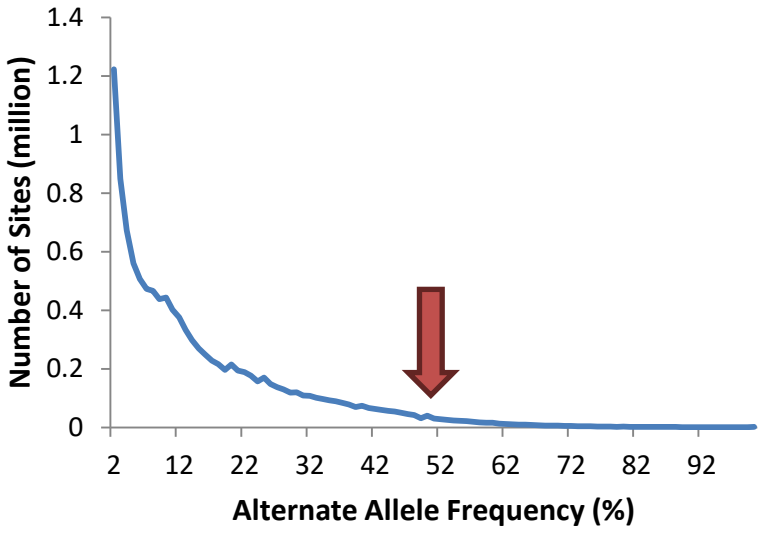


Figure 5.6: Alternate allele frequency of genomic sites with a total coverage between 60 and 400. Normally, a heterozygous peak should be found at the 50% range (red arrow). Reads originate from plants grown in the field which have undergone many cycles of desiccation and rehydration.

Given that the genome appears to be tetraploid, the three genomic coverage peaks observed in Figure 5.5 are presumably the 4n, 2n and 1n peaks. If that is indeed the case, then it will probably be better to look at each peak in isolation (Figure 5.8). These peaks represent merged duplicated sequences (4n), correctly assembled diploid sequences (2n) and heterozygous haplotypes which have been erroneously assembled independently. Each coverage region should therefore, under normal circumstances, have a different peak pattern. The 4n peak (180x-220x) should have alternate allele frequencies, around 25%, 50% and 75%. The 2n peak should have an alternate allele frequency of 50% and the 1n peak shouldn't have any alternate allele frequencies.

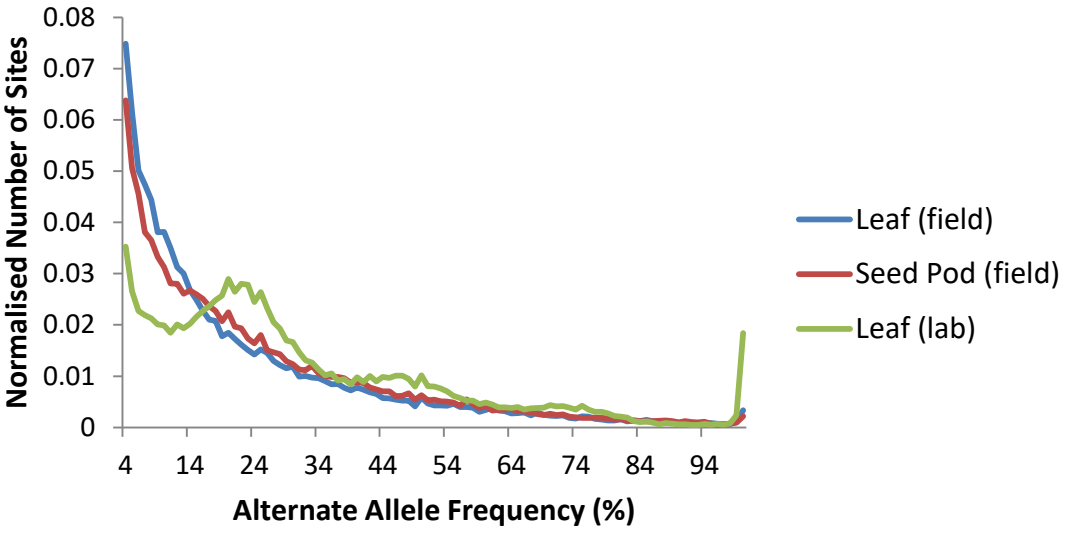


Figure 5.7: Alternate allele frequency of RNA-seq data from genomic sites with a total coverage between 60 and 400. All RNA had a coverage of at least 100 reads at these sites. RNA from plants grown in the field showed no heterozygous peak. The RNA from plants grown in the lab (which had never desiccated and rehydrated) had peaks at 25, 50 and 75% frequency.

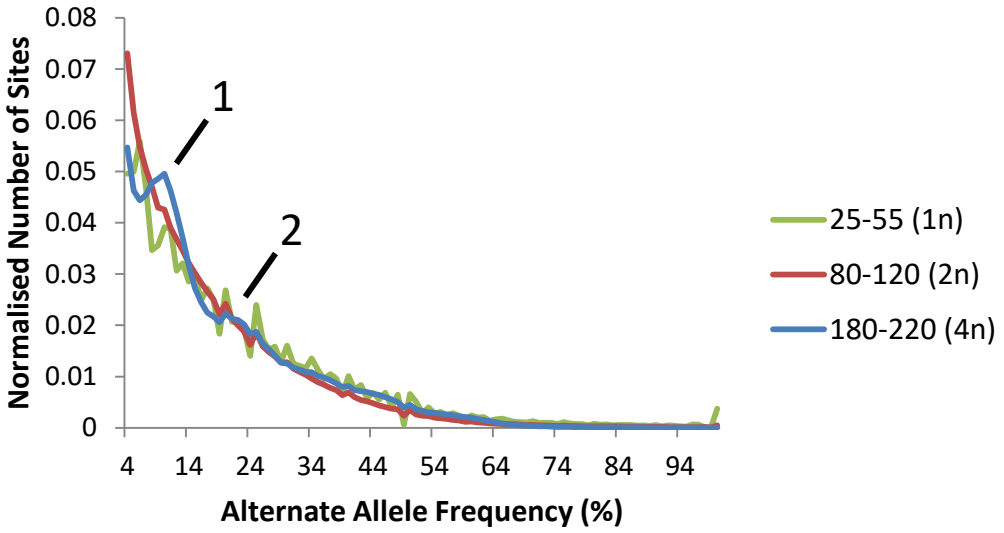


Figure 5.8: Alternate allele frequency of genomic sites from three chosen coverage ranges. There does not appear to be a peak in the 25-55 (1n) or 80-120 (2n) coverage range. The 180-220 (4n) coverage range appears to have two visible peaks. The first peak (1) is at the 10% frequency and the second, smaller peak (2) is around the 25% frequency range.

Instead of the expected alternate allele frequency peaks, the only visible peaks are in the 4n copy range (180x-220x coverage). But these peaks were not at the expected frequencies. Instead, the peaks are present at the 10% and 25% frequency range. It is also peculiar that neither of the peaks present in the 4n range are present in the 2n range. The alternate allele peaks from the 4n region also weren't present in the RNA-seq data from those regions (Figure 5.9).

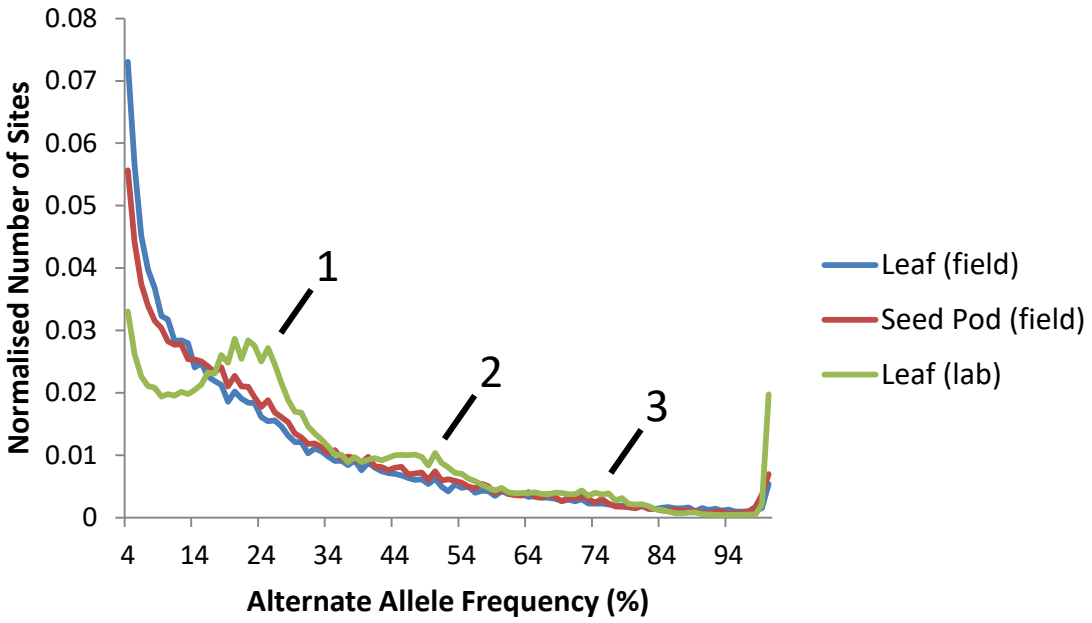


Figure 5.9: Alternate allele frequency of RNA-seq data from genomic sites with a total coverage between 180 and 220. RNA from leaf and seed pods from the field showed no heterozygous peak. The RNA from tissue that was grown in the lab and had never undergone dehydration showed 3 peaks at 25, 50 and 75%.

The alternate allele frequency generated from RNA-seq in the more focussed 4n regions (Figure 5.9), reproduced the pattern generated more broadly (Figure 5.7). There are no peaks in the leaf and seed pod RNA from field grown plants and the lab grown leaves had the same 25%, 50% and 75% peaks as before. There is no trace of the 10% peak that was observed in the genomic data (Figure 5.8) in any of the RNA-seq samples.

Verification of Genome Duplication

To further test the hypothesis that the genome had undergone a recent duplication, sites were found in the RNA data from the lab grown plants which had 3 or more alleles, with each allele having at least 10 reads. These sites should not exist (other than sequencing error) if the genome was not duplicated. In total, 499 of these sites were identified. Of these, 159 were obviously the result of sequencing error, with the frequency of the primary allele being >95%. The remaining 340 sites were compared to a null distribution of randomly generated frequencies for 3 alleles (Figure 5.10).

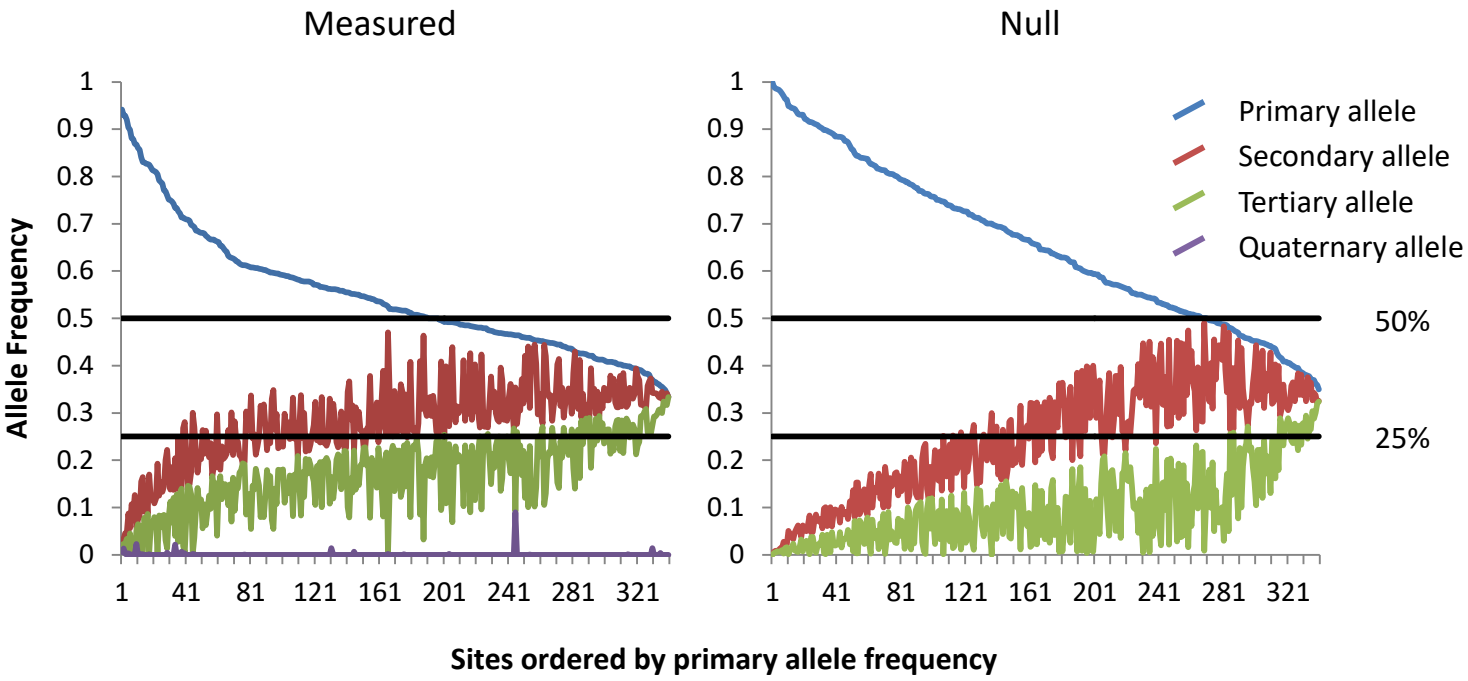


Figure 5.10: Allele frequencies within 340 sites with 3 or more alleles compared to a random dataset. The presence of sites with 3 or more alleles is a sign that the genome assembly process combined recently duplicated genomic sequences. It is predicted that sites of this character would have the primary allele around the 50% range and the secondary and tertiary alleles at the 25% frequency.

These 340 sites appeared to be associated with the 50% and 25% frequencies (Figure 5.10A), when compared to the null distribution (Figure 5.10B), as would be expected under the hypothesis that the three alleles are coming from sites representing 4 strands of DNA instead of the expected 2. This association was confirmed by classifying the allele frequencies at each site into 4 groups:

Homozygous sites with sequencing error (100, 0, 0), heterozygous sites with sequencing error (50, 50, 0), heterozygous site within duplicated sequences with sequencing error (75, 25, 0) and heterozygous sites in each duplicated area / heterozygous site in diverged duplicated areas (50, 25, 25). This final combination is the unambiguous pattern of frequencies expected from these sites which would suggest a merging of duplicated regions. Each site was classified as being closest to the allele frequency pattern of either 100/0/0, 75/25/0, 50/50/0 or 50/25/25 by squaring the difference between the observed and expected values and then seeing which pattern had the lowest combined value. The comparison of the classification process to the random dataset is shown in Table 5.5. As predicted, there are significantly more cases ($p<<0.001$, chi-square test) of the final allelic pattern than would be expected by chance.

Table 5.5: A comparison of the allelic patterns in 340 isolated sites with 3 or more alleles in the lab grown Xerophyta RNA-seq dataset compared to a random dataset. The genome duplication hypothesis predicts the existence of these sites in a 50:25:25 ratio.

Allelic Pattern	Measured	Null
100, 0, 0	11	53
50, 50, 0	31	51
75, 25, 0	61	135
50, 25, 25	**237**	**101**

Accounting for Sequencing Error

Under normal circumstances, real alternate alleles represent heterozygous sites. This means that real alternate alleles have a frequency of about 50% of the coverage, while sequencing errors have a much lower frequency, with a coverage of about 1 or 2 reads. Thus, it should be quite easy to differentiate sequencing errors from real alternate alleles at sites with high coverages. However, the current data does not appear to have an obvious delineation between real alternate alleles and sequencing errors, making this classification more complicated.

In order to get a sense of how the aligned data looks with minimal sequencing error at a site, a stringent filter was applied where a site's alternate allele wasn't counted unless the average sequencing quality score for the nucleotides was over 30 (this is a 1 in 1000 chance of an error). This threshold was chosen based off the difference in quality score distributions between alternate alleles with 1 read coverage and alternate alleles with 2 to 4 reads coverage (Figure 5.11).

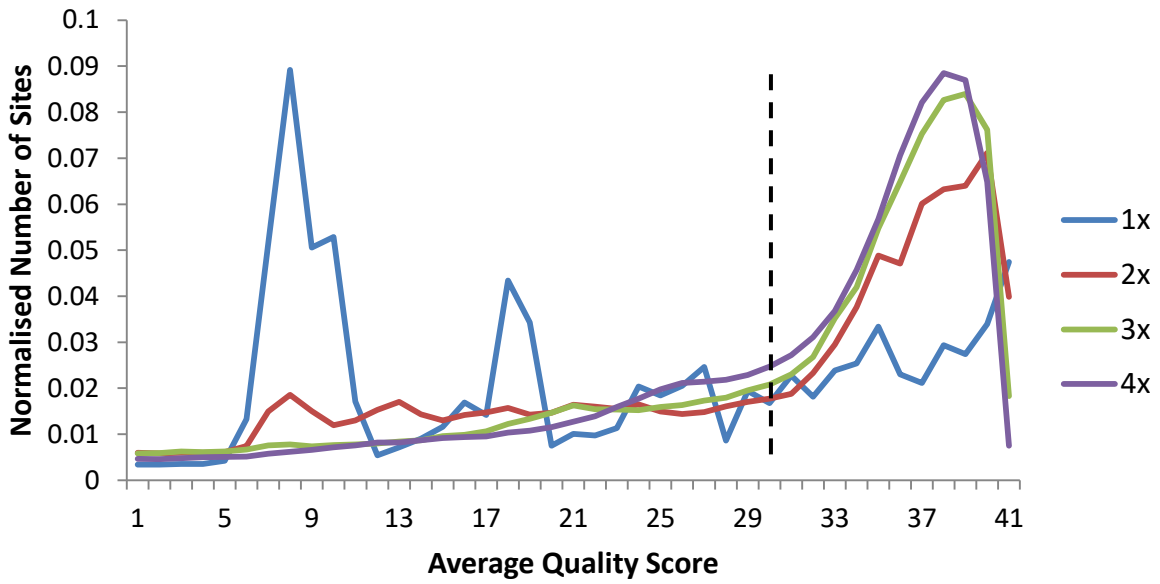


Figure 5.11: Average sequencing quality of alternate alleles with 1-4x coverage at genomic sites. The dotted line shows the threshold cut off selected. Alternate alleles to the left of the dotted line were filtered out.

When looking at the genome sequencing from plants grown in the field, 10.3% of sites had an alternate allele before this threshold quality score was applied. Once the stringent threshold was applied, excluding sites with an average quality score below 30 in the alternate allele reads, the total number of sites with alternate alleles dropped to 5.9% (Figure 5.12). The majority of filtered sites had a single read coverage over the alternate allele, which is to be expected when eliminating sequencing error. In total, 67.7% of these 1x sites were filtered. But, the overall shape of the distribution did not change much with the additional filtering. This suggests that the unusual alternate allele frequency distributions (Figure 5.6 and Figure 5.9) are probably not the result of sequencing errors.

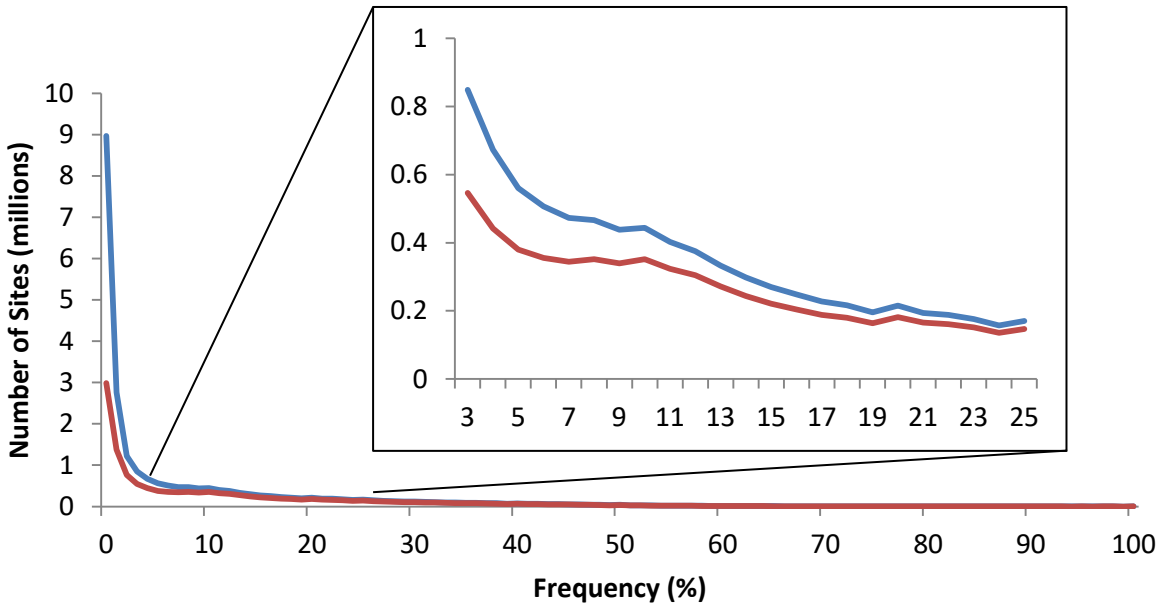


Figure 5.12: Alternate allele frequencies before and after filtering of sites. The blue line represents the raw frequency allele numbers, while the red line represents the numbers after all sites where an average sequencing quality below 30 in the alternate allele reads were removed.

Confirming the Biological Origin of the Degraded Signal

If the unexpected patterns in the k-mer frequency plot and alternate allele frequency plots aren't an artefact of some kind, then it is reasonable that it should leave a biologically relevant signal in the data. One biologically relevant marker that can be checked is whether the 3rd codon position shows a higher percentage of sites with an alternate allele when compared to the first 2 sites in the codon. One would predict that somatic mutations that change a protein structure are more likely to be deleterious for a cell than synonymous mutations which don't change the protein structure. This means that somatic mutations are more likely to accumulate in the more permissive 3rd codon position than the first 2 codon positions.

Over 250 000 presumptive 3rd codon positions were identified. These sites had an alternate allele 6.6% of the time (after low quality filtering). In comparison, the first 2 positions had an alternate allele 4.0% of the time (Table 5.6). Interestingly, the frequency in the 3rd codon position is even higher than the base genomic frequency which was 5.9%.

The codon relationship was reproduced in the RNA-seq datasets (Table 5.6) as expected. Unlike the genomic data, the overall alternate allele frequency was higher than the 3rd codon position in the RNA-seq datasets. This is not inconsistent per se. The genomic dataset features transcribed and untranscribed regions, whereas the RNA-seq datasets, by definition, are only transcribed regions. So the two statistics are not measuring comparable things in this case. But the difference is still noteworthy. The alternate allele frequency in the coding portion of transcripts should be an average of the three positions. So that will be closer to the average of the first two positions than the third position. This means that for the overall transcript average to be higher than codon average, it suggests that the alternate allele frequency in the non-coding regions is substantially higher than both the coding regions and the assembled regions of the genome that weren't transcribed.

Table 5.6: Alternate allele frequency of the first two codon positions compared to the third codon position in the 4 datasets. Note that due to the nature of the data, the "overall frequency" category represents different types of sequences in the "Leaf DNA" data versus the RNA-seq data.

	Leaf DNA (field)	Seed RNA (field)	Leaf RNA (field)	Leaf and root RNA (lab)
First 2 codon positions	4.0	3.0	3.5	3.5
3rd codon position	6.6	4.5	5.0	4.7
Overall frequency	5.9	5.0	5.3	5.2

A final note about the alternate allele frequencies (Table 5.6) is that these values should be considered relative to the other values in their dataset only. This is because the measured frequency is highly dependent on the coverage of the sample. Higher coverage samples will detect less frequent mutations in the samples and will also have more sequencing errors that pass the quality score threshold that was set. This means that higher coverage samples will have a higher alternate allele frequency. For example, it seems counter intuitive that the lab grown plants have a similar alternate allele frequency to plants harvested from the field and undergone multiple cycles of desiccation. But the RNA-seq from lab grown plants that had never desiccated had twice as many reads as the RNA-seq from plants from the field (Table 5.2).

Discussion

The whole genome sequencing of *X. humilis* had modest goals at its outset. It simply aimed to create a genome assembly that was of a high enough quality to link genes to their associated promoters. But this project was immediately hindered by unexpected and unprecedented problems. These problems appear to derive from two identifiable sources; a recent genome duplication and an apparent accumulation of somatic mutations. It is difficult to disentangle which of these factors is playing the major disruptive role to the genome assembly process, or if there are additional disruptive factors which have been obscured.

There was a Recent, Previously Unknown, Genome Duplication in *Xerophyta humilis*

The evidence for a recent genome duplication is persuasive. The coverage of genomic reads over the genome (Figure 5.5), the alternate allele frequency of RNA-seq from plants grown in the lab (Figure 5.7 and Figure 5.9), and the presence and frequency of sites with 3 or more alleles (Figure 5.10) are all consistent with the hypothesis that there was a genome duplication in *X. humilis* which was merged during genome assembly into a hybrid ancestral sequence. The fact that the genome assembly merged the duplicated sequences also tells us that the duplication probably happened recently, with sequences not having had enough time to diverge.

The genome size predicted by Hanson et al (2001) was 532 Mbp. This was consistent with the assembly size generated by SOAPdenovo2 (Table 5.4) and the size predicted by the 4n coverage statistics (Figure 5.5). This means that the genome duplication identified in the sequencing data was not present in Hanson et al's (2001) sample. This probably represents a sub-species of *X. humilis* different to that used by Hanson et al, which was collected in Botswana.

Repeated Dehydration Likely Causes an Accumulation of Somatic Mutations

The most irregular aspect of the data is the lack of expected peaks in the k-mer frequency and alternate allele frequency distributions (Figures 5.2, 5.8, 5.9 and 5.10). This does not appear to be the result of sequencing error, or some other artefact (Figure 5.2 and Table 5.6). The k-mer frequency plot (Figure 5.3) suggests a low coverage across the vast majority of 25-mers. However,

the coverage of aligned reads across the assembly (Figure 5.5) suggests that the lack of coverage in the 25-mers does not translate into a lack of genomic coverage.

There are several contributing factors that could explain the differences between the two figures. The k-mer frequency plot doesn't allow for any mismatches in the sequences, unlike the Bowtie2 alignment algorithm. To compound this effect, any change in a sequence will result in 'k' new k-mers, as the somatic mutation changes position within the k-mer. These k-mers, along with the original k-mers will all be at a lower frequency than they were without the somatic mutation. This results in the real peaks being eroded by 'k' k-mers for each somatic mutation, while the left hand shoulder builds up at twice the rate.

In comparison, the genomic coverage has excluded a lot of the more troublesome data. There is a survival bias in which regions were assembled, both due to the assembly being done on a reduced dataset and a bias as to what is actually able to be assembled, and which reads were aligned. This filters out a lot of the noise and maintains the main coverage peaks.

A possible explanation for the k-mer and alternate allele frequency plots is that there is an accumulation of somatic mutations in the sequenced plants. But somatic mutations have not been reported in such high numbers before (Dubrovina and Kiselev, 2016). Even in the 100 year old Oak tree, *Quercus robur*, which was considered to have a high level of somatic mutations as a result of its age, the genome sequencing (and k-mer frequency plots) were achieved without issue (Plomion et al, 2018). Given the uniqueness of mutations to this extent, and the unique trait of the plant, *X. humilis*, these mutations are presumably a result of the desiccation process. Whether the process of dehydrating and rehydrating itself is damaging (Jiang et al, 2014), or else being in a dehydrated state is damaging, possibly in combination with something like UV damage that can't effectively be repaired (Shibai et al, 2017), still has to be determined. With the exception of the lab grown plants used for RNA sequencing, all plants used for this study were harvested from the wild. This means it is difficult to say how many hours or cycles of desiccation were necessary to cause the extent of the mutations observed in the data. But the fact that the only data that did not show degraded peaks in the alternate allele frequency plot's was from lab grown plants which had never undergone dehydration is highly suggestive of this hypothesis.

Another question that needs to be answered is to what degree, if any, the genome is protected during this process from somatic mutation. While the rate of mutation appears exceptionally high, it is clear that the distribution of mutations across the genome is not even. Is this the result of selective insulation of portions of the genome or is it some sort of survival bias that occurs at the cell

level, where cells that get mutations in vital genes are unable to function and die. Notably, meristem cells that will go on to produce seeds do not appear to be subject to more protection at the genome level than any other cell (see seed pod data in Figure 5.7 and 5.10). This could suggest the plant is unable to protect even a small subset of cells from the overall process. It could also mean that the observed mutations do in fact make it into the germ line, which could lead to an accelerated rate of evolution in *X. humilis* or an accumulation of detrimental mutations faster than natural selection is able to remove them.

Other Desiccation Tolerant Plant Genomes

If the build-up of somatic mutations is the result of repeated dehydration and rehydration, it is reasonable to wonder why this has not been an issue for previous work. Multiple desiccation tolerant plants have been sequenced and assembled independently (VanBuren et al, 2015; Xiao et al, 2015), including *Xerophyta viscosa* (Costa et al, 2017), a close relative of *X. humilis*. But no pattern of somatic mutation was reported in any of the studies. This can be explained by the fact that all three studies used plants grown in the lab, whereas *X. humilis* was sequenced from a sample obtained from the field. This suggests that the *X. humilis* plant was probably older and had experienced harsher conditions than the plants used in the other studies.

Conclusion

The *X. humilis* genome proved incredibly challenging. While better than no assembly, the current genome assembly is highly fragmented and has a lot of room for improvement (with a scaffold N50 size of 7.5kbp and 17% N's). The genome duplication is interesting, but the real discovery of this work is the somatic mutation rate that was identified. This could have interesting ramifications for future work. There are some basic questions that still need to be answered: how widespread is the trend within the desiccation tolerant plants? And what exactly is the mechanism that is causing the mutation? Is there a system to protect vital regions of the genome? But these questions aside, the value of the system could be immense. This plant potentially represents the greatest mutation assay possible, with blue prints to every important genetic sequence in the genome just waiting to be identified. This could make for a valuable and unique resource in the field of plant genetics.

Chapter 6 – Concluding Remarks

This project aimed to use the recent advances in DNA sequencing technology to explore the genomes of several non-model systems. A diverse range of non-model species were used to identify and answer questions of evolutionary interest in four systems: Natal Long Fingered Bat, Yellow Baboon, Ruschioideae and *Xerophyta humilis*. In doing so, a wide assortment of methodologies were used and developed, taking full advantage of the versatility that whole genome sequencing can provide. The chosen species and questions are wide ranging, but the skills and ways of thinking proved remarkably transferable between the different systems. This resulted in an expansive exploration into what Next Generation Sequencing has to offer for non-model systems.

The Natal Long Fingered Bat, *Miniopterus natalensis*, was used as a representative system to investigate the genetic mechanisms responsible for the the development of the bat wing. We reasoned that a RNA-seq and ChIP-seq analysis comparing embryonic bat forelimb autopods to bat hindlimb autopods would explain how the highly elongated digits and the retained interdigital webbing which make up the bat wing evolved. In order to do this, an assembled genome was required to facilitate RNA-seq and ChIP-seq analysis. In addition to the genome assembly and annotation, dN/dS analysis and lncRNA prediction were conducted. The genome assembly and annotation were successful, with assembly metrics being comparable to other assembled bat genomes. The dN/dS analysis showed that all the differentially expressed signalling pathways were being selectively conserved, while the lncRNA analysis identified several putative transcripts that were differentially expressed, including Tbx5as1 and HoxA13as1 (Eckalbar et al, 2016).

The Amboseli National Park in Kenya has a local population of Yellow baboons (*Papio cynocephalus*) that has recently come into contact with a population of Olive baboons (*Papio anubis*). These populations appear to be hybridising. A genome assembly of *P. cynocephalus* was assembled and used to align low coverage sequencing from 45 baboons, including admixed individuals along with unadmixed individuals each species. By identifying SNPs that were predictive of species, hybrid individuals were confirmed in the Amboseli population. Furthermore, the observed frequency for *P. anubis* SNPs in the Amboseli population was higher than was expected, suggesting that this is not the first time admixture has occurred between the two populations (Wall et al, 2016).

The Ruschieae are a prolific Tribe of plants found along the West Coast of Southern Africa. Comprised of approximately 1500 species that have recently diverged, the Tribe has the highest rate of speciation known for land plants (Klak et al, 2004). An exploratory analysis sequenced and assembled two Ruschieae genomes (*Polymita steenbokensis* and *Faucaria felina*) along with a sister

taxon (*Cleretum herrei*) from a neighbouring tribe. The three plants, collectively in the sub-family Ruschioideae, were compared to each other in order to try and identify any genetic signatures which could explain the rapid speciation. The two Ruschieae species had increased levels of duplication within the genome without having undergone a genome duplication, as had previously been suggested (Illing et al, 2009). These duplications did not appear to be tandem in nature. Unfortunately, the Ruschieae genome assemblies were of poor quality, which limited the possible analyses and conclusions that could be drawn.

Desiccation tolerant plants are able to dry down and rehydrate without any apparent morphological harm. *X. humilis* is one such 'resurrection' plant which has been used to study the mechanism by which this is able to occur in Angiosperms. In order to further facilitate a RNA-seq analysis of hydrated and desiccating *X. humilis* leaves, the genome was sequenced and assembled. During the assembly process, it became clear that something extraordinary was occurring at the genetic level. Further analysis suggested that the process of dehydration and rehydration was resulting in rampant somatic mutations. The somatic mutations, combined with a previously unknown genome duplication made the genome assembly challenging and limited in applicability.

These apparently disparate topics explored the possibilities and limitations for whole genome sequencing in the study of non-model organisms. Mechanisms of genetic change were examined at the genomic scale, from adaptation and hybridisation to various forms of duplication and mutation. In this way, a large variety of events responsible for the evolutionary change of genomes in plants and animals were analysed in a diverse set of systems. While each genome presented its own unique set of problems, methods and insights would often transcend the genome they were developed for.

Of the systems explored, *M. natalensis* proved the most conventional. The normality of a mammalian genome, along with high coverage and multiple library insert length made the assembly relatively straight forward. But the genome assemblies and analyses grew increasingly more challenging. The lower coverage, larger genome and poorer sequencing quality made the baboon assembly worse quality overall than *M. natalensis*, although still adequate for the purposes of the snp analysis of individuals in the hybridizing zone. The two plant systems were a different story however, proving far more difficult than anticipated.

Plant Genomes

The transition from the mammalian genomes to the plant genomes represented a massive increase in assembly difficulty. While there is literature documenting that plants are generally a more difficult and variable system to work on compared to mammals, the magnitude of that difference was still a surprise.

A major difference in the assembly process was in the way in which the genomes improved with optimisation. As a general rule, the mammalian genome assemblies started off relatively mediocre, but then saw a steady improvement as variables were refined and data added. This behaviour was not observed in the plant systems, which often saw little improvement gained through optimisation.

While using Platanus, an assembler that was designed for plant genomes, helped with the Ruschieae, they were still not very successful. In comparison, *Cleretum herrei* assembled relatively well. Especially considering the assembly only had 2 short insert libraries. This drastic difference (*C. herrei* had an N50 value of 50 kbp, compared to the Ruschieae N50 of 5-6 kbp) between two closely related plants shows an inherent risk associated with genome sequencing. Namely, there is no cost efficient means of discovering the relevant qualities of a genome that determines how well it will assemble before sequencing it. This can also be seen in the difficulties that were encountered in *X. humilis.* The risk appears amplified in plant genomes, which, as evidenced by the 4 sequenced here, are highly variable and able to change over short evolutionary time frames.

K-mer Frequency Plots

While there isn't a cost effective means of determining how well a genome will assemble, it is worth noting that all of the major problems that were encountered in the plant genome analyses were observable from the k-mer frequency graphs of their data. The high heterozygosity of the Ruschieae was clearly visible and was in stark contrast to the *C. herrei* k-mer frequency plot. The duplicated sequences in the Ruschieae were less obvious, but still visible in the disproportionate levels of high frequency k-mers. And while unorthodox, the somatic mutations that were eventually verified in *X. humilis* were first hypothesised from the unusual k-mer frequency plot the data produced. The utility of k-mer frequency plots should not be underestimated. They can be a lot more useful than just helping with sequencing error corrections. K-mer frequency plots can predict how successful of an assembly can be generated and even inform entire aspects of the downstream analysis strategy.

The Future of Genome Sequencing

It is difficult to ascertain to what degree the difficulties in assembly of the plant genomes could have been avoided if long read technologies had been used instead. But, in the time since the sequencing was originally done, these technologies have shown promise in *de novo* sequencing of plant genomes (Paajanen et al, 2017). Long read technologies, such as Pacific Bioscience, would make the plant genome assemblies a lot less fragmented, sequencing through repetitive elements and joining duplicated regions with divergent regions. This would allow for disentanglement in either the local duplication case, as seen in the Ruschieae, or in the case of whole genome duplication, as seen in *Xerophyta humilis*. Hopefully this technology will enable *de novo* plant genome assembly in the same way that Next Generation Sequencing has enabled mammalian genome assembly, allowing high quality genomic resources for niche, non-model systems.

References

Alberts, S. C. and Altmann, J. 2012. The Amboseli Baboon Research Project: 40 Years of Continuity and Change. In: Kappeler, P. M. and Watts, D. P. (editors). 2012. *Long-Term Field Studies of Primates*. Berlin: Spring-Verlag

Alberts, S. C., Altmann, J. 2001. Immigration and hybridization patterns of yellow and anubis baboons in and around Amboseli, Kenya. *American Journal of Primatology* **53:** 139-154

Alpert, P. 2005. The limits and frontiers of desiccation-tolerant life. *Integrative and Comparative Biology* **45: 5:** 685-695

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. and Lipman, D. J. 1990. Basic local alignment search tool. *Journal of Molecular Biology* **215:** 403-410

Andrews, S. 2010. FastQC: a quality control tool for high throughput sequence data. *http://www.bioinformatics.babraham.ac.uk/projects/fastqc*

Arakaki, M., Christin, P.-A., Nyffeler, R., Lendel, A., Eggli, U., Ogburn, R. M., Spriggs, E., Moore, M. J. and Edwards, E. J. 2011. Contemporaneous and recent radiations of the world's major succulent plant lineages. *Proceedings of the National Academy of Sciences* **108: 20:** 8379-8384

Arnold, M. L. and Meyer, A. 2006. Natural hybridization in primates: one evolutionary mechanism. *Zoology* **109:** 261-276

Bentley, D. R., Balasubramanian, S., Swerdlow, H. P., et al. 2008. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456: 7218:** 53–59

Bolger, A. M., Lohse, M. and Usadel, B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30: 15:** 2114-2120

Born, J., Linder, H. P. and Desmet, P. 2007. The Greater Cape Floristic Region. *Journal of Biogeography* **34:** 147-162

Bourque, G., Burns, K. H., Gehring, M., et al. 2018. Ten things you should know about transposable elements. *Genome Biology* **19:** 199

Cantrell, M. A., Scott, L., Brown, C. J., Martinez, A. R. and Wichman, H. A. 2008. Loss of LINE-1 activity in the megabats. *Genetics* **178:** 393-404

Carroll, S. B. 2008. Evo-Devo and an Expanding Evolutionary Synthesis: A Genetic Theory of Morphological Evolution. *Cell* **134:** 25-36

Charpentier, M. J., Fontaine, M. C., Cherel, E., et al. 2012. Genetic structure in a dynamic baboon hybrid zone corroborates behavioural observations in a hybrid population. *Molecular Ecology* **21:** 715-731

Chin, C.-S., Peluso, P., Sedlazeck, F. J. et al. 2016. Phased Diploid Genome Assembly with Single Molecule Real-Time Sequencing. *Nature Methods* **13: 12:** 1050-1054

Clarke, J., Wu, H. C., Jayasinghe, L., Patel, A., Reid, S. and Bayley, H. 2009. Continuous base identification for single-molecule nanopore DNA sequencing. *Nature Nanotechnology* **4:** 265-270

Cooper, L. N., Cretekos, C. J. and Sears, K. E. 2012. The evolution and development of mammalian flight. *Wiley Interdisciplinary Reviews: Developmental Biology* **1:** 773–779

Cowling, R. M., Procheş, Ş. and Partridge, T. C. 2009. Explaining the uniqueness of the Cape flora: Incorporating geomorphic evolution as a factor for explaining its diversification. *Molecular Phylogenetics and Evolution* **51: 1:** 64-74

Cusack, B. P. and Wolfe, K. H. 2007. Not born equal: increased rate asymmetry in relocated and retrotransposed rodent gene duplicates. *Molecular Biology and Evolution* **24:** 679-686.

De Rocher, E. J., Harkins, K. R., Galbraith, D. W. and Bohnert, H. J. 1990. Developmentally regulated systematic endopolyploidy in succulents with small genomes. *Science* **250: 4977:** 99-101

Diekmann, B., Fälker, M. and Kuhn, G. 2003. Environmental history of the south-eastern South Atlantic since the Middle Miocene: evidence from the sedimentological records of ODP Sites 1088 and 1092. *Sedimentology* **50:** 511-529

Dolezel, J., Greilhuber, J. and Suda, J. 2007. Estimation of nuclear DNA content in plants using flow cytometry. *Nature Protocols* **2: 9:** 2233-2244

Dong, D., Lei, M., Hua, P., Pan, Y. H., Mu, S., Zheng, G., Pang, E., Lin, K., Zhang, S. 2017. The Genomes of Two Bat Species with Long Constant Frequency Echolocation Calls. *Molecular Biology and Evolution* **34: 1:** 20-34

Dubrovina, A. S. and Kiselev, K. V. 2016. Age-associated alterations in the somatic mutation and DNA methylation levels in plants. *Plant Biology* **18: 2:** 185-196

Duchene, D. and Bromham, L. 2013. Rates of molecular evolution and diversification in plants: chloroplast substitution rates correlate with species-richness in the Proteaceae. *BMC Evolutionary Biology* **13:** 65

Earl, D. A., Bradnam, K., John, J. S., et al. 2011. Assemblathon 1: A competitive assessment of de novo short read assembly methods. *Genome Research* **21:** 2224-2241

Eckalbar, W. L., Schlebusch, S. A., Mason, M. K., et al. 2016. Transcriptomic and epigenomic characterization of the developing bat wing. *Nature Genetics* **48:** 528-536

Eid, J., Fehr, A., Gray, J., et al. 2009. Real-time DNA sequencing from single polymerase molecules. *Science* **323:** 133-138

Fedurco, M., Romieu, A., Williams, S., Lawrence, I. and Turcatti, G. 2006. BTA, a novel reagent for DNA attachment on glass and efficient generation of solid-phase amplified DNA colonies. Nucleic Acids Research **9: 34: 3:** e22

Fisher, K. M. 2008. Bayesian reconstruction of ancestral expression of the LEA gene families reveals propagule-derived desiccation tolerance in resurrection plants. *American Journal of Botany* **95: 4:** 506-515

Foden, W. and Potter, L. 2005. *Xerophyta humilis* (Baker) T.Durand & Schinz. *National Assessment: Red List of South African Plants* **2017**: 1

Freeling, M. 2009. Bias in plant gene content following different sorts of duplication: tandem, whole-genome, segmental, or by transposition. Annual Review of Plant Biology **60:** 433-453

Gnerre, S., MacCallum, I., Przybylski, D., et al. 2011. High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proceedings of the National Academy of Sciences USA* **108: 4:** 1513-1518

Goldman, N. and Yang, Z. 1994. A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Molecular Biology and Evolution* **11: 5:** 725-36

Gregory, T. R. 2019. Animal Genome Size Database. Available at www.genomesize.com. Accessed May 18, 2019

Hanson, L., Mcmahon, K. A., Johnson, M. A. T. and Bennett, M. D. 2001. First Nuclear DNA C-values for 25 Angiosperm Families. *Annals of Botany* **87:** 251-258

Harrow, J., Frankish, A., Gonzalez, J. M., et al. 2012. GENCODE: the reference human genome annotation for the ENCODE Project. *Genome Research* **22:** 1760-1774

Hockman, D., Cretekos, C. J., Mason, M. K., Behringer, R. R., Jacobs, D. S. and Illing, N. 2008. A second wave of Sonic hedgehog expression during the development of the bat limb. *Proceedings of the National Academy of Sciences of the United States of America* **105:** 16982–16987

Hockman, D., Mason, M. K., Jacobs, D. S. and Illing, N. 2009. The role of early development in mammalian limb diversification: a descriptive comparison of early limb development between the Natal long-fingered bat (*Miniopterus natalensis*) and the mouse (*Mus musculus*). *Developmental Dynamics* **238: 4:** 965-979

Holland, P. W. H., Marlétaz, F., Maeso, I., Dunwell, T. L. and Paps, J. 2017. New genes from old: asymmetric divergence of gene duplicates and the evolution of development. *Philosophical Transactions of the Royal Society B: Biological Sciences* **372: 1713:** 20150480

Holt, C. and Yandell, M. 2011. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics* **12:** 491

Ihlenfeldt, H-D. 1994. Diversification in an Arid World: The Mesembryanthemaceae. *Annual Review of Ecology and Systematics* **25:** 521-546

Illing, N., Denby, K., Collett, H., Shen, A. and Farrant, J. M. 2005. The signature of seeds in resurrection plants, a molecular and physiological comparison of desiccation tolerance in seeds and vegetative tissues. *Integrative and Comparative Biology* **45:** 771-787

Illing, N., Klak, C., Johnson, C., Brito, D., Negrao, N., Baine, F., van Kets, V., Ramchurn, K. R., Seoighe, C. and Roden, L. 2009 Duplication of the Asymmetic Leaves1/Rough Sheath 2/Phantastica (ARP) gene precedes the explosive radiation of the Ruschioideae. *Development, Genes, and Evolution* **219:** 331-338

Jiang, C., Mithani, A., Belfield, E. J., Mott, R., Hurst, L. D. and Harberd, N. P. 2014. Environmentally responsive genome-wide accumulation of de novo *Arabidopsis thaliana* mutations and epimutations. *Genome Research* **24:** 1821-1829

Jiang, N., Bao, Z., Zhang, X., Eddy, S. R., Wessler, S. R. 2004. Pack-MULE transposable elements mediate gene evolution in plants. *Nature* **431:** 569-573

Johnson, A. D., Handsaker, R. E., Pulit, S. L., Nizzari, M. M., O'Donnell, C. J. and de Bakker, P. I. 2008. SNAP: a web-based tool for identification and annotation of proxy SNPs using HapMap. *Bioinformatics* **24:** 2938–2939

Johnsson, P., Lipovich, L., Grandér, D. and Morris, K. V. 2014. Evolutionary conservation of long noncoding RNAs; sequence, structure, function. *Biochimica et Biophysica Acta* **1840: 3:** 1063-1071

Jurka, J., Kapitonov, V. V., Pavlicek, A., Klonowski, P., Kohany, O. and Walichiewicz, J. 2005. Repbase Update, a database of eukaryotic repetitive elements. *Cytogenetic and Genome Research* **110: 1-4:** 462-467

Kajitani, R., Toshimoto, K., Noguchi, H., et al. 2014. Efficient de novo assembly of highly heterozygous genomes from whole-genome shotgun short reads. *Genome Research* **24: 8:** 1384-1395

Kapusta, A., Suh, A. and Feschotte, C. 2017. Dynamics of genome size evolution in birds and mammals. *Proceedings of the National Academy of Sciences of the United States of America* **114: 8:** E1460-E1469

Kellogg, E. A. 2016. Has the connection between polyploidy and diversification actually been tested? *Current Opinion in Plant Biology* **30:** 25-32

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R. and Salzberg, S. L. 2013. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biology* **14:** R36

Kimura, M. 1977. Preponderance of synonymous changes as evidence for the neutral theory of molecular evolution. *Nature* **267:** 275-276

Kingdon, J., Butynski, T.M. & De Jong, Y. 2008. *Papio anubis*. The IUCN Red List of Threatened Species 2008: e.T40647A10348950

Kingdon, J., Butynski, T.M. & De Jong, Y. 2016. *Papio cynocephalus*. The IUCN Red List of Threatened Species 2016: e.T92250442A92250811

Klak, C., Bruyns, P. V. and Hanáček, P. 2013. A phylogenetic hypothesis for the recently diversified Ruschieae (Aizoaceae) in southern Africa. *Molecular Phylogenetics and Evolution* **69: 3:** 1005-1020

Klak, C., Hanáček, P. and Bruyns, P. V. 2017. Out of southern Africa: Origin, biogeography and age of the Aizooideae (Aizoaceae). *Molecular Phylogenetics and Evolution* **109:** 203-216

Klak, C., Reeves, G. and Hedderson, T. 2004. Unmatched tempo of evolution in Southern African semi-desert ice plants. *Nature* **427: 6969:** 63-65

Kong, L., Zhang, Y., Ye, Z. Q., Liu, X. Q., Zhao, S. Q., Wei, L. and Gao, G. 2007. CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine. *Nucleic Acids Research* **35:** W345–W349

Krammer, R., Baumann, K.-H. and Henrich, R. 2006. Middle to late Miocene fluctuations in the incipient Benguela Upwelling System revealed by calcareous nannofossil assemblages (ODP Site 1085A). *Palaeogeography, Palaeoclimatology, Palaeoecology* **230:** 319-334

Kung, J. T. Y., Colognori, D. and Lee, J. T. 2013. Long Noncoding RNAs: Past, Present, and Future. *Genetics* **193: 3:** 651-669

Kyriakidou, M., Tai, H. H., Anglin, N. L., Ellis, D. and Strömvik, M. V. 2018. Current strategies of polyploid plant genome sequence assembly. *Frontiers in Plant Science* **9:** 1660

Langmead, B and Salzberg, S. L. 2012. Fast gapped-read alignment with Bowtie 2. *Nature Methods* **9:** 357-359

Li, C., Lin, F., An, D., Wang, W. and Huang, R. 2017. Genome Sequencing and Assembly by Long Reads in Plants. *Genes (Basel)* **9: 1:** 6

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R. and 1000 Genome Project Data Processing Subgroup. 2009. The Sequence alignment/map (SAM) format and SAMtools. *Bioinformatics* **25:** 2078-2079

Li, R., Zhu, H., Ruan, J., et al. 2010. De novo assembly of human genomes with massively parallel short read sequencing. *Genome Research* **20:** 265-272

Lisch, D. 2012. How important are transposons for plant evolution? *Nature Reviews Genetics* **14:** 49-61

Luo, R., Liu, B., Xie, Y. et al. 2012. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *Gigascience* **1:** 18

Lutz, K. A., Wang, W., Zdepski, A. and Michael, T. P. 2011. Isolation and analysis of high quality nuclear DNA with reduced organellar DNA for plant genome sequencing and resequencing. *BMC Biotechnology* **11:** 54

Magoč, T. and Salzberg, S. L. 2011. FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics* **27: 21:** 2957–2963

Mardis, E. R. 2011. A decade's perspective on DNA sequencing technology. Nature **470:** 198–203

Margulies, M., Egholm, M., Altman W. E., et al. 2005. Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437:** 376-380

Mason, M. K., Hockman, D., Curry, L., Cunningham, T. J., Duester, G., Logan, M., Jacobs, D. S. and Illing, N. 2015. Retinoic acid–independent expression of Meis2 during autopod patterning in the developing bat and mouse limb. *Evodevo* **6:** 6

McDonald, J. T., Rautenbach, L. T. and Nel, J. A. J. 1990. Roosting requirements and behaviour of five bat species at De Hoop Guano Cave, southern Cape Province of South Africa. *South African Journal of Wildlife Research* **20: 4:** 157-161

McKenna, A., Hanna, M., Banks, E., et al. 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research* **20:** 1297-1303

Mercer, T. R., Dinger, M. E. and Mattick, J. S. 2009. Long non-coding RNAs: insights into functions. *Nature Reviews Genetics* **10:** 155-159

Meredith, R. W., Janecka, J. E., Gatesy, J., et al. 2011. Impacts of the Cretaceous Terrestrial Revolution and KPg Extinction on Mammal Diversification. *Science* **34: 6055:** 521-524

Miller, J. R., Koren, S. and Sutton, G. 2010. Assembly algorithms for next-generation sequencing data. *Genomics* **95:** 315–327

Miller-Butterworth, C. M., Murphy, W. J., O'Brien, S. J., Jacobs, D. S., Springer, M. S. and Teeling, E. C. 2007. A family matter: conclusive resolution of the taxonomic position of the long-fingered bats, miniopterus. *Molecular Biology and Evolution* **24: 7:** 1553-1561

Monadjem, A., Griffin, M., Cotterill, F., Jacobs, D. and Taylor, P. J. 2017. Miniopterus natalensis. *The IUCN Red List of Threatened Species* **2017:** e.T44862A22073129

Muse, S. V. and Gaut, B. S. 1994. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. *Molecular Biology and Evolution* **11: 5:** 715-724.

Paajanen, P., Kettleborough, G., Lopez-Girona, E., et al. 2017. A critical comparison of technologies for a plant genome sequencing project. *bioRxiv* 201830

Panchy, N., Lehti-Shiu, M. and Shiu, S-H. 2016. Evolution of Gene Duplication in Plants. *Plant Physiology* **171:** 2294-2316

Parker, J., Tsagkogeorga, G., Cotton, J. A., Liu, Y., Provero, P., Stupka, E. and and Rossiter, S. J. 2013. Genome-wide signatures of convergent evolution in echolocating mammals. *Nature* **502:** 228–231

Parra, G., Bradnam, K. and Korf, I. 2007. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics* **23: 9:** 1061-1067

Paterson, A. H., Freeling, M., Tang, H. and Wang, X. 2010. Insights from the Comparison of Plant Genome Sequences. *Annual Review of Plant Biology* **61:** 349-372

Pevzner, P. A., Tang, H., and Waterman, M. S. 2001. An Eulerian path approach to DNA fragment assembly. *Proceedings of the National Academy of Sciences* **98: 17:** 9748-9753

Platt, R. N., Mangum, S. F. and Ray, D. A. 2016. Pinpointing the vesper bat transposon revolution using the *Miniopterus natalensis* genome. *Mobile DNA* **7:** 12

Plomion, C., Aury, J.-M., Amselem, J., et al. 2018. Oak genome reveals facets of long lifespan. *Nature Plants* **4:** 440-452

Quek, X.C., Thomson, D. W., Maag, J. L., Bartonicek, N., Signal, B., Clark, M. B., Gloss, B. S. and Dinger, M. E. 2015. lncRNAdb v2.0: expanding the reference database for functional long noncoding RNAs. *Nucleic Acids Research* **43:** D168–D173

Rabosky, D. L., Santini, F., Eastman, J., Smith, S. A., Sidlauskas, B., Chang, J. and Alfaro, M. E. 2013. Rates of speciation and morphological evolution are correlated across the largest vertebrate radiation. *Nature Communications* **4:** 1958

Ranwez, V., Harispe, S., Delsuc, F. and Douzery, E. J. P. 2011. MACSE: Multiple Alignment of Coding SEquences Accounting for Frameshifts and Stop Codons. *Plos One* **6: 9:** e22594

Richardson, J. E., Weitz, F. M., Fay, M. F., Cronk, Q. C. B, Linder, H. P., Reeves, G. and Chase, M. W. 2001. Rapid and recent origin of species richness in the Cape Flora of South Africa. *Nature* **412:** 181-183

Rieseberg, L. H. and Willis, J. H. 2007. Plant speciation. *Science* **317: 5840:** 910-914

Rinn, J. L., Kertesz, M., Wang, J. K., et al. 2007. Functional Demarcation of Active and Silent Chromatin Domains in Human HOX Loci by Non-Coding RNAs. *Cell* **129: 7:** 1311-1323

Rundle, H. D. and Nosil, P. 2005. Ecological Speciation. *Ecology Letters* **8: 3:** 336-352

Sanger, F., Nicklen, S. and Coulson, A. R. 1977. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences* **74: 12:** 5463–5467

Sankararaman, S., Patterson, N., Li, H., Pääbo, S. and Reich, D. 2012. The date of interbreeding between Neandertals and modern humans. Plos Genetics **8: 10:** e1002947

Schlebusch, S. and Illing, N. 2012. Next generation shotgun sequencing and the challenges of de novo genome assembly. South African Journal of Science **108: 11-12:** 62-70

Schmutz, J., Wheeler, J., Grimwood, J., et al. 2004. Quality assessment of the human genome sequence. *Nature* **429:** 365-368

Seim, I., Fang, X., Xiong, Z., et al. 2013. Genome analysis reveals insights into physiology and longevity of the Brandt's bat *Myotis brandtii*. *Nature Communications* **4:** 2212

Shibai, A., Takahashi, Y., Ishizawa, Y., Motooka, D., Nakamura, S., Ying, B. and Tsuru, S. 2017. Mutation accumulation under UV radiation in *Escherichia coli*. *Scientific Reports* **7:** 14531

Simpson, J. T. and Durbin, R. 2012. Efficient de novo assembly of large genomes using compressed data structures. *Genome Research* **22: 3:** 549-556

Simpson, J. T., Wong, K., Jackman, S. D., Schein, J. E., Jones, S. J. M. and Birol, İ. 2009. ABySS: A parallel assembler for short read sequence data. *Genome Research* **19: 6:** 1117–1123

Slater, G. S. and Birney, E. 2005. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* **6:** 31

Smit, A. F. A., Hubley, R. and Green, P. RepeatMasker Open-4.0. 2013-2015. *http://www.repeatmasker.org*

Snyder-Mackler, N., Majoros, W. H., Yuan, M. L., Shaver, A. O., Gordon, J. B., Kopp, G. H., Schlebusch, S. A., Wall, J. D., Alberts, S. C., Mukherjee, S., Zhou, X., Tung, J. 2016. Efficient Genome-Wide Sequencing and Low-Coverage Pedigree Analysis from Noninvasively Collected Samples. *Genetics* **203: 2:** 699-714

Springer, N. M., Anderson, S. N., Andorf, C. M., et al. 2018. The maize W22 genome provides a foundation for functional genomics and transposon biology. *Nature Genetics* **50:** 1282-1288

Stanke, M., Steinkamp, R., Waack, S. and Morgenstern, B. 2004. AUGUSTUS: a web server for gene finding in eukaryotes. *Nucleic Acids Research* **32:** W309–W312

Strauss, W. M. 1993. Preparation of Genomic DNA from Mammalian Tissue. *Current Protocols in Immunology* **10: 2**

Swartz, S. M. and Middleton, K. M. 2008. Biomechanics of the Bat Limb Skeleton: Scaling, Material Properties and Mechanics. *Cells Tissues Organs* **187:** 59-84

Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M. and Kumar, S. 2011. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Molecular Biology and Evolution* **10:** 2731-3739

The ENCODE Project Consortium. 2011. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489:** 57-74

Thompson, J. D., Higgins, D. G. and Gibson, T. J. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research* **22: 22:** 4673-4680

Tung, J., Charpentier, M. J., Mukherjee, S., Altmann, J. and Alberts, S. C. 2012. Genetic effects on mating success and partner choice in a social mammal. *American Naturalist* **180: 1:** 113-129

Valente, L. M., Britton, A. W., Powell, M. P., Papadopulos, A. S. T., Burgoyne, P. M. and Savolainen, V. 2014. Correlates of hyperdiversity in southern African ice plants (Aizoaceae). *Botanical Journal of the Linnean Society* **174: 1:** 110-129

Van de Peer, Y., Mizrachi, E. and Marchal, K. 2017. The evolutionary significance of polyploidy. *Nature Reviews Genetics* **18:** 411-424

Vukašinović, N., Cvrčková, F., Eliáš, M., Cole, R., Fowler, J. E., Žárský, V. and Synek, L. 2014. Dissecting a hidden gene duplication: the Arabidopsis thaliana SEC10 locus. *Plos One* **9: 4:** e94077

Wall, J. D., Lohmueller, K. E. and Plagnol, V. 2009. Detecting ancient admixture and estimating demographic parameters in multiple human populations. *Molecular biology and evolution* **26:** 1823-1827

Wall, J. D., Schlebusch, S. A., Alberts, S. C., et al. 2016. Genome-wide ancestry and divergence patterns from low-coverage sequencing data reveal a complex history of admixture in wild baboons. *Molecular ecology* **25: 14:** 3469-3483

Wang, K. C., Yang, Y. W., Liu, B., et al. 2011. A long noncoding RNA maintains active chromatin to coordinate homeotic gene expression. *Nature* **472:** 120–124

Wang, Y., Wang, X. and Paterson, A. H. 2012. Genome and gene duplications and gene expression divergence: a view from plants. *Annals of the New York Academy of Sciences* **1256:** 1-14

Wang, Y., Wang, X., Tang, H., Tan, X., Ficklin, S. P., Feltus, F. A. and Paterson, A. H. 2011. Modes of Gene Duplication Contribute Differently to Genetic Novelty and Redundancy, but Show Parallels across Divergent Angiosperms. *Plos One* **6: 12:** e28150

Weisenfeld, N. I., Kumar, V., Shah, P., Church, D. M. and Jaffe, D. B. 2017. Direct determination of diploid genome sequences. *Genome Research* **27:** 757-767

Xu, H., Luo, X., Qian, J., Pang, X., Song, J., Qian, G., Chen, J. and Chen, S. 2012. FastUniq: A Fast De Novo Duplicates Removal Tool for Paired Short Reads. *Plos One* **12:** e52249

Yang, Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution* **24: 8:** 1586-1591

Yang, Z. and Bielawski, J. P. 2000. Statistical methods for detecting molecular adaptation. *Trends in Ecology and Evolution*. **15: 12:** 496-503

Zhang, G., Cowled, C., Shi, Z., et al. 2013. Comparative analysis of bat genomes provides insight into the evolution of flight and immunity. *Science* **339: 6118:** 456-460

Zinner, D., Arnold, M. L. and Roos, C. 2011. The strange blood: natural hybridization in primates. *Evolutionary Anthropology: Issues, News, and Reviews* **20:** 96-103

Supplementary Information

Supplementary Table 2.1. Predicted bat lncRNAs. Adapted from Eckalbar et al (2016)

lncRNA database (DE)
lncRNA database (NDE)
conserved mammalian lncRNA (DE)
conserved mammalian lncRNA (NDE)
conserved bat lncRNA (DE)
conserved bat lncRNA (NDE)
M. natalensis lncRNA (DE)
M. natalensis lncRNA (NDE)

Transcript Code	Present in all bats	Present in humans	Present in mouse	Present in dog	Present in horse	Present in cat	Known lncRNA	Unique in Bats	In a bat	Unique in Mnat	CPS (average)	CPS (min)	Nearest Gene	Dist. to Gene (bp)	Mean normalsied read count	Adjusted p-value from expression
Mnat.G.20702	1	1	1	1	1	1	XIST	0	1	0	-0.72	-0.97	Unnamed	6815	28575.3	NA
Mnat.G.18271	1	1	1	1	1	1	HOXA11-AS	0	1	0	0.84	0.40	Hoxa10	836	839.6	2.17E-01
Mnat.G.18248	1	1	0	1	1	1	HOTTIP	0	1	0	-0.19	-1.07	Hoxa13	919	563.9	1.36E-17
Mnat.G.20575	1	1	1	1	1	1	LINC00643	0	1	0	1.42	1.42	Syt16	3265	94.5	3.28E-05
Mnat.G.16635	1	1	0	0	0	0	AC002310.7	0	1	0	0.65	0.65	Znf516	2132	97.5	8.00E-03
Mnat.G.21822	1	1	1	1	1	1	RMST	0	1	0	-0.88	-1.07	Nedd1	440521	40.3	3.20E-03
Mnat.G.12401	0	1	0	1	1	1	RP11-820L6.1	0	1	0	-0.81	-0.82	Unnamed	194141	65.4	5.91E-04
Mnat.G.14726	1	1	0	1	1	1	MEG3	0	1	0	0.44	-0.70	Unnamed	6399	53778.0	2.06E-01
Mnat.G.4633	1	1	1	1	1	1	RP11-463J10.3	0	1	0	1.13	1.13	C14orf166	32904	756.1	8.39E-03
Mnat.G.22244	1	1	0	1	1	1	CDR1	0	1	0	0.86	0.86	Unnamed	5312	333.7	5.74E-01
Mnat.G.19872	1	1	1	1	1	1	FBXL19-AS1	0	1	0	-1.16	-1.21	Fbxl19	1351	95.4	6.83E-03
Mnat.G.6547	1	1	0	1	1	1	RP11-346C20.4	0	1	0	-0.88	-0.88	Unnamed	465	21.1	5.74E-01
Mnat.G.14738	1	1	1	1	0	1		0	1	0	-0.79	-0.97	Unnamed	25936	191.7	2.75E-08
Mnat.G.17552	1	1	0	1	1	1		0	1	0	1.84	1.84	Terf1	2775	2959.1	7.94E-03
Mnat.G.14359	1	1	0	1	1	1		0	1	0	1.51	0.61	Mamdc4	842	144.8	6.78E-04

Mnat.G.22099	1	1	1	1	1	1		0	1	0	2.72	2.72	Samd1	749	2849.7	1.77E-01
Mnat.G.12279	1	1	1	1	1	1		0	1	0	0.70	0.70	Mppe1	275	448.2	8.48E-02
Mnat.G.5393	1	1	1	1	1	1		0	1	0	4.16	4.16	Acsl3	53083	525.1	1.49E-01
Mnat.G.8148	1	1	1	0	1	0	TBX5AS1	0	1	0	-0.38	-1.20	Tbx5	42468	382.0	4.60E-157
Mnat.G.11921	0	0	0	0	1	0		0	1	0	-0.83	-0.83	Scyl1	21208	4454.0	1.93E-13
Mnat.G.17493	1	0	0	1	0	1		0	1	0	-0.82	-0.82	Crispld1	656	202.3	2.08E-03
Mnat.G.10372	1	1	1	1	0	1		0	1	0	1.80	1.80	Slc1a2	19392	538.1	1.23E-05
Mnat.G.18236	0	0	0	0	0	0		0	1	0	0.00	-0.07	Unnamed	2834	56.9	9.38E-08
Mnat.G.17046	0	0	0	0	0	0		0	1	0	-0.51	-0.51	Slc16a3	2245	518.6	1.51E-02
Mnat.G.14830	1	1	0	1	0	0		0	1	0	1.67	1.67	Unnamed	1371	165.2	7.27E-02
Mnat.G.10970	1	1	1	1	1	1		0	1	0	1.88	1.88	Ret	8985	121.8	1.36E-02
Mnat.G.2914	0	0	0	0	0	0		0	1	0	-0.70	-0.82	Msantd1	1362	197.7	1.96E-01
Mnat.G.9156	1	1	1	1	1	1		0	1	0	-1.01	-1.01	Wnt5a	4462	113.0	7.09E-05
Mnat.G.19613	1	1	1	1	1	1		0	1	0	10.92	10.92	Slc25a13	13460	88.4	1.67E-02
Mnat.G.21131	1	1	1	1	1	1		0	1	0	2.26	2.26	Prkg1	14343	87.5	2.14E-01
Mnat.G.4609	1	1	0	1	1	1		0	1	0	0.62	0.62	Anp32a	581	318.2	7.03E-05
Mnat.G.6127	1	0	0	0	1	1		0	1	0	-1.18	-1.23	Unnamed	2614	84.9	4.40E-03
Mnat.G.3610	1	1	1	1	1	1		0	1	0	-0.16	-0.16	Erbb4	209905	68.9	4.99E-12
Mnat.G.16202	0	1	0	1	1	1		0	1	0	2.21	2.21	Fam209	1665	130.0	1.47E-04
Mnat.G.10903	1	0	0	0	0	1		0	1	0	-0.57	-0.57	Smarca5	15681	38.6	1.51E-03
Mnat.G.387	1	1	0	1	1	1		0	1	0	-0.55	-0.64	Spdyb	55573	54.0	5.48E-02
Mnat.G.15160	1	1	1	1	1	1		0	1	0	-0.29	-0.53	Unnamed	77423	25.4	1.08E-03
Mnat.G.10390	0	0	0	1	0	0		0	1	0	-0.03	-0.04	Unnamed	19676	53.1	1.04E-04
Mnat.G.11084	0	0	1	1	1	1		0	1	0	-0.86	-0.95	Unnamed	1935	9.2	3.32E-01
Mnat.G.5442	0	0	0	0	0	0		0	1	0	-1.00	-1.00	Unnamed	3996	34.0	2.66E-01
Mnat.G.15837	0	0	0	0	0	0		0	1	0	-0.83	-0.83	Klf4	14274	6.5	NA
Mnat.G.4151	1	0	0	0	1	0		0	1	0	0.33	0.33	Catsper3	25417	15.5	2.53E-21
Mnat.G.1325	1	1	0	0	1	1		0	1	0	-0.64	-0.64	Unnamed	1498	4.6	NA
Mnat.G.17860	1	0	0	1	1	0		0	1	0	-0.64	-0.64	Mrpl23	38461	51242.5	7.71E-01
Mnat.G.15243	1	1	0	0	0	0		0	1	0	0.80	0.61	Ccr6	38827	3261.5	9.22E-01
Mnat.G.4899	1	1	1	1	1	1		0	1	0	1.87	1.87	Kremen1	155	760.8	6.05E-01
Mnat.G.13087	1	1	0	1	1	1		0	1	0	-0.72	-1.05	Glyctk	5908	1326.1	7.00E-01
Mnat.G.16533	1	1	1	1	1	1		0	1	0	10.42	10.42	Epha8	38743	641.5	2.40E-01
Mnat.G.14923	1	1	1	1	1	1		0	1	0	3.35	2.75	Unnamed	9133	546.3	5.09E-01
Mnat.G.20151	1	0	0	0	0	1		0	1	0	0.80	0.73	Kmt2e	2015	1921.1	1.47E-01

Mnat.G.7568	1	1	1	1	1	1	0	1	0	5.36	5.36	Nr1d1	3967	1021.8	5.64E-01
Mnat.G.14725	0	1	0	1	1	1	0	1	0	-1.11	-1.42	Unnamed	383	659.5	1.30E-02
Mnat.G.22728	1	1	1	1	1	1	0	1	0	9.82	9.82	Slc28a3	189547	65.7	4.58E-01
Mnat.G.12320	1	0	0	0	0	1	0	1	0	0.69	0.69	Ippk	95	709.0	3.54E-01
Mnat.G.19010	1	1	1	0	1	1	0	1	0	-0.94	-0.94	Itga5	10	10.2	3.95E-01
Mnat.G.23783	1	1	0	1	1	1	0	1	0	0.68	-0.37	Unnamed	9047	700.4	6.35E-01
Mnat.G.13746	0	0	0	0	0	0	0	1	0	-0.97	-0.97	Lmx1b	656	11.6	2.76E-03
Mnat.G.16930	1	1	1	1	0	1	0	1	0	-0.73	-0.94	Kcna6	67	491.5	4.73E-02
Mnat.G.5144	0	1	1	1	1	1	0	1	0	-0.44	-0.99	Unnamed	111089	471.4	9.51E-01
Mnat.G.10524	0	1	0	1	1	0	0	1	0	-0.97	-1.04	Unnamed	18868	33.2	3.98E-01
Mnat.G.4569	1	1	0	1	1	1	0	1	0	-0.99	-1.27	Unnamed	110	469.9	1.72E-01
Mnat.G.9677	1	1	1	1	1	1	0	1	0	10.65	10.65	Hivep3	1136	5.5	NA
Mnat.G.18622	1	0	0	0	1	0	0	1	0	1.12	0.80	Unnamed	26488	9.1	1.18E-02
Mnat.G.5464	1	0	0	0	1	1	0	1	0	1.90	1.87	Cyp4x1	13527	37.8	5.96E-01
Mnat.G.15391	1	1	0	1	1	1	0	1	0	2.94	2.83	Unnamed	61946	362.8	2.41E-01
Mnat.G.16103	1	1	0	1	1	1	0	1	0	-1.27	-1.37	Tmem255a	14769	47.2	9.18E-01
Mnat.G.14185	0	0	0	1	0	1	0	0	0	0.08	0.08	C1orf106	330	281.6	3.30E-02
Mnat.G.2597	1	1	1	1	1	1	0	1	0	14.07	14.07	Cul1	123238	193.1	4.33E-02
Mnat.G.5822	0	0	0	1	1	0	0	1	0	0.46	0.46	Gzmk	31250	53.4	9.54E-02
Mnat.G.12937	1	0	0	1	1	1	0	1	0	0.14	0.03	Epb41l5	8226	212.0	3.96E-01
Mnat.G.23733	1	1	1	1	1	1	0	1	0	2.00	2.00	Plek	24117	134.4	3.06E-04
Mnat.G.4083	1	1	1	1	1	1	0	1	0	-0.99	-1.07	Pde3a	2190	41.3	1.57E-01
Mnat.G.10786	1	1	0	1	1	1	0	1	0	1.34	1.01	Unnamed	14695	191.1	1.08E-01
Mnat.G.5356	0	0	0	1	1	1	0	0	0	-0.92	-1.15	Unnamed	6191	164.9	9.74E-01
Mnat.G.2101	0	0	0	0	0	0	0	1	0	-1.22	-1.26	Slc19a1	787	18.1	1.23E-02
Mnat.G.21130	1	1	0	0	0	0	0	1	0	1.01	1.01	Gm14446	570	37.3	5.81E-01
Mnat.G.17480	0	0	0	0	1	0	0	1	0	2.93	2.93	Tg	7041	5.4	NA
Mnat.G.2950	1	1	1	1	1	1	0	1	0	0.59	0.59	Alpi	12931	361.0	1.00E-01
Mnat.G.17269	1	1	0	1	1	1	0	1	0	-0.67	-0.77	4732465j04rik	15120	167.0	8.01E-01
Mnat.G.15820	0	0	0	0	1	0	0	1	0	-0.91	-0.91	Cyth3	11	237.6	1.16E-01
Mnat.G.22165	1	1	0	1	1	1	0	1	0	1.61	1.61	Plekha8	28671	302.8	4.64E-01
Mnat.G.21175	1	0	1	1	1	1	0	1	0	2.05	2.05	Hcst	159	8.6	5.68E-01
Mnat.G.20239	0	1	0	0	0	1	0	1	0	1.61	1.61	Cep170b	291	8.9	3.26E-04
Mnat.G.10604	0	1	0	1	1	1	0	1	0	-0.92	-0.92	Mcmdc2	931	266.9	6.92E-01
Mnat.G.3983	1	0	0	0	0	1	0	1	0	-0.08	-0.08	Prex1	1843	634.6	3.70E-01

Mnat.G.10451	0	0	0	0	0	0	0	1	0	-0.98	-1.11	Unnamed	2307	110.3	6.67E-01
Mnat.G.1440	1	1	1	1	1	1	0	1	0	0.12	0.12	Unnamed	14334	54.8	2.48E-01
Mnat.G.3442	0	0	0	0	1	0	0	1	0	-0.97	-0.97	Asb6	7546	10.1	1.06E-01
Mnat.G.9161	0	1	0	1	1	1	0	1	0	-1.04	-1.08	Unnamed	2440	63.7	2.43E-01
Mnat.G.12987	1	1	1	1	1	1	0	1	0	8.06	3.63	Slc6a6	4356	23.3	1.97E-01
Mnat.G.22075	0	1	0	1	1	1	0	1	0	0.13	0.07	Nrxn1	327619	4.5	NA
Mnat.G.5310	1	1	1	1	1	1	0	1	0	-0.53	-1.30	Sfrp4	14652	101.9	6.16E-01
Mnat.G.7370	1	0	0	1	0	0	0	1	0	-1.02	-1.02	Gemin4	112	107.6	1.08E-01
Mnat.G.11333	0	0	0	0	0	0	0	1	0	-0.49	-0.49	Cd34	46	32.8	1.00E-01
Mnat.G.11560	1	1	0	1	1	1	0	1	0	1.48	1.48	Unnamed	3117	45.7	4.48E-01
Mnat.G.18190	0	0	0	0	0	0	0	1	0	-1.04	-1.05	Unnamed	3297	116.3	1.20E-01
Mnat.G.585	0	0	0	0	0	0	0	1	0	-0.94	-0.99	Unnamed	12289	97.8	3.57E-01
Mnat.G.6887	0	1	0	0	1	1	0	0	0	-0.59	-0.59	Dbx2	55593	59.4	7.50E-01
Mnat.G.10315	1	1	1	1	1	1	0	1	0	1.67	1.67	0610012h03rik	87201	5.7	NA
Mnat.G.15693	1	1	1	1	1	1	0	1	0	2.51	2.45	Cstb	3189	54.0	2.83E-01
Mnat.G.1306	1	1	0	1	0	0	0	1	0	-0.95	-0.95	Unnamed	5453	23.7	9.28E-01
Mnat.G.17428	0	0	0	0	0	0	0	1	0	-0.81	-0.81	Myct1	910	62.8	5.79E-01
Mnat.G.3468	1	1	1	1	1	1	0	1	0	2.71	2.71	Adamtsl2	8342	47.0	8.30E-01
Mnat.G.3285	1	1	1	1	1	1	0	1	0	-0.38	-0.38	Dgat2	95	18.4	6.95E-01
Mnat.G.14825	1	0	0	0	1	1	0	1	0	0.62	0.62	Slc16a5	14	48.0	7.88E-01
Mnat.G.15392	1	0	1	1	1	1	0	1	0	-1.15	-1.15	Rnpc3	1741	25.5	7.28E-01
Mnat.G.13334	1	1	1	1	1	1	0	1	0	-1.08	-1.08	Xrcc4	13	247.5	8.82E-01
Mnat.G.1303	0	0	0	0	0	0	0	1	0	-0.68	-1.10	Kcnu1	1061	2.8	NA
Mnat.G.18281	1	1	1	1	1	1	0	1	0	-1.08	-1.08	Tra2a	2470	53.7	1.45E-01
Mnat.G.15262	0	0	0	0	0	0	0	1	0	-0.07	-0.12	Adap1	426	53.7	8.50E-01
Mnat.G.4967	0	0	0	0	0	0	0	1	0	-0.37	-0.57	Unnamed	3390	36.4	7.25E-01
Mnat.G.2662	1	1	1	1	1	1	0	1	0	1.51	1.51	Tspan14	101998	10.8	2.61E-01
Mnat.G.14577	0	0	0	0	0	0	0	1	0	-1.00	-1.00	Tasp1	14277	31.6	9.93E-01
Mnat.G.19143	1	1	1	0	1	0	0	1	0	-1.02	-1.10	Ccdc149	25623	76.6	1.21E-01
Mnat.G.5243	0	0	0	1	0	0	0	1	0	-0.55	-0.60	Prkrip1	2381	46.2	4.03E-01
Mnat.G.2043	0	0	0	0	0	0	0	1	0	-0.95	-0.95	Fam207a	36547	47.2	7.99E-01
Mnat.G.6984	0	0	0	1	0	1	0	1	0	-1.42	-1.42	Cpne8	344	6.0	NA
Mnat.G.16272	0	0	0	0	0	0	0	1	0	-0.81	-0.81	Unnamed	202780	38.9	1.75E-01
Mnat.G.6902	1	1	0	1	1	1	0	1	0	0.83	0.83	Unnamed	166	51.1	2.37E-02
Mnat.G.9635	0	0	0	1	1	1	0	0	0	-1.03	-1.03	Dars	21	14.6	4.35E-01

Mnat.G.6480	0	0	1	1	1	1	0	1	0	-0.12	-0.12	Unnamed	546	17.5	1.21E-01
Mnat.G.18269	1	1	0	1	1	1	0	1	0	-1.05	-1.05	Unnamed	17453	31.1	6.20E-01
Mnat.G.19865	0	0	0	0	0	0	0	1	0	-0.89	-1.05	Prr14	1436	22.7	6.70E-01
Mnat.G.12098	1	0	0	0	1	0	0	1	0	-0.53	-0.53	Vwa9	828	34.2	1.24E-02
Mnat.G.22016	1	1	0	0	1	1	0	1	0	1.25	1.25	Chid1	512	45.6	1.17E-01
Mnat.G.1153	1	1	0	1	1	1	0	1	0	-0.85	-0.85	Kank4	12146	26.4	7.48E-01
Mnat.G.19959	0	0	0	0	0	0	0	1	0	-0.23	-0.23	Aar2	6309	167.8	6.72E-01
Mnat.G.21437	0	0	0	0	1	0	0	1	0	-0.95	-1.29	Nvl	1003	36.3	3.78E-01
Mnat.G.16636	0	0	1	0	0	0	0	1	0	0.51	0.51	Zfp236	5330	10.0	8.01E-01
Mnat.G.14902	1	0	0	1	1	0	0	1	0	-0.96	-1.31	Ppp4r4	3388	26.4	4.11E-01
Mnat.G.7102	1	0	0	1	1	1	0	1	0	-0.18	-0.18	Trpv4	18	7.3	NA
Mnat.G.14138	0	0	0	0	0	0	0	1	0	-0.98	-1.07	Tmem200c	764	50.0	7.60E-01
Mnat.G.23486	0	0	0	0	0	0	0	1	0	-0.86	-0.86	Tab2	4300	54.0	4.27E-01
Mnat.G.6524	1	1	1	1	1	1	0	1	0	-0.43	-0.43	Unnamed	465	27.0	1.40E-01
Mnat.G.22789	0	0	0	0	0	0	0	1	0	-1.22	-1.33	Anp32b	3686	58.7	8.58E-01
Mnat.G.14401	0	1	0	1	0	1	0	1	0	-0.51	-0.69	Zmynd19	4092	17.9	4.11E-01
Mnat.G.13877	1	1	1	1	1	1	0	1	0	-0.24	-0.24	Smg6	22974	8.2	1.36E-01
Mnat.G.24065	1	1	1	1	1	1	0	1	0	0.33	-0.29	Skida1	913	30.9	7.32E-01
Mnat.G.4000	0	0	0	1	1	1	0	1	0	-0.18	-0.18	Cidea	4015	40.6	6.56E-01
Mnat.G.3184	0	0	0	0	0	0	0	1	0	-0.28	-0.28	Unnamed	102	94.8	1.14E-01
Mnat.G.11952	0	0	1	1	0	1	0	1	0	-1.07	-1.08	Ehd1	8134	27.6	7.11E-01
Mnat.G.12119	0	0	0	0	0	0	0	1	0	0.30	0.30	Tgfb3	86	16.3	5.02E-02
Mnat.G.10423	0	0	0	0	0	0	0	1	0	-0.87	-0.87	Cbx2	835	42.8	8.68E-01
Mnat.G.5817	0	0	0	0	0	0	0	1	0	-0.95	-0.99	Rab3c	16069	5.7	NA
Mnat.G.9896	0	0	0	0	0	0	0	1	0	-0.24	-0.24	Hesx1	59431	28.5	1.94E-01
Mnat.G.9045	0	0	0	0	0	0	0	1	0	-0.67	-0.69	Snx19	11762	17.1	5.84E-01
Mnat.G.11252	0	0	0	1	1	0	0	1	0	-0.62	-0.72	Hmgn3	1593	26.3	8.07E-01
Mnat.G.6947	1	1	0	1	1	1	0	1	0	-0.72	-0.72	Kif26b	2051	11.6	1.34E-01
Mnat.G.13694	0	0	0	0	0	0	0	1	0	-0.42	-0.42	Lzts1	348	46.6	6.91E-01
Mnat.G.11202	0	0	0	0	1	1	0	1	0	-1.18	-1.24	Col12a1	193641	26.2	4.21E-02
Mnat.G.14175	0	1	0	1	1	1	0	1	0	-0.08	-0.08	Tnni1	563	36.9	9.73E-01
Mnat.G.823	1	1	0	0	1	1	0	1	0	-1.10	-1.10	Stmn3	2384	4.4	NA
Mnat.G.2985	0	0	0	1	0	1	0	0	0	-1.08	-1.08	Znf521	1049	18.4	5.23E-04
Mnat.G.3195	0	0	0	0	0	0	0	1	0	-0.49	-0.49	Unnamed	102	46.3	3.95E-01
Mnat.G.24051	0	0	0	1	0	0	0	1	0	-0.78	-0.78	Commd3	59971	7.1	NA

Mnat.G.4049	0	0	0	0	0	0	0	1	0	-0.93	-0.93	Ifltd1	3851	28.9	4.54E-01
Mnat.G.16773	1	1	1	1	1	1	0	1	0	-0.18	-0.18	Krt27	2625	11.8	9.11E-01
Mnat.G.18258	0	0	0	0	0	0	0	1	0	-1.13	-1.13	Twist1	2045	19.9	3.43E-02
Mnat.G.1414	0	0	0	0	0	0	0	1	0	-1.04	-1.09	Irx2	2422	27.7	1.18E-01
Mnat.G.20615	0	0	0	0	1	0	0	1	0	-0.48	-0.48	Hace1	55385	2.2	NA
Mnat.G.17483	0	0	0	1	1	1	0	1	0	0.55	0.55	Unnamed	5791	4.7	NA
Mnat.G.1998	1	1	1	1	1	1	0	1	0	-1.33	-1.33	Irf2bpl	2504	18.2	9.05E-01
Mnat.G.2247	1	1	0	0	0	0	0	1	0	-0.02	-0.02	Elfn2	87	10.1	2.04E-01
Mnat.G.21081	0	0	0	1	0	0	0	0	0	-1.07	-1.07	Ide	7648	10.7	6.29E-01
Mnat.G.20245	1	1	0	1	1	1	0	1	0	-0.93	-0.93	Tbx3	1362	9.4	5.18E-04
Mnat.G.23022	0	1	0	0	0	1	0	1	0	-1.15	-1.15	Kcnc4	1433	5.2	NA
Mnat.G.4131	0	0	0	0	1	1	0	0	0	-1.04	-1.18	Loh12cr1	2261	8.5	3.48E-01
Mnat.G.8858	0	0	0	0	0	0	0	1	0	-0.78	-0.78	Six3	10698	9.4	4.03E-02
Mnat.G.18326	0	0	0	0	0	0	0	1	0	-0.95	-0.95	Mcur1	81898	21.4	3.03E-01
Mnat.G.5583	0	1	0	1	0	1	0	1	0	1.24	1.24	Unnamed	10453	17.0	2.18E-01
Mnat.G.10839	0	0	0	0	0	0	0	1	0	-0.77	-0.77	Grm5	37690	6.1	NA
Mnat.G.2755	1	1	0	1	0	1	0	1	0	0.58	0.58	Csgalnact1	147	27.7	9.95E-01
Mnat.G.14430	0	0	0	0	0	0	0	1	0	-0.28	-0.28	Unnamed	3789	5.9	NA
Mnat.G.8076	0	0	0	0	0	0	0	1	0	-0.78	-0.78	Actrt3	573	6.2	NA
Mnat.G.18311	0	1	0	0	0	0	0	0	0	-0.66	-0.66	Cttn	2618	5.2	NA
Mnat.G.7927	0	0	0	0	0	1	0	1	0	0.00	0.00	Nradd	20618	6.2	NA
Mnat.G.2327	0	1	0	1	1	1	0	1	0	0.07	-0.34	Egflam	5798	9.2	8.34E-01
Mnat.G.19004	1	1	0	1	1	0	0	1	0	-0.76	-0.76	Hoxc4	164	6.9	NA
Mnat.G.22673	0	0	0	0	0	1	0	1	0	-0.95	-0.95	Anks6	6098	13.0	7.82E-01
Mnat.G.6136	0	0	0	0	0	0	0	1	0	-0.93	-0.93	Hand2	1760	13.4	1.98E-01
Mnat.G.550	0	0	0	0	0	0	0	1	0	-0.88	-0.88	Rassf3	18861	10.9	7.49E-01
Mnat.G.5112	0	0	0	0	0	0	0	1	0	-1.02	-1.02	Susd5	3609	13.2	6.79E-01
Mnat.G.23661	0	0	0	0	0	0	0	1	0	-0.72	-0.73	Eef1b	86283	3.2	NA
Mnat.G.14872	1	1	1	1	1	1	0	1	0	-0.88	-0.88	Unnamed	1515	5.2	NA
Mnat.G.20473	0	0	0	0	0	0	0	1	0	-0.75	-0.75	Hpse2	3459	2.6	NA
Mnat.G.5592	0	0	0	0	0	0	0	1	0	-1.02	-1.02	Unnamed	10453	4.9	NA
Mnat.G.1686	0	0	0	0	0	0	0	1	0	-1.17	-1.17	Cdk18	56991	2.3	NA
Mnat.G.11568	0	1	0	0	1	0	0	1	0	-0.62	-0.62	Unnamed	3117	3.5	NA
Mnat.G.9028	1	0	0	0	0	0	1	1	0	-1.10	-1.10	2610318n02rik	1610	110.7	1.15E-01
Mnat.G.18995	1	0	0	0	0	0	1	1	0	-0.56	-0.56	Hoxc8	915	2.3	NA

Mnat.G.18496	1	0	0	0	0	0	1	1	0	-1.03	-1.03	Ankh	122589	19.8	1.64E-01
Mnat.G.77	1	0	0	0	0	0	1	1	0	-0.99	-1.02	Atp5g3	83100	10.1	9.54E-01
Mnat.G.14017	1	0	0	0	0	0	1	1	0	-0.75	-0.78	Dnase1	2059	53.3	9.57E-01
Mnat.G.19902	0	0	0	0	0	0	0	0	1	-0.33	-0.40	Unnamed	24793	914.8	9.43E-01
Mnat.G.18192	0	0	0	0	0	0	0	0	1	-1.36	-1.43	Unnamed	3297	72.2	1.34E-04
Mnat.G.13126	0	0	0	0	0	0	0	0	1	-0.64	-0.64	Nup50	8169	302.2	3.78E-01
Mnat.G.17529	0	0	0	0	0	0	0	0	1	-0.41	-0.41	Znf704	508	311.9	5.83E-01
Mnat.G.17476	0	0	0	0	0	0	0	0	1	-0.43	-0.43	Pag1	114	113.6	3.67E-05
Mnat.G.24119	0	0	0	0	0	0	0	0	1	-0.59	-0.69	Unnamed	37169	12.5	9.31E-02
Mnat.G.2844	0	0	0	0	0	0	0	0	1	-0.36	-0.36	Mitf	1832	11.3	2.94E-01
Mnat.G.24066	0	0	0	0	0	0	0	0	1	-1.20	-1.20	Arhgap21	2529	22.3	1.24E-02
Mnat.G.18468	0	0	0	0	0	0	0	0	1	1.23	1.23	Tmem206	8901	13.1	8.84E-01
Mnat.G.18473	0	0	0	0	0	0	0	0	1	3.00	3.00	Unnamed	21980	144.3	9.42E-01
Mnat.G.17241	0	0	0	0	0	0	0	0	1	-1.02	-1.02	Unnamed	1044	4.6	NA
Mnat.G.24092	0	0	0	0	0	0	0	0	1	-0.63	-0.63	Unnamed	6266	10.0	3.21E-01
Mnat.G.14736	0	0	0	0	0	0	0	0	1	-0.69	-0.69	Tecpr2	23	182.9	8.23E-02
Mnat.G.12299	0	0	0	0	0	0	0	0	1	-0.51	-0.51	Unnamed	25319	58.1	1.53E-01
Mnat.G.15151	0	0	0	0	0	0	0	0	1	-0.89	-0.89	Runx2	12590	3.4	NA
Mnat.G.22328	0	0	0	0	0	0	0	0	1	-0.16	-0.16	Gas7	2478	119.4	5.27E-01
Mnat.G.6454	0	0	0	0	0	0	0	0	1	-0.76	-0.76	Unnamed	2605	14.0	6.99E-02
Mnat.G.5368	0	0	0	0	0	0	0	0	1	-0.64	-0.64	Unnamed	1969	21.3	4.33E-01
Mnat.G.23400	0	0	0	0	0	0	0	0	1	-0.54	-0.54	Plekhh1	253	177.2	5.31E-01
Mnat.G.21715	0	0	0	0	0	0	0	0	1	-1.30	-1.30	C14orf80	3942	19.8	2.35E-02
Mnat.G.6130	0	0	0	0	0	0	0	0	1	-1.08	-1.08	Unnamed	2614	31.6	7.08E-01
Mnat.G.4316	0	0	0	0	0	0	0	0	1	-0.43	-0.82	Mctp2	33795	23.4	8.80E-01
Mnat.G.10561	0	0	0	0	0	0	0	0	1	-1.35	-1.35	Foxc1	4724	19.2	3.94E-02
Mnat.G.3788	0	0	0	0	0	0	0	0	1	-0.67	-0.67	Usp9x	10961	22.9	5.09E-01
Mnat.G.21006	0	0	0	0	0	0	0	0	1	-1.22	-1.22	Fam195a	9272	14.5	3.79E-02
Mnat.G.19511	0	0	0	0	0	0	0	0	1	-1.06	-1.06	Slc25a29	9173	5.5	NA
Mnat.G.6617	0	0	0	0	0	0	0	0	1	0.00	0.00	C17orf100	29	35.2	6.16E-01
Mnat.G.2295	0	0	0	0	0	0	0	0	1	-1.21	-1.21	Osmr	225	4.3	NA
Mnat.G.8514	0	0	0	0	0	0	0	0	1	-0.95	-1.19	Unnamed	9781	12.3	4.33E-01
Mnat.G.14498	0	0	0	0	0	0	0	0	1	-0.57	-0.57	Adam33	6226	8.9	3.11E-01
Mnat.G.2652	0	0	0	0	0	0	0	0	1	-0.61	-0.61	Zfp282	3403	10.7	9.53E-01
Mnat.G.17284	0	0	0	0	0	0	0	0	1	-1.12	-1.12	Arl14	13375	7.6	NA

Mnat.G.18867	0	0	0	0	0	0	0	0	1	-1.20	-1.20	Snx2	14744	9.0	9.10E-01
Mnat.G.10554	0	0	0	0	0	0	0	0	1	-0.74	-0.74	Foxf2	46996	7.3	NA

Supplementary Table 3.1: Sites with significantly disproportionate representation within the Amboseli population. Rows are coloured blue and red, according to whether they were disproportionately of *P. anubis* or *P. cynocephalus* ancestry, respectively. The 'Missing Data', 'Heterozygous', '*P. anubis*' and '*P. cynocephalus*' columns display the number of Amboseli individuals that were classified into group. The final column contains the adjusted p-values, after multiple testing correction. Rows are ordered by species and then by p-value.

Scaffold No.	Base Position	Missing Data	Heterozygous	*P. anubis*	*P. cynocephalus*	Adj. Prob.
scaffold1237	644944	2	0	21	0	8.17E-07
scaffold1815	3721118	2	0	21	0	1.01E-06
scaffold599	2981992	3	0	20	0	4.46E-06
scaffold1979	1090419	3	0	20	0	6.76E-06
scaffold414	3395466	2	1	20	0	2.26E-05
scaffold6730	350301	3	0	20	0	2.58E-05
scaffold5212	64419	3	0	20	0	3.09E-05
scaffold100	176007	3	0	19	1	3.3E-05
scaffold210	1129233	4	0	19	0	4.88E-05
scaffold773	1716708	4	0	19	0	7.05E-05
scaffold2893	1983678	1	1	20	1	8.97E-05
scaffold158	56009	3	0	19	1	0.00015
scaffold942	70896	3	0	19	1	0.00018
scaffold575	36179	4	0	19	0	0.000188
scaffold1805	770364	4	0	19	0	0.000234
scaffold127	1408966	5	0	18	0	0.000239
scaffold4807	16662	4	0	19	0	0.000365
scaffold820	672190	4	0	19	0	0.000485
scaffold1817	483076	3	0	19	1	0.000682
scaffold100	175008	4	0	18	1	0.000827
scaffold2005	478061	4	0	19	0	0.000862
scaffold16	1334630	5	0	17	1	0.00106

scaffold341	2370585	5	0	18	0	0.001195
scaffold227	6967057	5	0	18	0	0.00148
scaffold2791	167352	3	1	19	0	0.001905
scaffold693	1189362	5	0	18	0	0.001974
scaffold271	503930	5	0	18	0	0.002509
scaffold1691	249974	3	0	19	1	0.002898
scaffold659	1983449	4	0	18	1	0.00356
scaffold187	455374	5	0	17	1	0.003691
scaffold337	1991376	5	1	17	0	0.004205
scaffold21	2415702	5	0	17	1	0.004286
scaffold16	1364416	4	2	16	1	0.004759
scaffold16	1159065	3	2	16	2	0.005209
scaffold1407	1407553	3	0	18	2	0.005387
scaffold3353	108204	3	2	18	0	0.006739
scaffold1827	102187	2	2	18	1	0.006876
scaffold575	70040	3	2	18	0	0.007623
scaffold1693	27379	3	1	18	1	0.009372
scaffold4711	2070631	5	0	18	0	0.009593
scaffold1541	1161457	4	0	18	1	0.009614
scaffold890	545869	4	1	17	1	0.01047
scaffold16	1426523	5	1	16	1	0.010789
scaffold2119	78006	5	0	18	0	0.010894
scaffold1390	261315	3	0	18	2	0.012347
scaffold2779	212446	3	1	18	1	0.01267
scaffold16	1334709	4	1	16	2	0.014815
scaffold2337	47099	2	1	18	2	0.016374
scaffold1089	210852	5	1	17	0	0.017558
scaffold4006	1773416	4	1	18	0	0.018907
scaffold1523	26896	3	0	18	2	0.019989

scaffold4954	236585	3	0	18	2	0.020347
scaffold6470	77230	3	1	18	1	0.022295
scaffold367	39521	5	1	17	0	0.024403
scaffold138	730522	1	1	0	21	0.005258
scaffold806	1577476	2	0	0	21	0.013801
scaffold763	1523405	0	2	0	21	0.014162
scaffold2116	500838	1	0	1	21	0.024995

Supplementary Table 4.1: Phylogenetic patterns observed in Ruschioideae sequences which aligned to the first and last 50 amino acids of 30 *Arabidopsis thaliana* genes. The genes were filtered such that they had at least 2 copies in each of the Ruschioideae species. An "early" divergence pattern indicated that the gene was duplicated before the divergence of the three species. A "middle" divergence indicates that the gene was duplicated after the divergence of *Cleretum herrei* from the two Ruschieae species. A "late" duplication happened in each species independently. Note that one gene (AT1G23490.1) had the last 50 amino acids fail to align in a multiple sequence alignment and is therefore excluded from the analysis totals.

	First 50 amino acids			Last 50 amino acids		
Gene Name	Early	Middle	Late	Early	Middle	Late
AT1G02500.1	1	0	0	1	0	0
AT1G08350.1	1	0	0	1	0	0
AT1G10350.1	1	0	0	1	0	0
AT1G13370.1	1	1	1	1	0	1
AT1G16920.1	1	1	0	1	0	0
AT1G19890.1	1	1	1	1	0	1
AT1G21530.1	1	1	0	1	0	0
AT1G23490.1	1	0	0	N/A	N/A	N/A
AT1G31340.1	1	0	0	1	0	0
AT1G35550.1	0	0	0	0	0	0
AT1G41920.1	1	0	1	1	1	1
AT1G50920.1	0	1	0	0	0	0
AT1G52150.3	1	0	0	1	0	0
AT1G55060.1	1	0	0	1	1	0
AT1G59725.1	1	0	0	1	0	0
AT1G62020.1	1	0	0	0	0	0
AT1G75600.1	1	1	1	1	0	1
AT1G78360.1	1	0	0	1	0	0
AT1G80530.1	1	0	0	1	0	0
AT2G03520.1	0	0	0	1	1	0
AT2G03530.1	1	0	0	1	1	0
AT2G21220.1	0	0	0	1	0	0
AT2G21390.1	1	0	0	1	0	0
AT2G27030.2	1	0	0	1	0	0
AT2G32220.1	0	0	0	0	0	0
AT3G08510.1	1	0	0	1	0	0
AT3G15060.1	1	0	1	1	0	0
AT3G17390.1	1	0	0	1	0	0
AT3G21460.1	1	0	0	1	0	0
AT3G22230.1	0	0	0	0	0	0
Total	23	6	5	24	4	4