Genome-wide Analysis of Selection in Insects, Mammals and Fungi

Kate E. Ridout

St John's College

Department of Plant Sciences

Supervisor

Dr. Dmitry A. Filatov

Secondary Supervisor

Prof. John Pannell

D.Phil. Thesis submitted to the University of Oxford for the degree of Doctor of Philosophy of Science

Oxford, October 2011

ACKNOWLEDGEMENTS

The research in this thesis was supported by a grant from the Biotechnology and Biological Sciences Research Council (BBSRC), whom I would like to thank for allowing me to undertake this work.

I would like to thank my supervisor Dmitry Filatov for always being honest, and for allowing me so much freedom in my research whilst keeping me on a sensible track. I would also like to thank my second supervisor John Pannell for providing guidance whenever it was needed.

My lasting thanks to those members of the Filatov lab who were always patient and willing to answer my stupid questions; to Chris, for making me work faster than I had thought was humanly possible and numerous scientific discussions, to Graham, for helping me to understand population genetics and how to eat healthily, to Maxim, for helping me understand PAML, to Mark, for constant reassurance that my chapters were OK, and to Tonya, for feeding me.

Thank you to the members of the ecology lab who kept me diverted and sane, including Gill, Nodoka, Lizzie, Steven, Kirsty, Tonya and Christina, and to the expended members of the department who have helped me along the way.

Thanks also to my singing teacher Lesley and to Steve, Nalin, Chris, Doug, Paul B, Alex, Paul K, Jon, Rich, Ant and Duncan of the IMMposters, who gave me purpose when chapters were rocky, and to Dave and Rachel for letting me call them any time with any problem.

Finally, and most importantly, to Mom and Dad, Josie (and Sam who proved to be a distraction even before he was born!), Gungy and Grandad and Aunty Gilly, who have provided me with unquestionable support, food and lodgings whenever they were needed, diversions when they were (and weren't) needed, and for all the enforced marches my mother made me take 'for my health'.

GENOME-WIDE ANALYSIS OF SELECTION IN MAMMALS, INSECTS AND FUNGI

KATE E. RIDOUT
ST JOHN'S COLLEGE

THESIS SUBMITTED FOR THE DEGREE OF DOCTOR OF PHILOSOPHY OF SCIENCE

Trinity term 2011

Characterising and understanding factors that affect the rate of molecular evolution in proteins has played a major part in the development of evolutionary theory. The early analyses of amino acid substitutions stimulated the development of the neutral theory of molecular evolution, which later evolved into the nearly neutral theory. More recent work has lead to a better understanding of the role selection plays at the molecular level, but there is still limited understanding of how higher levels of protein organisation affect the way natural selection acts. The investigation of this question is the central aim of this thesis, which is addressed via the analysis of selective pressures in secondary protein structures in insects, mammals and fungi. The analyses for the first two groups were conducted using publically available datasets. To conduct the analyses in fungi, genome sequence data from the fungal genus *Microbotryum* (sequenced in our laboratory) was assembled and annotated, resulting in the development of a number of bioinformatics tools which are described here. The fungal, insect and mammalian datasets were interrogated with regard to a number of structural features, such as protein secondary structure, position of a site with regard to adaptively evolving sites, hydropathy and solvent-accessibility. These features were correlated with the signals of positive and purifying selection detected using phylogenetic maximum likelihood and Bayesian approaches. I conclude that all of the factors examined can have an effect on the rate of molecular evolution. In particular, disordered and hydrophilic regions of the protein are found to experience fewer physiochemical constraints and contain a higher proportion of adaptively evolving sites. It is also revealed that positively selected residues are 'clustered' together spatially, and these trends persist in the three taxa. Finally, I show that this variation in adaptive evolution is a result of both selective events and physiochemical constraint.

TABLE OF CONTENTS

1. INTRODUCTION

Kate E. Ridout

It has long been the goal of genetics researchers to understand the molecular basis of evolution. Molecular evolution begins at the nucleotide level. Understanding why a nucleotide variant exists at a particular position at a given locus is the start to understanding how different proteins, systems and organisms arose, and how they work. Investigating the effects of changes in the DNA is therefore the key to understanding why and how proteins function in different ways, and how those functions can change.

Using nucleotide data it has become possible to study the molecular footprints of natural selection (Kreitman and Akashi, 1995). These approaches are derived from the neutral theory of evolution; that is that the evolution of the genome is mostly caused by the accumulation of neutral mutations due to genetic drift and stochastic processes (Kimura, 1983). Using this expected distribution of neutral mutations, statistical tests and mathematical models have arisen to determine the degree of digression from neutrality (Nielsen, 2001). Molecular approaches for classifying and quantifying selection can be broadly grouped into two categories. Population genetics approaches are designed to look at mutations that segregate within a species, while phylogenetic approaches look for mutations that have become fixed in multiple species. Here, positive selection is explored in a phylogenetic context, that is, between species. Although selection in non-coding regions does happen it is beyond the scope of this thesis; unless specified otherwise, all further mentions of positive selection will therefore be with regard to protein-coding genes.

By performing genome-wide analyses, we can characterise factors that affect the rate of evolution, and understand how organisms evolve. This thesis is divided into two major areas, both of which investigate large-scale evolution. The first is the assembly and annotation of the

genomes of seven fungal species. The second is the use of this, and other publicly available genomic data, to investigate broad trends that affect the rate of evolution of a protein.

The main aim of this thesis is to investigate factors – primarily with relation to gene composition – that affect the rate of evolution in insects, mammals and fungi. In order to perform this analysis, genomes of the genus *Microbotryum* were assembled and annotated, and a bioinformatics genome annotation pipeline was developed.

1.1 UNDERSTANDING NATURAL SELECTION THROUGH GENETICS

Modern understanding of evolution is the culmination of centuries of work by numerous scientists; perhaps the most famous of which is Charles Darwin, who published the theory of evolution through natural selection. Following Darwin's theories, debates raged over the mode of natural selection. It wasn't until the rediscovery of work on inheritance performed by Gregor Mendel that genes were widely considered to be the answer. Mendel's experiments lead to the understanding that variation could persist in a population, and therefore confer selective advantages in different individuals through the medium of genes. Mutations were thought of as uncommon, stable, and stochastic events — continuously arising in populations. From these mutations beneficial types would persist, and deleterious mutations would eventually be removed from the population (for a review see Charlesworth and Charlesworth (2009)). These theories further developed the concept of 'fitness', whereby those mutations that increased an individual's fitness over that of the wild-type would confer selective advantage and spread through the population.

Based on this theory, selection can be divided into three categories; positive, purifying and balancing selection. Positive selection is the maintenance of a mutation that confers a selective advantage. Purifying selection is the selective removal of deleterious mutations from the population. Finally, balancing selection is the selective maintenance of multiple alleles at a given locus in a population. The notion that different types of selection are the predominant determinants of changes between species is known as the selectionist theory.

Since this time, the concept of stable fitness altering mutations has itself begun to evolve. In 1983, Motoo Kimura presented the theory of neutral evolution, whereby mutations were continuously arising but most resulting in only a small effect on fitness; the frequency of these variants could then fluctuate within a population with no large affect on fitness (Kimura, 1983). Using this theory, it became possible to attempt to quantify selection based on the digression from neutrality of a given mutation. This was at the forefront of a new understanding of evolution. Mutations began to be thought of as more frequent, stochastic processes, often having little to no effect. These mutations would either be lost from a given population due to chance, or spread through the population, also by chance; this spread of random mutations is known as genetic drift. However, the neutral theory did not sufficiently explain the rate of evolution in different species, as truly neutral mutations would lead to a rate of evolution that mirrored mutation rate. It was therefore suggested that if most mutations might be only slightly deleterious, then the fixation rate of these mutations (due to drift) depended on the effective population size and generation time, allowing the rate of molecular evolution to be constant with generation time and not entirely dependent on mutation rate (Ohta and Kimura, 1971). This theory has since been developed into the nearly neutral theory of molecular evolution, which allows for neutral, slightly deleterious and non-neutral, deleterious mutations. Both the nearly neutral and the selectionist theories are able to explain many aspects of evolution, and there is no consensus view to the most likely of the two.

We now know that the fixation of mutations is yet more complicated, and is affected by additional genetic components. When considering mutations that confer a selective advantage, it was initially predicted that only the fitness of the mutation affected the likelihood of fixation. Now however, studies have shown that phenotypic patterns do not always reflect genotypic patterns (Hadfield, et al., 2007). At the population level, a large contributor to this effect is that different alleles are not necessarily independent; a phenomenon known as linkage. Genotypic variation can therefore arise through large-scale processes driven by linkage, such as genetic hitchhiking and background selection. These processes can be detected through population genetics methods, and result in the segregation of mutations through means other than positive selection, purifying selection, balancing selection or genetic drift. In sexually-reproducing organisms, recombination plays a particularly important role in the distribution of genetic variation, as this is the major form of linkage. It is also though that recombination has an effect on mutation rate. For a review of these and other factors, see Hurst (2009).

The factors above represent genomic processes that shape variation between populations. Mutations that become fixed in this way have not been selected for or against, but are not necessarily neutral. Using population data one can demonstrate the movement of variation within a species, but cannot describe selective events with a broader signal, happening on a longer time scale. In considering genomes of different species, patterns begin to emerge that reveal synonymous mutations to in fact be under longer-term selective pressures, such as selection of translational efficiency (e.g. Stenøien and Stephan, (2005); Wang and Roossinick, (2006); Mukhopadhyay, et al., (2007)). Many correlations have also been detected where it is not clear

whether this is due to selection events or nearly neutral events, such as gene 'essentiality', intron-number and tissue-biased expression (Larracuente, et al., 2008).

To examine factors affecting rates of evolution, investigations into synonymous sites have become prominent. Polymorphisms in coding regions are classified into two categories; those that change an amino acid are known as non-synonymous, and those that do not change an amino acid are synonymous. Since synonymous substitutions do not change the amino acid, they are assumed to be invisible to selection, and are therefore assumed to evolve neutrally. Given the nearly neutral view of evolution, we would expect synonymous sites to be evolving neutrally as they do not change the amino acid. Contrary to this, investigations into codon (Duret, 2002) and nucleotide (Bernardi, 2005) composition have detected biased usage, and these are not the only gene compositional factors that are correlated with rates of site and protein evolution. Understanding whether phenomena such as these are due to nearly neutral factors (e.g. biased nucleotide usage due to biased DNA repair mechanisms), or selective events (e.g. nucleotide bias due to selection for chemical stability) will help us to understand the true role of natural selection in evolution. Identifying additional factors that affect the rate of evolution is also important to fully grasp the dynamics of fixation events between species.

1.2 FACTORS WITHIN A GENE THAT AFFECT THE RATE OF EVOLUTION

There are a number of known compositional factors that affect the rate of evolution of a gene. As previously discussed, the rate of synonymous substitution is known to be affected by nucleotide bias, which is the variation in nucleotide composition found both within and between species (Bernardi, 2005). Possible causes of nucleotide bias include biases in DNA repair, particularly

translational repair (Frank and Lobry, 1999), chemical stability (as the bases G and C contain more bonds and are therefore more energetically stable), and cost of the nucleotide (Rocha and Danchin, 2002). Within a species, the physical location within the genome and within the gene can also affect nucleotide bias (Wong, et al., 2002), as can the environment of the organism (Foerstner, et al., 2005; Singer and Hickey, 2003). Nucleotide bias has also been detected at non-synonymous sites, and can be traced up to the protein level (Singer and Hickey, 2000).

Organisms can also experience biased codon usage; where different codons coding for the same amino acid are not used equally (Duret, 2002). Codon usage can be extremely similar at close genetic distances (e.g. Vicario, et al., (2007)) but increases with phylogenetic distance. Within species, codon bias can vary between genes and tissues, and has been linked to mRNA and tRNA abundance (e.g. Chen, et al., (2004), Toshimichi, (1982)) gene expression (e.g. Duret and Mouchiroud (1999), Horn, (2008); Jia and Li, (2005)), gene length (e.g. Eyre-Walker, (1996); Duret and Mouchiroud, (1999); Chen, et al., (2004); Wang and Hickney, (2007)), the hydrophobicity of the gene (e.g. Zhong, et al., (2007)), amino acid composition of a gene (e.g. Chen, et al., 2004) and codon position within a gene (e.g. Gilchrist, (2007)). It has also been revealed that specific codons are favoured around exonic splice sites (Chamary and Hurst, 2005). There are two leading views on the cause of codon bias. Firstly, that it is caused by mutational bias/nucleotide bias in the codon region (Chen, et al., 2004), and secondly that it is caused by selection on transcription/translation (e.g. Chiusano, et al., (1999); Chen, et al., (2004); Rocha, (2004); Marquez, et al., (2005)).

Amino acid biases can also exist, and are the result of the preferential content of specific amino acids within the organisation of an organism (i.e. within a gene, gene system, network, whole

organism). The concentration of amino acids has been shown to correlate with environment, cellular localisation, physical properties, translational constraints/selection and the metabolic costs of the amino acids (e.g. Chiusano, et al., (1999); Bharanidharan, and Gautham, (2005); Kahali, et al., (2007)). It has also been suggested that selection would favour amino acid producing enzymes that contained minimal amounts of those amino acids that the enzyme produced (Alves and Savageau, 2005; Perlstein, et al., 2007).

With the recent increase in the volume of sequence data, it has become possible to examine many such factors affecting the rate of selection. Recent studies, for example (Larracuente, et al., 2008) have investigated and found correlations between the rate of evolution in a gene and the level of gene expression, the location of a gene in the genome, the length of a gene, the number of introns in a gene, the number of solvent exposed residues in a gene and many more features common to all genes. It is becoming increasingly clear that very few mutations within a gene have absolutely no effect. Given that the rate of evolution is variable and can be affected by many factors, it is reasonable to assume that the secondary structure of a protein might also affect the rate of evolution of a gene, due to the differing physiochemical properties of different structures. In mammals, proteins that are under strong structural constraints have been found to be under strong purifying selection and will tolerate less change in the amino acid sequence (Liu, et al., 2008). It has also been found that areas of the protein that are exposed to solvent, form looped regions, are natively disordered, of low complexity or in signalling regions of secreted proteins are more tolerant to mutations, i.e. are more polymorphic (Chiusano, et al., 1999). The causes and extent of these results are not known, nor is the effect in non-mammalian species (Tao and Dafu, 1998). Changes in hydropathy have been found to be structure specific, and in some cases have been strongly correlated with amino acid content (Das, et al., 2005); this increases the likelihood that secondary structure may play a role in the rate of protein evolution, as solvent accessible surfaces

are the most variable (Lin, et al., 2007). Overall, little is known about how positively selected sites might be correlated with protein secondary structure. From the above studies, there is strong reason to believe that positive selection might be affected by protein secondary structure and other structural aspects.

Using genomic data taken from the sequencing and assembly of the *Drosophila* 12 genomes project (Consortium, 2007), the alignments of genes from 6 mammalian species (Kosiol, et al., 2008) and the sequencing and assembly of the *Microbotryum* species performed in this thesis, factors, in particular the effect of secondary structure on the rate of evolution will be explored.

1.3 REFERENCES

Alves, R. and Savageau, M.A. (2005) Evidence of selection for low cognate amino acid bias in amino acid biosynthetic enzymes, *Molecular Microbiology*, **56**, 1017-1034.

Bernardi, G. (2005) Structural and evolutionary genomics — natural selection in genome evolution (review on nucleotide contents), *Elsevier Science*, 204:434.

Bharanidharan, D. and Gautham, N. (2005) Amino acid variation in cellular processes in 108 bacterial proteomes, *Archives of Microbiology*, **184**, 168-174.

Chamary, J.-V. and Hurst, L.D. (2005) Biased codon usage near intron-exon junctions: Selection on splicing enhancers, splice-site recognition or something else?, *Trends in Genetics*, **21**, 256-259.

Charlesworth, B. and Charlesworth, D. (2009) Darwin and genetics, *Genetics*, **183**, 757-766.

Chen, S.L., *et al.* (2004) Codon usage between genomes is constrained by genome-wide mutational processes, *Proceedings Of The National Academy Of Sciences Of the United States of America*, **101**, 3480-3485.

Chiusano, M.L., *et al.* (1999) Correlations of nucleotide substitution rates and base composition of mammalian coding sequences with protein structure, *Gene*, **238**, 23-31.

Consortium, D.G. (2007) Evolution of genes and genomes on the Drosophila phylogeny, *Nature*, **450**, 203-218.

Das, S., *et al.* (2005) Codon and amino acid usage in two major human pathogens of genus Bartonella — optimization between replicational-transcriptional selection, translational control and cost minimization, *DNA Research*, **12**, 91-102.

Duret, L. (2002) Evolution of synonymous codon usage in metazoans, *Current Opinion in Genetics and Development*, **12**, 640-649.

Duret, L. and Mouchiroud, D. (1999) Expression pattern and, surprisingly, gene length shape codon usage in Caenorhabditis, Drosophila, and Arabidopsis, *Proceedings of the National Academy of Sciences*, **96**, 4482-4487.

Eyre-Walker, A. (1996) Synonymous codon bias is related to gene length in Escherichia coli: Selection for translational accuracy?, *Molecular Biology and Evolution*, **13**, 864-872.

Foerstner, K.U., *et al.* (2005) Environments shape the nucleotide composition of genomes, *EMBO Rep*, **6**, 1208-1213.

Frank, A.C. and Lobry, J.R. (1999) Asymmetric substitution patterns: a review of possible underlying mutational or selective mechanisms, *Gene*, **238**, 65-77.

Gilchrist, M.A. (2007) Combining models of protein translation and population genetics to predict protein production rates from codon usage patterns, *Molecular Biology and Evolution*, **24**, 2362-2372.

Hadfield, J.D., *et al.* (2007) Testing the phenotypic gambit: Phenotypic, genetic and environmental correlations of colour, *Journal of Evolutionary Biology*, **20**, 549-557.

Horn, D. (2008) Codon usage suggests that translational selection has a major impact on protein expression in trypanosomatids, *BMC Genomics*, **9**, 2.

Hurst, L.D. (2009) Genetics and the understanding of selection, *Nat Rev Genet*, **10**, 83-93.

Jia, M. and Li, Y. (2005) The relationship among gene expression, folding free energy and codon usage bias in Escherichia coli, *FEBS Letters*, **579**, 5333-5337.

Kahali, B., Basak, S. and Ghosh, T.C. (2007) Reinvestigating the codon and amino acid usage of S. cerevisiae genome: A new insight from protein secondary structure analysis, *Biochemical and Biophysical Research Communications*, **354**, 693-699.

Kimura, M. (1983) The neutral theory of molecular evolution, *Cambridge University Press, Cambridge, MA.*

Kosiol, C., *et al.* (2008) Patterns of positive selection in six Mammalian genomes, *PLoS Genet*, **4**, e1000144.

Kreitman, M. and Akashi, H. (1995) Molecular evidence for natural selection, *Annual Review of Ecology and Systematics*, **26**, 403-422.

Larracuente, A.M., *et al.* (2008) Evolution of protein-coding genes in Drosophila, *Trends in genetics : TIG*, **24**, 114-123.

Lin, Y.-S., *et al.* (2007) Proportion of solvent-exposed amino acids in a protein and rate of protein evolution, *Molecular Biology and Evolution*, **24**, 1005-1011.

Liu, J., *et al.* (2008) Natural selection of protein structural and functional properties: a single nucleotide polymorphism perspective, *Genome Biology*, **9**, R69.

Mukhopadhyay, P., Basak, S. and Ghosh, T.C. (2007) Nature of selective constraints on synonymous codon usage of rice differs in GC-poor and GC-rich genes, *Gene*, **400**, 71-81.

Nielsen, R. (2001) Statistical tests of selective neutrality in the age of genomics, *Heredity*, **86**, 641-647.

Ohta, T. and Kimura, M. (1971) On the constancy of the evolutionary rate of cistrons, *Journal of Molecular Evolution*, **1**, 18-25.

Perlstein, E., *et al.* (2007) Evolutionarily conserved optimization of amino acid biosynthesis, *Journal of Molecular Evolution*, **65**, 186-196.

Rocha, E.P.C. (2004) Codon usage bias from tRNA's point of view: Redundancy, specialization, and efficient decoding for translation optimization, *Genome Research*, **14**, 2279-2286.

Rocha, E.P.C. and Danchin, A. (2002) Base composition bias might result from competition for metabolic resources, *Trends in Genetics*, **18**, 291-294.

Singer, G.A.C. and Hickey, D.A. (2000) Nucleotide Bias Causes a Genomewide Bias in the Amino Acid Composition of Proteins, *Molecular Biology and Evolution*, **17**, 1581-1588.

Singer, G.A.C. and Hickey, D.A. (2003) Thermophilic prokaryotes have characteristic patterns of codon usage, amino acid composition and nucleotide content, *Gene*, **317**, 39-47.

Stenøien, H. and Stephan, W. (2005) Global mRNA stability is not associated with levels of gene expression in Drosophila melanogaster but shows a negative correlation with codon bias, *Journal of Molecular Evolution*, **61**, 306-314.

Tao, X. and Dafu, D. (1998) The relationship between synonymous codon usage and protein structure, *FEBS Letters*, **434**, 93-96.

Toshimichi, I. (1982) Correlation between the abundance of yeast transfer RNAs and the occurrence of the respective codons in protein genes: Differences in synonymous codon choice patterns of yeast and Escherichia coli with reference to the abundance of isoaccepting transfer RNAs, *Journal of Molecular Biology*, **158**, 573-597.

Vicario, S., Moriyama, E. and Powell, J. (2007) Codon usage in twelve species of Drosophila, *BMC Evolutionary Biology*, **7**, 226.

Wang, H.-C. and Hickey, D. (2007) Rapid divergence of codon usage patterns within the rice genome, *BMC Evolutionary Biology*, **7**, S6.

Wang, L. and Roossinck, M. (2006) Comparative analysis of expressed sequences reveals a conserved pattern of optimal codon usage in plants, *Plant Molecular Biology*, **61**, 699-710.

Wong, G.K.-S., *et al.* (2002) Compositional gradients in Gramineae genes, *Genome Research*, **12**, 851-856.

Zhong, J., *et al.* (2007) Mutation pressure shapes codon usage in the GC-Rich genome of foot-and-mouth disease virus, *Virus genes*, **35**, 767-776.

2. ASSEMBLING THE GENOMES OF SEVEN SPECIES IN THE *M. VIOLACEUM* SPECIES COMPLEX

KATE E. RIDOUT

2.1 INTRODUCTION

Basidiomycete smut fungi are important study species for a number of reasons. One of the most well known is the maize pathogen *Ustilago maydis*, which has recently been sequenced leading to important discoveries about genes encoding aspects of pathogenicity (Kamper, et al., 2006). *Microbotryum violaceum* is a Basidiomycete fungal parasite that can cause anther smut in the plant family Caryophyllaceae (Thrall, et al., 1993). Although *Microbotryum* is a smut fungus, it is more closely related to yeasts such as Rhodotorula (see Chapter 4), and rust fungi; many of which, like *Ustialgo*, also parasitize essential food crops. *Microbotryum* has very little interaction with humans, and so has never been controlled by pesticides or other targeted campaigns. It is therefore an ideal species for studying natural host-pathogen co-evolution (Refregier, et al., 2008; Yockteng, et al., 2007), host-range shifts (Vercken, et al., 2010) and host-race differentiation, as hybrids become increasingly less viable with host species divergence (Le Gac, et al., 2007), as well as other aspects of fungal genetics (Hood, et al., 2005). The existence of the separate mating types provides the opportunity to examine the evolution of sex and mating types in a heterothallic fungal species (Devier, et al., 2009; Votintseva and Filatov, 2011), and features such as evolutionary strata (Votintseva and Filatov, 2009). Current research into *Microbotryum* has led to the sequencing of portions of the mating-linked region (Votintseva and Filatov, 2009), and the creation of EST libraries (Yockteng, et al., 2007).

The major portion of the life-cycle of *Microbotryum* is spent in the dikaryotic stage, and only a short amount of time is spent in the diploid phase and then haploid phase where mating occurs (Giraud, et al., 2008). In order to reproduce, the fungus has two mating types: A1 and A2, which co-exist in a single fungal cell as separate nuclei, this phenomenon is known as heterothallism. Mating between strains of the same type is incompatible, i.e. A1 cannot mate with A1. Though this system should tend towards outcrossing, the fungus preferentially mates with haploids

produced by the same dikaryon, which more closely resembles selfing (Giraud, et al., 2008). After mating, the sporidia undergo conjugation, resulting in the reinstating of the dikaryotic phase. During this time, *Microbotryum* produce hyphae, which will infect the host tissues. The fungus then overwinters in the meristematic tissues of the plant, until it is again ready to flower. Infection in this region ensures that only infected flowers will be produced; these flowers will then produce spores instead of pollen, leaving the host effectively sterilised and the parasite able to colonise other plants (Giraud, et al., 2008).

Comparisons between the genomes of *Microbotryum* and other species will provide valuable insights into many aspects of fungal evolution. To create further avenues of study concerning *Microbotryum*, multiple genomes of individuals from different host species have been sequenced. With these genomes, comparative analysis will help to better resolve the taxonomy of the fungus, perform evolutionary genetic analysis, identify gene differences and polymorphisms between *Microbotryum* adapted to different host species, determine species-specific adaptations and to gain further insights into the evolution of the mating-type loci. By sequencing whole genomes, large scale evolutionary analysis can be performed, and some, such as overall genome composition and gene content can only be accurately compared on a genomic scale. Sequencing projects that focus on whole genomes are able to identify novel structural and functional elements, protein coding genes, genomic duplications and losses, regulatory elements, mating strategies and more.

Here, we create a reference genome upon which other species of *Microbotryum* can be mapped on to. In order to do this, three *M. violaceum* var. *latifolia* individuals for each of the two mating types (A1 and A2) were sequenced, creating a total of six genomes of one species. The most

successful assembly proved to be an effective template upon which to map all other species genomes. The taxonomy of *Microbotryum violaceum* is currently under revision, as isolates from different plant hosts are sufficiently diverged to be considered different species (see Chapter 4), therefore a naming strategy is adopted with regard to the plant host. The genomes assembled in this chapter are from seven plant host species; *M. violaceum* var. *caroliniana* (from *Silene caroliniana*), *M. violaceum* var. *dioica* (from *Silene dioica*), *M. violaceum* var. *latifolia* (from *Silene latifolia*), *M. violaceum* var. *flos-cuculi* (from *Lychnis* (*Silene*) *flos-cuculi*), *M. violaceum* var. *vulgaris* (from *Silene vulgaris*), *M. violaceum* var. *saponaria* (from *Saponaria officinalis*) and *M. violaceum* var. *dianthus* (from *Dianthus carthusionorum*).

2.1.1 WHOLE GENOME SHOTGUN SEQUENCING

Shotgun sequencing gives rise to a complex computational problem. How are the small sequenced genomic fragments, or reads, to be put into the correct order? A single genome is sequenced many times so that multiple overlapping reads exist for a single sequence; the amount of over-sampling is referred to as the coverage of the genome. During the shotgun process the sequence is broken at random locations so that coverage will vary across the genome. The aim is to sequence a genome with high enough coverage to represent the maximum amount of accurate data at the lowest cost in the shortest time. As the genome is covered multiple times and the break points are random, reads will overlap each other, allowing us to use the overlaps to piece back together the original sequence (Sanger, et al., 1982). Though this may seem fairly straightforward, the problem is compounded by repetitive elements in the DNA, areas of low coverage, sequencing error, the lengths of the reads, polymorphic variants (in comparative assemblies or pooled assemblies) and the amount of sequencing data. This has created a need for sophisticated assembly algorithms that attempt to re-create the genome. Genome assembly can be a complicated task, with no single

right way to produce a finished genome. Different assembly software are likely to produce different quality output depending on the content of the genome sequenced (e.g. repeat and GC content), the quality and length of the reads, the use of paired end read information and the type of sequencing technology. The combination of multiple sequencing technologies and organism genomes will also affect the manner and result of the assembly process. Many different methods and software exist for assembling data from the various sequencing technologies, and assembly can be divided into two broad categories; assembling with a reference and without a reference.

2.1.2 ASSEMBLING WITH A REFERENCE

Where a reference genome is available Next Generation Sequencing (NGS) can be used to create comparative assemblies, usually increasing the assembly size relative to *de novo* assembly and massively reducing the assembly process, making this an effective tool for whole genome Single Nucleotide Polymorphism (SNP) analysis (Wang, et al., 2008). Contigs sequenced from the new genome are mapped onto the reference based on sequence homology; the more accurate and closely related the reference, the better the resultant assembly (Gnerre, et al., 2009). Alternatively, the reference can be used to guide the assembly and test sequence overlaps, rather than a direct mapping. Comparative re-sequencing (reference guided assembly) has become extremely popular; for example the human 1000 genomes project (www.1000genomes.org), and the *Arabidopsis* 1001 genomes project (www.1001genomes.org) are both using this strategy. Many mapping assembly projects are able to take advantage of increasingly cheap and accurate short reads, as longer reads are less crucial when there is a template involved (Hunt, et al., 2010; Tong, et al., 2010; Turner, et al., 2010). Unfortunately, due to repetitive elements and genome differences such as rearrangements, there will often be stretches of un-mapped reads that must be assembled *de novo*. Caution should be taken with mapping assemblies and reference guided

assemblies that rearrangements and novel genomic elements are not lost. Here, we have used the mapping module of commercial assembler CLC Bio Genomics Workbench (CLC Bio, Aarhus, Denmark), referred to from here onwards as CLC.

2.1.3 ASSEMBLING WITHOUT A REFERENCE

Assemblies without a template – *de novo* assemblies – will often incorporate longer reads, such as Sanger or 454 reads (e.g. Steuernagel, et al., (2009)); the increased read length facilitates a better assembly as longer reads are more likely to span short repeats and can use longer overlaps. Short read *de novo* assemblies are less effective on complex genomes, and are likely to become more viable as read lengths produced by different technologies increases over time (MacLean, et al., 2009). Scaffolding – the process of arranging, orientating and joining contigs based on additional sources of data – is a crucial step in creating reasonable *de novo* assemblies. Scaffolding is usually performed using sequence reads of a known distance apart, such as paired end or mate pair reads; this method not only reduces errors but also helps the resolution of misassembled regions, repetitive regions and areas of low coverage. Varying the distance between paired reads will also improve the overall assembly (Schatz, et al., 2010; Siegel, et al., 2000).

2.1.4 COMMON *DE NOVO* ASSEMBLY TECHNIQUES

Comparing all reads against all other reads is an NP hard problem (a class of complex algorithmic problems with no current efficient solution; therefore, assembly is carried out heuristically. Three main heuristic approaches are commonly taken by *de novo* assembly software developers,

depending on the type and lengths of the reads to be assembled. These methods are: the string based overlap approach, the overlap layout consensus (OLC) method and the *de Bruijn* method.

To reduce memory costs, assembly software usually employ an indexing step for more efficient sequence storage and access. A commonly used approach taken by all types of assembly software is to divide reads into smaller, uniform length sections called *k-mers*, which can be used for indexing. The '*k-mer* spectrum' represents the entire *k-mer* content of the genome. Individual *k-mers* can represent a single read, whereby the prefix and/or suffix of a read may be taken, or be a part of the entire spectrum, where each *k-mer* begins at the next base in a read and ends where there are *k*-1 bases left. Time demands are reduced if only reads with matching *k-mers* need to be compared, and storage can be made more efficient by only storing one instance of a given *k-mer*. This approach does not save time in repeat regions, and they are therefore often excluded from the process (Pop, 2009). The time taken for assembly is affected by the length of the *k-mers*, which should be large enough to avoid chance matches and small enough to reduce the effect of sequencing error and assembly time (Miller, et al., 2010). To further reduce the time and memory costs of sequence assembly, greedy heuristics can be used. The greedy approach chooses a starting read (or seed) – usually at random – and adds sequences one at a time to the growing contig, based on the highest scoring overlap. Greedy heuristics vastly reduce the time taken to compute overlaps, and simplify the overlap graph (discussed below) by considering only the highest scoring graph edges (Miller, et al., 2010). An example of a greedy string based assembler is QSRA, which computes overlaps allowing for imperfect matches and uses quality scores to detect read errors (Bryant, et al., 2009; Jeck, et al., 2007).

Both the OLC and the *de Bruijn* assembly methods are graph-based. Graph-based approaches rely on a computational graph made up of nodes and edges, where the edges usually represent sequence overlaps and the nodes represent the read or unique sequence. In the case of the *k-mer* spectrum, the edges might represent the single unique base and the nodes the overlap. A graph with only one possible direction between nodes is referred to as a directed graph, and a path through the graph where each node is visited only once is a simple path. Sequence assembly aims to find the longest simple path possible for the data provided, where a path represents a contig or a scaffold. An analysis of *de novo* assembly software is given by (Zhang, et al., 2011) and (Bao, et al., 2011).

The OLC method runs in three stages. First, all pair-wise overlaps are calculated in the data using a greedy step, and stored in an overlap graph. The region to be searched for an overlap is usually the same length as the *k-mer*, so that the graph node is the overlap and the edge is the remainder of the sequence. Secondly, an approximate read layout is calculated, identifying the possible paths through the genome (the graph). This is the most crucial step and begins by identifying unambiguous paths before using internal thresholds to choose the most effective contigs and path extensions. Due to the complexity and time required for overlap computation, OLC methods are often restricted to the longer and lower coverage 454 reads. An example of this is Roche's own assembly programme, Newbler, which was designed for 454 sequencing projects and functions both *de novo* and on mapping assemblies (Margulies, et al., 2005). The final step of the OLC method is to determine the consensus sequence of the chosen path. The goal of the OLC is to produce a single path that visits all nodes exactly once and uses all reads, called a Hamiltonian path (Pop, 2009). Software that use the OLC method include Celera Assembler (Denisov, et al., 2008; Myers, et al., 2000) and MIRA (Chevreux, et al., 1999).

Most new graph assembly methods are based on the *de Bruijn* graph, which was designed to represent all possible strings in a finite alphabet. Where the *k-mers* within an overlap graph represent the overlapping regions of the sequence, the *k-mers* in the *de Bruijn* graph make up the entire sequence, where each *k-mer* begins at the next base (Figure 2.1). Creating the graph in this way eliminates the need for expensive overlap calculations, as the contigs simply become a product of making the graph. A perfect outcome of this method would be to find a Eulerian path: a path through the graph which uses every edge in the graph only once. This approach was first developed by (Pevzner, et al., 2001), in the assembler EULER, and is also used by the popular short read assembler Velvet, which relies on the deep and even coverage of reads produced by short read assemblers (Zerbino and Birney, 2008). A downside to any *k-mer* based *g*raph approach is that it is less tolerant of sequencing error, due to shortened regions for comparison and searching for implicit overlaps, requiring sensitive correction methods and often using coverage to collapse weakly supported paths (caused by error-generated false overlaps; (Miller, et al., 2010). Shortened regions of comparison will also reduce the power to resolve repeats, as *k-mers* are less likely to be repeat spanning. Repeats, sequencing error and areas of low coverage can lead to assembly graphs becoming tangled, where false paths, bubbles in the graph, cycling, and convergence and divergence will mask the correct path through the data. Reducing the tangles in a graph is currently an NP hard problem, therefore local heuristics, approximations and scaffolding are employed to resolve these regions (Miller, et al., 2010).

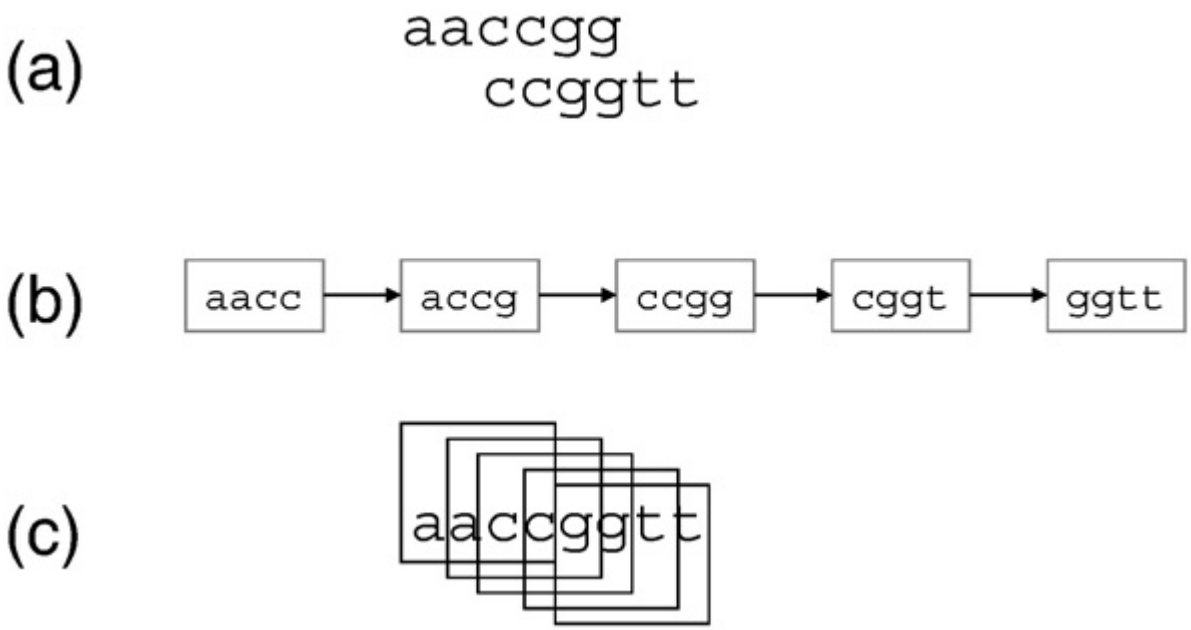

Figure 2.1. Sequence overlap represented by the *k-mer* spectrum from (Miller, et al., 2010). A) Two reads with a 4 bp exact overlap. B) The reads lain out as a *k-mer* graph, where the *k-mers* are 4 bp long. C) The contig consensus from the reads.

2.1.5 *DE NOVO* SCAFFOLDING

Scaffolding is an essential step in genome assembly. Contigs are usually scaffolded using paired end and known sequence data. Read pairs of a known distance apart and relative orientation can be mapped to or located in the assembly. Some of these reads will be 'broken', meaning that they are on separate contigs, and so the approximate distance between these contigs can be calculated and the contigs 'scaffolded' together, using 'N' characters to represent gaps. The orientation and distance between read pairs will vary based on the sequencing technology and method of read pair creation. Most scaffolders require at least one join to connect two contigs, thus accounting for sequence error. Scaffolding algorithms usually use greedy based heuristics, where the most

reliable joins are used first (Pop, 2009). Many programs contain their own scaffolding step using paired end data, including the 454 read assemblers Newbler (Margulies, et al., 2005) and Celera/CABOG (Denisov, et al., 2008; Miller, et al., 2008), and the short read assembler EULER-SR. The Eulerian assembler Velvet performs scaffolding using its *de Bruijn* graph, where paired reads determine the path through the graph. Stand-alone scaffolding programs take in assembly information and use paired end or mapping data to scaffold contigs; examples of this include Bambus (Pop, et al., 2004) and SSPACE (Boetzer, et al., 2010). Bambus connects contigs based on external links in addition to paired end data; the order in which paired end libraries and additional links are used can be specified by the user. Links can be detected from a range of additional data, including transcript information, protein homology, homologous genomes and mapping information. This linking data is created through sequence alignment of the known evidence using Nucmer, from the MUMMER package (Delcher, et al., 2002). Several scaffolding programs, including Celera and the stand-alone scaffolding program SSPACE will also incorporate overlaps between contigs and/or unassembled reads that were not sufficiently supported to assemble in the contig building stage.

2.1.6 CONTROLING ASSEMBLY QUALITY

Controlling the quality of an assembly without a reference sequence is an uncertain task. Assembly software can be benchmarked based on *de novo* assemblies of finished genomes, giving a measure of how well a specific data set could be assembled, but not necessarily all data. Fortunately, there are some guidelines; often, software designed for a specific sequencing technology should perform better on data created by that technology. Sequencing technologies are prone to platform biases; 454 technology often generates errors when determining the length of homopolymer regions, and can insert extra bases around these regions (Huse, et al., 2007).

Illumina biases tend to be substitution based, with different substitutions occurring at different rates, and a decrease in read quality towards the read ends. Illumina coverage of a genome is also biased, sequencing more GC rich regions (Dohm, et al., 2008). Sequence platform differences and biases are discussed by (Harismendy, et al., 2009). Due to the use of heuristics, even different assemblies of the same data using the same program may vary. Many assembly software and pipelines address assembly quality by first filtering reads using sequence quality information and correcting errors (DiGuistini, et al., 2009; Margulies, et al., 2005). Others, predominantly short read methods such as the *de Bruijn* based approaches, address errors through the graph by removing weakly supported or ambiguous paths (Simpson, et al., 2009; Zerbino and Birney, 2008). Coverage of the genome can also be used as a method of error correction by removing reads and graph paths with low sequence density. Unfortunately, coverage itself can vary with sequencing platform bias, cellular copy number variation and by chance (Miller, et al., 2010).

2.1.7 ASSESSING ASSEMBLY QUALITY

Measuring the quality of a final assembly is extremely challenging, as this cannot be explicitly determined by assembly software even when reference data is available. Without a reference genome it becomes even more difficult, and assembly quality must be estimated. Assuming that the correct assembly is one that uses the data to create the most complete contigs (i.e. fewest reads are left unassembled and contigs are long), an approximation of quality can be obtained through the number of contigs and scaffolds, the total base count, the average contig length, the longest contig and scaffold lengths, and the N_{50} value of the assembly. The N_{50} is the contig length at which 50% of the total bases in the assembly are represented within the longest contigs. Consequently, N_{50} values are not comparable between assemblies were the total length and shortest contig size are different. It is possible to estimate assembly quality from multiple genome

sequences of the same species or a close relative, transcriptome and known protein sequences (DiGuistini, et al., 2009), mapping data (e.g. via the optical scaffolding program SOMA (Nagarajan, et al., 2008)) and paired end data. Paired end information can be used to estimate the quality of an assembly using the number of joined contigs; if two contigs should be joined based on mate pair content, but the orientation and/or length constraints of the distances between the pairs do not allow it, this is likely evidence of a mis-assembly (DiGuistini, et al., 2009). In most *de novo* cases there will always be an unknown quantity of mis-assemblies and mis-called bases, and it is important to be aware of this fact when performing down-stream analysis.

In this chapter multiple genomes of *M. violaceum* var. *latifolia* are assembled in a number of different ways and the results compared to determine the best result. Assemblies are produced by combining different sequencing technologies, assembly software and genomes, and scaffolding using paired end and mate pair data of various insert sizes. Sequences from the remaining *Microbotryum* genomes were mapped to the 'best' *M. violaceum* var. *latifolia* assembly.

2.2 METHODS

2.2.1 DATA

Three types of paired read data have been used in the *Microbotryum* assemblies performed in this chapter. The first is the 454 paired end data (sometimes called mate pair). These reads are created through circularization, can be ~3 kb, ~8 kb or ~20 kb apart, and are oriented to face the same direction on the same strand (Ng, et al., 2006). The Illumina paired reads are either paired end or mate pair. Paired end read distances are up to ~600 bp including the length of the read, with both

reads facing inwards on opposite strands. Illumina mate pair reads can be up to 7 kb apart, and face in the same direction. Two different gap distances were used for the 454 sequencing as it is known that combining different gap distances of paired end data improves genome assembly (Schatz, et al., 2010).

Microbotryum sequences were generated by Roche/454 (Margulies, et al., 2005) in Liverpool, and by Solexa/Illumina (Bennett, 2004) in Hinxton, Cambridge. The data provided included six individuals from *M. violaceum* var. *latifolia*; three of mating type A1 and three of mating type A2. Two individuals of mating types A1 and A2 were sequenced using the 454 method including paired end data. The remaining four genomes; two with the A1 mating type and two with A2 were sequenced using the Illumina method; a single A2 genome contained paired end information. The additional six species genomes: *M. violaceum* var. *caroliniana*, *M. violaceum* var. *dioica*, *M. violaceum* var. *flos-cuculi*, *M. violaceum* var. *vulgaris*, *M. violaceum* var. *saponaria* and *M. violaceum* var. *dianthus* were all of the A2 mating type and sequenced using Illumina technology. All genomes contained paired end data.

Transcriptome data from *M. violaceum* var. *latifolia* was obtained via Illumina sequencing (Filatov, unpublished). The data comprised 9,131 unique Expressed Sequence Tag (EST) sequences.

2.2.2 ASSEMBLING THE GENOMES

Genome assembly was performed using both *de novo* and mapping approaches. The six *M. violaceum* var. *latifolia* genomes were the first to be sequenced. The closest sequenced relative at the time of assembly was *Ustilago maydis*, which was found to be around 20-30% divergent to *Microbotryum* when using highly conserved ITS data (Almaraz, et al., 2002); too divergent to be used for mapping. The genomes were assembled *de novo* using a variety of software. The six genomes of the different *Microbotryum* species were then mapped to the *M. violaceum* var. *latifolia* reference genome.

2.2.3 *M. VIOLACEUM* VAR. *LATIFOLIA* SINGLE GENOME ASSEMBLIES

The six genomes from *M. violaceum* var. *latifolia* were assembled separately using two different methods for each sequencing technology in order to find the best assembly method for the data. Repetitive regions of the genome were expected to be the key feature that might break an assembly; therefore programs were chosen that used different methods to handle repetitive stretches, as well as using a combination of assembly algorithms.

454 Genomes

The 454 reads were assembled using two OLC assembly programs; Newbler (Margulies, et al., 2005) and Celera (Denisov, et al., 2008). Newbler was chosen as it was designed by Roche for Roche data, and therefore should perform optimally. Celera assembler was created by Celera Genomics for the assembly of large eukaryotic genomes using Sanger data, and has since been extended to accept 454 reads (Miller, et al., 2008). Celera and its extension pipeline CABOG are

popular with large genomics institutions such as the J. Craig Venter Institute, and are frequently revised and upgraded. Each program contains an internal scaffolding module designed for 454-style read pairs and different methods of handling repetitive elements. Both Celera and Newbler were designed to handle 454 reads, and as such are robust to possible errors in homopolymer runs, common to 454 sequencing (Huse, et al., 2007).

Newbler (Roche WGS-assembler) was specifically designed for the 454 data, therefore filtering and masking was carried out by the software using data raw SFF (Standard Flowgram Format) data, thus incorporating base quality. The assembly was performed by Roche. In this method, the software attempts to scaffold across repeat regions using paired end information.

Celera assembler was designed to read in raw 454 sequencing data (SFF) and also performs filtering and masking of the data internally, whilst using sequence quality information to aid assembly. Celera handles repeats by collapsing single base runs into a single base, and estimates the likelihood that a region of the genome represents a collapsed repeat using a Poisson approximation based on coverage information. Scaffolding is performed internally using paired end data.

Illumina Genomes

Reads produced by Illumina sequencing were assembled using Velvet (Zerbino and Birney, 2008) and QSRA (Bryant, et al., 2009), which are both designed for short read data. Velvet was created at the European Bioinformatics Institute (EMBL-EBI) and was specifically designed for read data such as Illumina. QSRA is a string based assembler, and was chosen for its ability to include

quality information. The Illumina sequencing method does not include the sequencing of vectors meaning that data did not have to be masked. However, before assembly the four Illumina data sets were filtered for reads containing bases that could not be called, represented as "N"s; these reads were not considered to be sufficiently reliable to include in the assembly.

Velvet is a graph based assembler that creates a *de Bruijn* graph from the *k-mer* spectrum of the data and then searches for a Eularian path through that graph. Repeats are automatically collapsed as a product of *k-mer* indexing. A Poisson approximation is used to estimate the likelihood of a repetitive region based on the *k-mer* frequency using the bespoke algorithm 'Rock Band'. Where paired end data are available Velvet will use a second bespoke algorithm 'Pebble' to construct scaffolds across the *de Bruijn* graph, thereby resolving repeats (Zerbino, et al., 2009). Genomes were assembled with various *k-mer* lengths and the length producing the highest N_{50} value was chosen. The expected coverage of the reads and the coverage cutoff (below which coverage is not sufficiently high to be reliable) were estimated using the Velvet optimiser script distributed with Velvet.

QSRA is a greedy string based overlap assembler and, like Velvet, collapses repeats as a product of indexing when creating a hash table of *k-mers*. This index is used to detect overlaps and assemble reads in a greedy fashion, until the assembly process is interrupted where ambiguities occur. QSRA does not have a scaffolding module; therefore performance is not expected to be as good as Velvet where paired read data is available. Unlike Velvet, QSRA is able to use the quality information generated by the Illumina Genome Analyzer during sequencing, which might benefit the assembly by resolving ambiguities. QSRA was run with default parameters. Results from QSRA are not provided in assembly format, therefore to recover paired end information

contigs were scaffolded using SSPACE (Boetzer, et al., 2010), which maps paired reads onto contigs.

Scaffolding

The transcriptome sequences from *Microbotryum violaceum* were used to provide additional contig linking information for scaffolding. Both the 454 and Illumina genomes were scaffolded using this additional information plus paired reads where available using the stand-alone scaffolder Bambus (Pop, et al., 2004). Bambus considers linking information in the order specified by the user, therefore transcriptome links were considered to be the highest priority information. Two or more read pairs were required to join contigs, but only a single transcriptome, as only single copies of each gene were present in the transcriptome data set.

2.2.4 *MICROBOTRYUM VIOLACEUM* HYBRID GENOME ASSEMBLIES

It is possible that combining the *M. violaceum* var. *latifolia* genomes into a single consensus assembly for each mating type might increase the size and accuracy of the assemblies and create high coverage consensus assemblies. To test this, hybrid assemblies for the A1 and A2 mating types were produced; each consensus genome comprised two Illumina genomes and one sequenced with 454 technology. Two approaches were taken: the first was to simply use hybrid assembly software that could combine raw 454 and Illumina data. The second approach was to assemble either the 454 or the Illumina genomes, and to pipe the resulting contigs/scaffolds into an assembly with the remaining raw data. Illumina data sets were filtered for reads containing bases that could not be called, represented as "N"s. The exceptions were CLC and Celera, which contain their own quality control methods.

Single program assemblies

At the time of assembly, five popular hybrid assemblers were able to combine reads from different sequencing technologies. These were Celera, Velvet, CLC, EULER-SR and Ray.

Celera assembler is able to perform hybrid assemblies of 454 and Illumina raw sequence reads, although reads shorter than 75 bp are not officially supported. Sequences were directly imported into Celera using scripts provided with the software, this allowed information on quality and vector regions to be passed directly to the program. Both the Illumina and 454 paired end data were used for internal scaffolding.

Velvet is able to take in long sequences such as 454 data, but is not designed explicitly for 454 reads. Illumina reads were filtered for low complexity (for example homopolymer runs) using the *filterIlluminaReads* script from the EULER-SR program to mitigate the effects of decreasing the coverage cutoff (to allow for the reduced coverage of the 454 data). 454 reads were filtered by quality and paired data was split at the linker using the *qualityTrimmer* and *splitLinkedClones* scripts, also from EULER-SR. Illumina reads are more accurate and are not sequenced with a linker, therefore this step was not previously necessary. The orientation of 454 mate pair reads is different to that of Illumina paired end reads; therefore the forwards read of every 454 mate pair was reverse complemented, to mimic the Illumina style. Velvet was able to use Illumina and 454 paired read information to scaffold the contigs. A *k-mer* value of 31 was taken, as it resulted in the best single genome assembly.

CLC Bio Genomics Workbench (CLC bio, Aarhus, Denmark) is designed for both 454 and Illumina raw sequence data and was used with default parameters. CLC functions using a *de Bruijn* method, but does not contain a scaffolding module. The stand-alone scaffolder Bambus was used to scaffold data using both 454 and Illumina paired end information.

EULER-SR (Pevzner, et al., 2001) is a short read extension of EULER; the first *de Bruijn* graph assembler. As with Velvet, Illumina data was filtered for low complexity regions, and 454 data were filtered for low quality bases and split at the linker sequence. EULER-SR is able to create scaffolds from both the 454 and the Illumina paired end data, however scaffolding in this case was extremely poor, therefore SSPACE was used to scaffold and extend the contigs using default values. The single EULER-SR assembly containing Illumina paired end data could not be assembled; therefore the single genome assembly of the Illumina paired end data using Velvet was included in addition to the raw reads to assist the assembler. The highest *k-mer* value available (25) was used for the assembly.

Ray (Boisvert, et al., 2010) is a *de Bruijn* assembler which uses seeds rather than Eulerian walks to determine the path through an assembly. Illumina data were filtered to remove low complexity regions and 454 data was filtered for low quality bases and paired ends split at the linker. Ray was able to scaffold contigs using paired end information from the 454 data only. Ray was scaffolded using the highest *k-mer* value at the time of assembly (35 in version 1.6.0).

Combining prediction programs

A number of programs are able to include pre-assembled data into a raw sequence assembly. The programs Newbler, Velvet, Celera and MIRA were chosen. Newbler, Velvet and Celera are some of the most popular and well established assembly programs available, and were shown in the single genome assemblies to perform well on the data (see Results). MIRA was chosen as a different type of assembly method, and was used in the hybrid assemblies as it was unable to assemble the raw Illumina sequences.

A Newbler/Velvet hybrid assembly was performed, following a method similar to (Nowrousian, et al., 2010): raw 454 data were assembled using Newbler, which automatically masks and filters the 454 reads. The assembled 454 contigs and scaffolds were then added to the raw Illumina reads and the original raw 454 reads and assembled using Velvet. Raw 454 reads were filtered for low quality regions and any paired sequences were split at the linker before being used by Velvet. Raw Illumina sequences were filtered for reads containing "N" characters and low complexity regions. Newbler was chosen over Celera for use on the 454 data because the single genome assemblies revealed that a higher proportion of reads were included into the Newbler assemblies (Table 2.3). A *k-mer* value of 31 was used for the assembly, as this was the best value for the Illumina data. 454 data was scaffolded using Newbler before assembly with Velvet, and Velvet parameters were altered to allow for long contigs and scaffolds with a copy number of one. The raw 454 data and Illumina reads were scaffolded using Velvet. As before, the forward facing read pairs of the 454 data were reverse complemented.

Illumina data was assembled using Velvet; as with the single genome assemblies, the Illumina data was filtered for reads containing bases that could not be called and low-complexity regions,

and the *k-mer* value of 31 was used. The Velvet contigs were assembled to the raw 454 data using Celera assembler. Celera was not able to cope with scaffolded data as input, therefore scaffolds and long contigs were shredded, maintaining overlaps for re-assembly. Scaffolding of the Illumina data was carried out by Velvet, and the 454 data was scaffolded using Celera.

Velvet contigs and scaffolds from the single genome assemblies were assembled to the raw 454 data using MIRA (Chevreux, et al., 1999). SFF (Standard Flowgram Format) output from the 454 sequencing was run through the program *sff_extract* from the AMOS package which enabled the retention of information on base qualities and paired end regions (including linker regions), which could be passed to MIRA, therefore 454 data was not filtered or masked. At the time of assembly, MIRA did not contain a scaffolding module, and the assembly was scaffolded using Bambus with the 454 data. The Illumina sequences were already scaffolded upon inclusion into MIRA.

Hybrid assembly scaffolds

Due to the complicated assembly processes used in the hybrid genomes it was not always possible to create data in the correct assembly format for Bambus. Because all assemblies were scaffolded using paired end data, and the Bambus untangling process often reduced contig sizes, it was decided not to use Bambus to assemble these genomes. This also meant that transcript joining information is not available to assess the scaffolds.

2.2.5 EST MAPPING

The transcriptome sequences of *M. violaceum* var. *latifolia* were mapped onto all single and hybrid scaffolded assemblies using GMAP (Wu and Watanabe, 2005) as a measure of assembly quality. Three measures were taken: the number of complete transcript alignments, the number of transcripts broken across multiple scaffolds or duplicated regions and the number of transcripts missing from the genome assembly.

2.2.6 ADDITIONAL SCAFFOLDING

The final approach taken for genome assembly was to iteratively scaffold a single genome using paired end information from multiple genomes. The A2 Illumina sequenced genome containing paired end data (563a2) was chosen for this purpose as it contained the longest contigs. The 454 paired end data from the single A2 and A1 454 genomes of *M. violaceum* var. *latifolia* was mapped onto genome 563a2 using the mapping module of CLC. The mapped reads were then used to scaffold this genome (for the second time) using Bambus. Additional mate pair sequences were obtained at a later date from Illumina with a paired end distance of up to 7 kb for both the A1 and A2 genomes of *M. violaceum* var. *latifolia.* As with the 454 data, the new mate pair information was mapped onto the scaffolded 563a2 genome and the genome was scaffolded with Bambus for the third time.

2.2.7 MAPPING ASSEMBLIES

The additional six genomes of different *Microbotryum* species (all A2 mating type) were sequenced by Solexa/Illumina. All genomes were assembled separately using a mapping approach. Sequences were mapped using CLC Genomics Workbench onto the A2 reference genome from *M. violaceum* var. *latifolia*. This reference genome was taken from a single individual sequenced using Illumina paired end sequencing and assembled using Velvet (see Discussion). A single genome was chosen over the hybrid assemblies to avoid duplications of genomic segments generated by heterozygous regions. The A2 genome was chosen predominantly because it had the highest N_{50} value and most complete EST matches.

2.3 RESULTS

2.3.1 RAW DATA

In total, seven different *M. violaceum* species from different host plant species including three A1 genomes and nine A2 genomes were sequenced. Of the six *M. violaceum* var. *latifolia* strains, the four Illumina sequenced genomes had consistent read lengths of 51 bases with around 40× coverage. The 454 sequenced genomes contained reads of around 300 bases in length, and coverage of 11× and 13×. The GC content of the Illumina data was consistently higher than the 454 data, ranging from 51.95-53.37% compared to 48.68% and 49.25% in the 454 data. The single A2 Illumina sequenced genome containing paired end data was created using a gap distance with a maximum of 600 bp. The two 454 sequenced genomes also contained paired end data, with average gap distances of around 3,000 bp and 7,500-8,500 bp. Detailed results of sequencing are provided in Table 2.1. The remaining six species were sequenced with Illumina

technology using the same paired end distances as *M. violaceum* var. *latifolia* and a slightly increased read length of 91 bp.

Table 2.1. Summary of the raw sequencing output for *M. violaceum* var. *latifiolia*.

Individual Genome: Host Species (ID)	Location Collected	Mating Type	Technology	Paired End Data	Number Of Reads	Average Read Length	Total Bases	Coverage (25Mb genome)	GC% Content
Silene latifolia (1075a1)	Sardinia	A1	Illumina	No	19,686,467	51	1,004,009,817	40x	51.95%
Silene latifolia (562a1)	Poland	A1	Illumina	No	19,398,252	51	989,310,852	39x	52.00%
Silene latifolia (1076a2)	Sardinia	A2	Illumina	No	18,648,368	51	951,066,768	38x	51.97%
Silene latifolia (563a2)	Poland	A2	Illumina	Yes	21,892,872(x2)	51(x2)	1,116,536,472 (x2)	44x(x2)	52.53% 53.37%
Silene latifolia	Sweden	A1	454	Yes	920,205	320	294,711,198	11x	48.68%
Silene latifolia	Pyrennes	A2	454	Yes	1,007,004	339	342,362,706	13x	49.25%

Results of the filtered Illumina sequences are presented in Table 2.2; over 98.9% of all single end sequences remained after filtering, and over 99.7% of the forward paired end data and 98.7% of the reverse.

Table 2.2. Illumina sequence data before and after filtering.

Species ID	Number of Reads Before Filtering	Total Bases Before Filtering	Number After Filtering	Basecount After Filtering	% Reads Remaining
1075a1	19,686,467	1,004,009,817	19,487,045	993,839,295	98.987%
562a1	19,398,252	989,310,852	19,199,654	979,182,354	98.976%
1076a2	18,648,368	951,066,768	18,456,980	941,305,980	98.974%
563a2 Forwards	21,892,872	1,116,536,472	21,839,254	1,113,801,954	99.755%
563a2 Reverse	21,892,872	1,116,536,472	21,615,718	1,102,401,618	98.734%

2.3.2 *MICROBOTRYUM VIOLACEUM* SINGLE GENOME ASSEMBLIES

The GC contents of the 454 assemblies are slightly lower than the Illumina; ~54% compared to ~55% respectively, but far higher than the overall GC content of the raw 454 data, which was 48-49% and slightly higher than Illumina's 51-53%.

The 454 genomes were assembled using unfiltered data, and paired reads were available for all assemblies. Results of the contigs are provided in Table 2.3 and scaffolds in Table 2.4. Newbler assemblies (performed by Roche) contained 11,000 - 12,000 contigs; a similar number to the Velvet Illumina assemblies. The genome size assembled was 24-25 mb, which is the expected genome size of *Microbotryum.* The longest contigs and N_{50} values were comparable to the Velvet

Table 2.3. Contig assemblies of the 454 and Illumina genomes using Newbler, Celera, Velvet and QSRA. Velvet and QSRA input were filtered by removing reads containing N characters.

Genome (P = Paired)	Assembly Program	Bases in raw sequence data	Sequenced bases assembled (%)	Number of Contigs	Total Bases in Contigs	Mean Contig Length	Shortest Contig	Longest Contig	N50	Contig %GC content
454 A1 P	**Newbler**	294,711,198	96.34	12,731	24,422,291	1,918	100	46,184	4,319	54.42
	Celera	294,711,198	84.77	6,036	24,911,784	4,127	129	36,179	6,214	54.50
454 A2 P	**Newbler**	342,362,706	96.91	11,768	25,011,795	2,125	100	34,621	5,455	54.48
	Celera	342,362,706	86.75	5,360	24,274,282	4,528	192	49,739	7,463	54.40
Illumina 1075a1	**Velvet**	993,839,295	-	15,367	21,972,639	1,429	100	25,765	4,082	55.59
	QSRA	993,839,295	-	44,551	24,193,686	543	100	13,047	1,475	55.47
Illumina 1076a2	**Velvet**	941,305,980	-	14,251	21,944,497	1,539	100	22,680	4,645	55.60
	QSRA	941,305,980	-	42,355	24,074,380	568	100	13,347	1,662	55.49
Illumina 562a1	**Velvet**	979,182,354	-	14,707	22,041,422	1,498	100	25,384	4,390	55.60
	QSRA	979,182,354	-	42,955	24,125,913	561	100	11,229	1,593	55.46
Illumina 563a2 P	**Velvet**	2,216,203,572	-	11,224	21,975,666	1,957	100	59,933	7,405	55.65
	QSRA	2,216,203,572	-	101,644	27,860,946	274	100	3,726	384	55.83

assemblies. For both the A1 and A2 genomes, Newbler was able to create ~3000 scaffolds with N_{50} values of 7 kb and 8 kb, which is short in comparison to the Velvet paired end assembly and the Celera assemblies. The longest contig was 607 kp in the A1 genome and 165 kb in the A2 genome. Celera produced larger N_{50} values than Newbler (6-7 kb rather than 4-5 kb) and a smaller number of contigs overall (5,000-6,000 compared to 11,000-12,000 for Newbler). The number of bases assembled was similar to Newbler; 24 mb, as were the long contigs; 36 kb and 49 kb. A slightly larger percentage of the total read data is included into the Newbler assemblies (~96%) compared to the Celera assemblies (84% and 86%). The Celera scaffolding process resulted in a larger number of scaffolds than the Newbler assemblies, and hence a more fragmented assembly. N_{50} values were increased to 16 kb and 10 kb, and the longest contig was 510 kb in the A1 genome (slightly shorter than the Newbler assembly) and 182 kb in the A2 genome (slightly longer than the Newbler assembly).

Filtered Illumina data (Table 2.2) was used for the following assemblies (contig results of the Illumina assemblies are presented in Table 2.3 and scaffolds in Table 2.4). The *de Bruijn* assembler Velvet performed well on the data. Numbers of contigs ranged from 11,000, where paired-end data was available, to 15,000. The genome size was consistently considered to be 21-22 mb. The paired end data set had almost double the number of reads and double the single end N_{50} value (7 kb compared to 4 kb), with a longest contig length of 59 kb. Where Velvet scaffolding was performed, the N_{50} was increased to 33 kb, and the longest scaffold to 196 kb. The simple string assembler QSRA performed considerably less well than Velvet. Final contig counts ranged from 42,000 to 101,000 contigs in the paired end data set. The N_{50} values were found to be around 1 kb, and 384 bp for the paired end sequences with longest contigs from 3 kb (paired end) to 13 kb. SSPACE scaffolding increased the N_{50} to 3 kb and the longest scaffold to 32 kb, a far poorer result than the Velvet scaffolding. Scaffolds contained a greater number of bases (~30 mb) than the expected content of the *Microbotryum* genome (~25 mb).

Table 2.4. Scaffolds produced from contig assemblies. * Minimum contig length of the 454 genomes is 2000bp, and Illumina 100bp.

Genome (P = Paired)	Assembly Program	Scaffolding Program	Number of Paired Joins	Number of Transcript Joins	Number of Scaffolds	Bases in Scaffolds	Shortest Scaffold	Longest Scaffold	N50*	Number of Incorrect/Unchecked Transcript Joins
454 A1 P	**Newbler**	**Newbler**	101,671	0	3,313	19,596,117	2,001	607,788	7,171	7
		Bambus	32,152	2,458	9,643	24,608,510	100	136,027	9,884	
	Celera	**Celera**	53,550	0	4,849	26,030,595	548	510,547	16,608	0
		Bambus	12,779	828	6,349	26,664,926	800	1,297,964	11,118	
454 A2 P	**Newbler**	**Newbler**	72,108	0	3,158	20,897,573	2,003	165,618	8,176	122
		Bambus	23,066	1,191	5,794	24,039,930	500	75,666	10,108	
	Celera	**Celera**	39,394	0	5,150	25,434,786	279	182,071	10,615	23
		Bambus	7,154	727	6,935	27,243,369	800	68,912	11,396	
Illumina 1075a1	**Velvet**	**Bambus**	0	6,116	8,647	14,761,024	100	23,832	3,843	0
	QSRA	**Bambus**	0	18,643	18,472	18,396,230	100	28,658	2,069	0
Illumina 1076a2	**Velvet**	**Bambus**	0	4,876	7,284	13,566,947	100	23,298	4,270	0
	QSRA	**Bambus**	0	16,453	16,921	17,826,452	100	28,276	2,208	0
Illumina 562a1	**Velvet**	**Bambus**	0	5,646	8,504	14,719,081	100	25,592	4,013	0
	QSRA	**Bambus**	0	18,249	18,681	18,703,666	100	20,678	2,128	0
Illumina 563a2 P	**Velvet**	**Velvet**	-	0	20,557	24,248,496	21	196,899	33,861	-
		SSPACE	1,644,377	0	16,490	24,754,868	61	198,704	37,373	
	QSRA	**Bambus**	0	44,423	31,533	19,327,907	100	17,255	973	0
		SSPACE	5,841,330	0	50,669	30,184,866	100	32,587	3,647	

Additional scaffolding of single genome assemblies

The results of the Bambus scaffolds are presented in Table 2.4. Additional scaffolding with Bambus increased the N_{50} value of the 454 assemblies by ~1-2 kb in all but the Celera A1 assembly, and considerably decreased the longest scaffold length. The one exception was once again the A1 Celera assembly, where the longest scaffold assembled by Bambus was greater than that produced by Celera (1 mb compared to 500 kb). Bambus increased the total base count of all the 454 assemblies when creating scaffolds, something that is expected given that scaffolds are joined by "N" characters.

All single-end Illumina assemblies contained fewer total bases after scaffolding with Bambus, probably due to the lack of joining information resulting in discarded contigs. Running Bambus on the Velvet assembled Illumina single end data reduced the number of contigs/scaffolds but made little difference to the longest contig length. The N_{50} value was consistently decreased in all single end Velvet runs after Bambus scaffolding; however the large differences in total base count mean that a comparison of the N_{50} values is not reliable in this instance. Using Bambus with the same single end data assembled using QSRA, almost doubles the longest scaffold length, and the N_{50} value, in comparison to the QSRA contigs. A greater number of transcript joins are used by Bambus for single end scaffolding of the QSRA assemblies than in the Velvet assemblies, almost certainly because there were more gaps to close in the QSRA assembly. All single-end Illumina QSRA assemblies produce lower N_{50} values than the Velvet assemblies; however, two genomes (1075a1 and 1076a2) contain larger longest scaffolds than their Velvet equivalents. No genomes contain false transcript joins. The paired end Illumina genome assembled with QSRA was also scaffolded using Bambus (treated as a single end assembly), resulting in the smallest N_{50} and shortest long scaffold length of the paired Illumina assemblies.

Scaffolding the Illumina paired end assemblies with SSPACE (which does not use transcript data) improved upon Velvet's own scaffolding module, slightly increasing the N_{50} value, the longest scaffold length and the total base count, while decreasing the total contig number. SSPACE scaffolding of the QSRA contigs produced a large improvement to the assembly, but did not come close in size to the Velvet SSPACE assembly in terms of N_{50} and longest contig length (Table 2.4).

Assessing the scaffolds of single genome assemblies

Wherever possible, the quality of the scaffolds assembled was assessed using the number of failed transcript contig joins determined using Bambus; i.e. the number of joins suggested by transcript data that could not be resolved using the current assembly (Table 2.4). An additional measure of the number of complete EST matches through GMAP mapping to the assemblies was also taken. The results of the EST mappings are presented in Table 2.5. Out of the 454 assemblies, Celera consistently produced fewer incorrect transcript joins than Newbler. The Illumina assemblies using both QSRA and Velvet produced zero incorrect joins.

Table 2.5. ESTs mapped to the genomes using GMAP.

Genome (P = Paired)	Assembly Program	Scaffolding Program	Number of Complete EST Alignments	Number of Absent ESTs	Number of ESTs Broken Across Scaffolds / Duplicated
454 A1 P	**Newbler**	**Newbler**	4,959	2,371	1,801
		Bambus	5,522	148	3,461
	Celera	**Celera**	5,184	484	3,463
		Bambus	5,400	316	3,415
454 A2 P	**Newbler**	**Newbler**	5,561	1,573	1,997
		Bambus	5,310	252	3,569
	Celera	**Celera**	5,029	204	3,898
		Bambus	4,877	135	4,119
Illumina 1075a1	**Velvet**	**Bambus**	4,859	2,862	1,410
	QSRA	**Bambus**	5,002	1,371	2,758
Illumina 1076a2	**Velvet**	**Bambus**	4,517	3,465	1,149
	QSRA	**Bambus**	5,023	1,722	2,386
Illumina 562a1	**Velvet**	**Bambus**	5,020	2,660	1,451
	QSRA	**Bambus**	5,067	1,170	2,894
Illumina 563a2 P	**Velvet**	**Velvet**	6,635	201	2,295
		SSPACE	6,667	165	2,299
	QSRA	**Bambus**	5,182	1,566	2,383
		SSPACE	4,940	226	3,965

EST mapping revealed that both the (A1 and A2) 454 Celera assemblies were missing fewer ESTs, but contained a greater number of broken or duplicated ESTs than the Newbler assemblies. The A1 genome matched a greater number of complete ESTs in the Celera assembly than the Newbler assembly, but fewer in the A2 assembly. For both programs, scaffolding with Bambus consistently increased the number of ESTs that could be mapped onto the data. In the Newbler assemblies, this resulted in a large increase in the number of duplicated and broken EST alignments from 1,801 to 3,461 in the A1 genome and from 1,997 to 3,569 in the A2 genome. The number of complete EST alignments was increased in the A1 Newbler assembly, but decreased in the A2. In the Celera assemblies, the number of broken/duplicated EST alignments

did not differ a great deal when scaffolded with Bambus. The number of complete EST alignments was increased in the A1 assembly but decreased in the A2.

The single end QSRA assembles consistently contained a greater number of complete EST alignments and fewer absent ESTs, as well as a greater number of duplicated or broken genes than the Velvet assembly. This result is likely to be due to the greater base count in all QSRA assemblies. For the paired end assemblies, the QSRA scaffolds contained fewer complete EST alignments than the Velvet assembly scaffolds, with a greater number of missing ESTs. Scaffolding the QSRA assembly with Bambus (compared to SSPACE) increased the number of complete EST matches, but also increased the number of absent ESTs, which is consistent with the reduction of genome size in the Bambus assembly. Using SSPACE on the Velvet assemblies increased the number of total and complete EST alignments, and increased the number of duplicated and broken EST alignments by only four genes.

Comparing the single genome assemblies

Despite the 454 data containing a far smaller amount of read information, the increased read length led to higher N_{50} values and longer contigs than most of the Illumina assemblies, while creating genomes with a comparable or greater base content to the Illumina genomes. However, when considering the EST mapping and transcript joins it would appear that the Illumina paired end assembly contains fewer false joins and more complete EST matches. The genomes assembled using Velvet contain fewer bases than the 454 genomes, possibly due to stringent pruning of the *de Bruijn* graph, and the genomes assembled with QSRA contain similar numbers of bases to the 454 assembly. The maximum scaffold length produced by Velvet does not exceed that of the longest A1 454 genome, however the N_{50} value is far higher in the A2 paired Illumina

Velvet assembly than any of the 454 assemblies. The shorter maximum scaffold length is likely to be due to the shorter paired end distance and read length used in sequencing, as Velvet might not be able to resolve very long repeats. The number of broken or duplicated ESTs was slightly higher in the Velvet paired end assembly than in the Newbler assembly, but less than the Celera assembly. Consistent with this, the Velvet assembly comprised of a greater total number of bases than the Newbler assembly, but less than the Celera assembly. Overall, the Velvet paired-end assembly was the best *Microbotryum* single genome assembly achieved using the programs selected.

2.3.3 *MICROBOTRYUM VIOLACEUM* HYBRID GENOME ASSEMBLIES

Assembling the genomes together had one major drawback which is illustrated in the MIRA results, which due to a high sensitivity to errors and SNPs resulted in a chimeric assembly with a too high base count. Assembling several genomes together *de novo*, rather than iteratively mapping or scaffolding them to each other also causes genuine SNPs to break the assembly. The contigs produced from the hybrid assemblies are provided in Tables 2.6A and 2.6B and scaffolds in Tables 2.7A and 2.7B. Data was filtered wherever possible to reduce the effects of SNPs and sequencing errors, as discussed in the Methods. *De Bruijn* graph methods are expected to be the most disrupted, as the *k-mer* spectrum indexing will pick up all alternative graph paths, and programs with aggressive pruning algorithms like Velvet may discard substantial data. Performance of the different assembly methods was extremely varied, ranging from longest contigs shorter than the single assemblies to double those same assemblies. Assemblies that were true hybrids and assemblies that had been pre-assembled were also varied, with no clear best approach.

Table 2.6. A) Contigs from the combined genome assemblies for the A1. B) Combined genome assemblies for the A2 mating type.

A. Mating type A1

Assembly Method	Total reads after filtering		Number of Contigs	Total Bases in Contigs	Smallest Contig	Longest Contig	N50 (with minimum contig length 1000)	%GC
	454	Illumina						
Newbler → Velvet	896,113	38,684,880	15,038	25,710,787	100	107,484	11,414	53.77
Velvet → Celera	920,205	38,686,699	3,767	25,867,252	1000	107,253	17,525	54.75
Velvet → MIRA	920,205	38,686,699	15,770	35,387,049	25	52,564	8,565	53.35
Celera	920,205	39,084,719	3,822	25,771,243	1000	96,464	17,859	54,76
Velvet	896,113	38,684,880	10,831	22,397,797	100	43,915	9,471	55.54
CLC	920,205	39,084,719	10,274	24,653,653	34	67,319	14,269	54.72
Ray	896,113	38,684,880	20,118	23,046,565	100	19,451	4,306	54.93
EULER-SR	896,113	38,684,880	39,724	25,159,388	26	57,979	10,046	54.64

B. Mating type A2

Assembly Method	Total reads after filtering		Number of Contigs	Total Bases in Contigs	Smallest Contig	Longest Contig	N50 (with minimum contig length 1000)	%GC
	454	Illumina						
Newbler → Velvet	980,095	61,903,430	19,698	23,206,886	100	55,899	5,594	54.42
Velvet → Celera	1,007,004	61,911,952	3,857	26,235,516	157	102,592	16,758	54.66
Velvet → MIRA	1,007,004	61,911,952	17,154	36,833,148	27	49,309	9,308	53.64
Celera	1,007,004	62,434,112	4,234	27,226,310	157	99,788	15,578	54.68
Velvet	980,095	61,903,430	16,832	21,768,812	100	28,488	6,499	55.64
CLC	1,007,004	62,434,112	10,895	24,534,397	30	94,799	10,590	54.70
Ray	980,095	61,903,430	21,982	24,174,789	100	21,982	4,307	54.75
Velvet pairs → EULER-SR	980,095	61,903,430	36,710	24,505,799	26	66,755	10,939	54.72

Table 2.7. A) Scaffolds from the combined genome assemblies for the A1. B) Combined genome assemblies for the A2 mating type.

A. Mating type A1

Assembly Method	Scaffolding Program(s)	Number of Complete EST Alignments	Number of Absent ESTs	Number of ESTs Broken Across Scaffolds	Number of Scaffolds	Total Bases in Scaffolds	Smallest Scaffold	Longest Scaffold	N50 (with minimum contig length 1000)
Newbler → Velvet	Newbler, Velvet	6,319	428	2,384	14,675	26,212,052	100	482,592	8,839
Velvet → Celera	Velvet, Celera	5,834	156	3,141	3,583	26,821,198	1,000	217,273	21,380
Velvet → MIRA	Velvet, Bambus	5,167	168	3,796	6,721	32,290,284	700	115,975	11,813
Celera	Celera	5,824	175	3,132	2,997	27,189,479	1,000	234,066	38,554
Velvet	Velvet	6,705	313	2,113	8,802	23,600,834	100	112,807	17,928
CLC	Bambus	6,093	362	2,676	8,844	24,739,453	34	155,137	20,637
Ray	Ray	5,370	472	3,289	20,118	23,046,565	100	19,451	4,306
EULER-SR	SSPACE	6,132	819	2,180	39,724	25,164,003	26	57,979	10,046

B. Mating type A2

Assembly Method	Scaffolding Program	Number of Complete EST Alignments	Number of Absent ESTs	Number of Partial EST Alignments	Number of Scaffolds	Total Bases in Scaffolds	Smallest Scaffold	Longest Scaffold	N50 (with minimum contig length 1000)
Newbler → Velvet	Newbler, Velvet	6,705	313	2,113	21,374	23,457,149	100	84,292	3,331
Velvet → Celera	Velvet, Celera	5,525	151	3,455	3,687	26,442,564	1,000	102,592	18,184
Velvet → MIRA	Velvet, Bambus	4,759	92	4,280	7,029	34,394,841	550	109,415	10,492
Celera	Celera	5,470	122	3,539	4,075	27,429,383	1,000	102,510	17,223
Velvet	Velvet	5,916	1,073	2,142	16,821	21,769,320	100	28,488	6,490
CLC	Bambus	5,857	125	3,149	10,156	24,578,737	29	106,293	15,885
Ray	Ray	5,408	295	3,428	19,309	25,982,152	100	42,333	10,632
Velvet pairs → EULER-SR	Velvet, SSPACE	6,787	858	1,486	30,650	25,173,218	26	147,391	28,625

Single program hybrid assemblies

For both mating types Celera performed well on the data; the longest contig in the A1 data (Table 2.6A) was 96 kb and the N_{50} was 17kb. The A2 data (Table 2.6B) produced a long contig of 99 kb with an N_{50} of 15 kb. Both the long contigs and N_{50} values are both larger than the single genome assemblies. The %GC was similar to that produced by the 454 single genome assemblies, suggesting that the GC bias was reduced here. The total base count of the A1 genome was 25 mb, which is an acceptable value, but the A2 genome was 27 mb, which is slightly too high. It was not possible to sufficiently alter Celera's parameters to account for this duplication. Celera scaffolding was also extremely effective, with long contigs of 234 kb and 102 kb for the A1 and A2 genomes respectively and N_{50} values of 38 kb and 17 kb. The A1 N_{50} value was higher than any single or hybrid genome assembly.

Velvet performed less well than Celera, with contig N_{50} values of only 9 kb and 6 kb and longest contigs of 43 kb and 28 kb for the A1 and A2 genomes respectively. In addition, the genome sizes of both the contigs and scaffolds were 21-23 mb; lower than the expected genome size. Scaffolding of the A1 genome produced a longest scaffold of 112 kb and an N_{50} value of 17 kb. The A2 genome produced a longest scaffold of only 28 kb with an N_{50} value of 6 kb; shorter than many of the single genome assemblies. The GC content of the genome is more similar to the single genome Illumina assemblies than the hybrid or 454 assemblies, suggesting that much 454 data might have been discarded. Altering parameters to increase the number of 454 reads included only decreased the overall contig lengths due to ambiguities and errors.

The performance of CLC on the data fell in the middle of the assembly programs; the base count in all assemblies of over 24 mb was acceptable, and the longest contig lengths of 67 kb (A1) and

94 kb (A2) were greater than most other *de Bruijn* methods but smaller than those created using Celera's OLC approach. Both the A1 and A2 contig N_{50} values were shorter than those produced by Celera. Scaffolding was performed with Bambus: the A1 assembly contained shorter contigs than the Celera assembly, but was still reasonable, with a longest scaffold length of 155 kb and an N_{50} value of 20 kb. Scaffolding of the A2 genome created one of the best A2 assemblies, with the longest contig length of 106 kb beating Celera's 102 kb. The N_{50} value was 15 kb, slightly lower than Celera's 17 kb. From this we might surmise that CLC was able to optimize its parameters to avoid confusion from the multiple genomes.

The contigs of the A1 EULER-SR assembly were longer than those created by Velvet, although not so long as the CLC and Celera assemblies. The longest contig was 57 kb, with an N_{50} value of 10 kb — longer than the single genome assemblies. The A2 paired Illumina genome was pre-assembled using Velvet, as EULER-SR was not able to produce assemblies using the raw data. Unfortunately, EULER-SR either shredded or discarded the longest scaffolds; effectively losing the Illumina paired read data. Despite this the longest contig was 66 kb and the N_{50} value was 10 kb, once again larger than Velvet but not CLC or Celera. The base contents of the EULER-SR contig assemblies were 25 mb and 24 mb for the A1 and A2 genomes respectively, which are acceptable values, and the GC content was ~54% compared to Velvet's 55% suggesting that more of the 454 data was used. SSPACE was used to scaffold the genomes. The A1 genome was not altered from the contig result, as few read pairs could be mapped to the scaffolds, and any that were had already been satisfied by EULER-SR. The A2 genome, scaffolded using both the Illumina and 454 paired reads, was improved, and the long contig length was increased to 147 kb; the longest of any A2 scaffolds, and the N_{50} value to 28 kb, also the longest created in any A2 genome. Despite these values being high, they were still smaller than the scaffold lengths of the Velvet single genome assembly.

Ray performed the worst of all the assemblers, with longest contig lengths of 19 kb (A1) and 21 kb (A2) and N_{50} values of 4 kb. The base counts of 23-24 mb were slightly low. Scaffolding did not greatly improve the results; the A1 genome doubled in longest contig length (21 kb to 42 kb) but was small in comparison to other program assemblies, as was the N_{50} value of 10 kb. The A2 genome remained the same.

Combined program hybrid assemblies

Adding Newbler assemblies to Velvet produced some of the longest contigs of any of the assemblies (107 kb (A1), almost identical to Velvet/Celera's 107 kb (A1)); however the N_{50} for the same genome was around 6 kb shorter than both of the Celera A1 assemblies. The A2 assembly was much poorer; the longest contig was only 55 mb with an N_{50} value of 5 kb. The longest scaffold of the A1 genome was the largest of any hybrid assembly, but the N_{50} value was only 8 kb — a similar result to the single genome 454 assemblies. Indeed, the contig GC contents are the lowest of all programs with the exception of the MIRA assemblies, suggesting that these assemblies incorporated more of the 454 data. The A2 longest scaffold was 82 kb, with an N_{50} of just 3 kb; the worst of any of the hybrid assemblies.

The combined assembly of Velvet and Celera performed similarly to the Celera assembly. For both genomes, the longest contigs were longer than those assembled using only Celera. The A1 N_{50} values were extremely similar (~17 kb), and the A2 N_{50} value was 16 kb, compared to 15 kb when using Celera only. The contig data, therefore, suggests that this combination might be better than the Celera only assemblies. All genome base counts were also very similar. Scaffolding the data produced a slightly different result. In both genomes (A1 and A2) the base content was reduced by ~1 mb, from 27 to 26 mb, which is potentially a result of Velvet's graph pruning.

Both the longest contig and the N_{50} value are reduced for the A1 assembly compared to the Celera assembly. The A2 genome shows the opposite effect, with the longest contig and N_{50} value only very slightly increased over the Celera assembly.

The Velvet/MIRA assemblies performed badly on the data, as it would appear that MIRA is too sensitive to assemble heterozygous genomes, and instead assembled polymorphic regions separately. This is reflected in the extremely large genome sizes of 35 and 36 mb for the A1 and A2 genomes respectively. The low GC content of these assemblies would also suggest that the 454 data was assembled separately from most of the Illumina contigs. The contig assemblies contained longest contigs of 52 kb (A1) and 49 kb (A2), which was short for the hybrid assemblies but still longer than the single genomes assemblies. The N_{50} values were low, which is likely to be due to a large number of short heterozygous contigs that were not assembled together. Scaffolding was performed using Bambus, as MIRA does not contain a scaffolding module. The longest scaffolds were large; 115 kb in the A1 genome and 109 kb in the A2 genome, however, the N_{50} values remain low, at 11 kb and 10 kb.

Assessing the hybrid assembly scaffolds

The Velvet assembly had the greatest number of complete ESTs aligned in the A1 consensus genome, and EULER-SR in the A2, with 6,705 and 6,787 complete matches respectively. In both genomes, EULER-SR and Velvet were the two programs with the greatest number of complete alignments. Ray generally produced a genome with the fewest EST alignments, with Celera performing slightly better. CLC was once again in the middle of the programs assessed. The number of absent ESTs followed a different pattern again, with Celera missing the fewest EST matches. Velvet performed well on the A1 data set, although worse than Celera, however the A2

data set, which it had difficulty assembling was extremely poor, with 1,073 missing ESTs, possibly due to the low base count of the overall assembly. The EULER-SR assemblies generally contained a high number of missed ESTs, despite a reasonably high genome base count. CLC produced genomes with a relatively low number of missed ESTs, with 362 in the A1 data set, and only 125 in the A2 genome. The Celera and Ray assemblies contained the highest number of broken and duplicated EST matches. For Celera, this is likely to be due to the large genome size produced by the assemblies (~27 mb) leading to duplicates. It is likely that the poor assembly of Ray resulted in a high number of broken ESTs. In the A1 data set, the Velvet assembly aligned to the fewest broken/duplicated ESTs. For the A2 data set, EULER-SR performed the best, with only 1,486 broken/duplicated genes aligned to the genome. The performance of CLC was average, containing the third highest number of broken/duplicated ESTs in both the A1 and A2 genomes.

The multiple program assemblies did not perform noticeably better on the data than the single program assemblies. In both the A1 and A2 genomes, the Newbler/Velvet assemblies contained the greatest number of complete EST alignments, though not more than the top program from each of the single program assemblies. For both genomes, the Velvet/MIRA assembly contained the fewest complete ESTs, and some of the least of all the hybrid assembly approaches. The numbers of absent ESTs were highest in the Newbler/Velvet assembly for both the A1 and A2 genomes. In the A1 genome, the Velvet/Celera assembly contained the fewest missing ESTs; the fewest of any of the assembly methods, and in the A2 genome the Velvet/MIRA assembly contained the fewest of any assembly method in any genome, with only 92 ESTs missing, this might be due to the increased genome size of 34 mb. For both the A1 and A2 genomes, the MIRA assemblies contained the greatest number of broken and duplicated genes, and the Newbler/Velvet assembly the least, exactly following the trend in genome base count. The MIRA

assemblies contained the highest number in any assembly of any genome, again reflecting the massive genome sizes.

Comparing the assemblies

There was no clear best approach when producing the hybrid genomes, and after assembly, there was no clear best assembly. Overall, the *de Bruijn* methods produced some of the most accurate assemblies, as determined by the EST alignments, but with reasonably short scaffold lengths. The OLC methods on the other hand often created long scaffolds, but this came at the cost of accuracy. The single program assemblies were not obviously better or worse than the multiple program assemblies. The hybrid genome assemblies generally contained more complete EST matches than the single genome assemblies, with the exception of the Velvet paired end assembly which was comparable to the hybrid assemblies with the most complete matches.

2.3.4 ITERATIVE SCAFFOLDING

Though the hybrid assemblies were able to improve slightly on the single genome assemblies, it is obvious from the data that the balance between polymorphisms and errors is extremely difficult to find. Thus, assembling the separate genomes might always contain either chimeric data or repeated contigs, and may also involve the removal of valid contigs from assembly. To avoid this problem an iterative scaffolding approach was taken whereby only a single genome for the A2 mating type was included in the assembly, and then scaffolded against other genomes. The results of the scaffolding are presented in Table 2.8. Performing this type of mapping was found to vastly improve the assembly, and is likely to be the best approach for the assembly of a more complete

Table 2.8. Iterative mapping of paired end reads onto the A2 Illumina *M. violaceum* genome contigs over 500 bp. The *M. violaceum* 454 A2 genome plus additional Illumina mate pair reads were used for mapping. Each row of the data builds on the previous scaffold output.

Mapping data	Scaffolding Program(s)	Number of Paired Joins	Number of Transcript joins	Number of Complete EST Alignments	Number of Absent ESTs	Number of ESTs Broken Across Scaffolds	Number of Scaffolds	Total Bases in Scaffolds	Smallest Scaffold	Longest Scaffold	N50 (with minimum contig length 500)
Original Velvet Assembly	Velvet	-	-	6,955	829	1,347	2,185	22,530,967	501	196,899	33,450
Original A2 Paired reads	Bambus	105,990	-	7,048	816	1,267	1,336	22,581,907	502	248,513	47,699
A2 454 Paired end reads	Bambus	1303	1	7,028	938	1,165	796	22,223,440	502	308,231	72,683
Illumina mate pair (A2)	Bambus	11,590	-	7,080	832	1,219	349	22,513,558	517	1,520,898	201,086
Illumina Mate pair (A2)	Bambus	3,747	-	7,053	922	1,156	210	22,521,898	517	1,863,588	476,125

reference genome. The number of complete ESTs that could be mapped to the scaffolds also generally increased with each scaffolding to around 7,000; higher than any of the previous assemblies. The number of absent ESTs was unfortunately also high (800-900) due to the low base count of the genome (22 mb) which resulted from the loss of contigs that were not mapped on to the genome. This reduction can therefore be easily remedied in future assemblies.

2.3.5 MAPPNIG ASSEMBLIES

In accordance with the expectation, the mapping assemblies produced longer contigs and higher N_{50} values than any of the *de novo* assemblies. The detailed results are provided in Table 2.9. The results provided represent the consensus assembly for each new genome mapped to the *de novo* single genome Velvet assembly of the A2 paired end Illumina genome from *M. violaceum* var. *latifolia*. Generating a consensus in this way has one major drawback — sequences that did not map onto the reference genome were lost from the consensus genomes. For the purpose of our evolutionary analysis this did not prove to be a problem, as only genes present in all genomes were required. The genome sizes recovered from the mapping assemblies varied, with *M. violaceum* var. *Dianthus* containing the fewest bases that could be mapped to the reference and *M. violaceum* var. *flos-cuculi* the most. *M. violaceum* var. *Dianthus* was able to create 1,875 of the 2,185 scaffolds in the reference genome, which was the fewest of any of the genomes, and *M. violaceum* var. *flos-cuculi* was closest to the reference number, re-creating 2,181 of the 2,185 scaffolds. The maximum contig lengths and N_{50} values were also found to follow this pattern; the more bases that could be mapped onto the genome, the greater the longest contig length and N_{50} value.

Table 2.8. *Microbotryum violaceum* genomes mapped onto the A2 Microbotryum genome from *Silene latifolia* sequenced using Illumina technology. Contig assemblies produced using CLC.

Host Species / Mating Type	Number of Contigs	Total Length of Contigs	Total Bases in Contigs	Shortest Contig	Longest Contig	Mean Contig Length	N50	%GC
Silene latifolia	2,185	21,639,957	21,639,957	136	196,530	9,903	33,064	55.64
Silene vulgaris A2	2,150	19,836,354	19,836,351	16	188,764	9,226	31,938	56,28
Silene dioica A2	3,978	22,665,570	22,665,570	180	120,512	5,697	19,979	55.61
Silene caroliniana A2	2,157	18,691,907	18,691,905	16	178,073	8,665	30,375	56.48
Saponaria officinalis A2	2,152	20,021,254	20,021,250	16	188,916	9,303	31,850	56.48
Lychnis flos-cuculi A2	2,181	21,054,092	21,054,091	16	195,416	9,653	32,904	55.82
Dianthus carthusianorum A2	1,875	16,551,794	16,551,345	18	157,884	8,827	26,778	56.26

2.4 DISCUSSION AND CONCLUSIONS

Many different approaches were taken to assemble the data using both individual programs and combinations. Performing a single assembly of a single genome resulted in highly fragmented assemblies, but with fewer duplicated regions than the hybrid genome assemblies. The most effective way to assemble the data appeared to be by iteratively scaffolding a single assembled genome by mapping paired end and mate pair reads onto the assembly.

The highly fragmented nature of the single genome *de novo M. violaceum* var. *latifolia* assemblies is likely due to short read lengths, lack of paired end data in the single assemblies, and repeat regions in the genome. The assemblies of the single 454 genomes revealed that Celera and Newbler performed similarly on the data, with no obvious affect of the difference in repeat handling. Both programs were potentially worth using for any 454 assembly. Celera was able to re-create more of the genome by including repetitive and ambiguous regions, and contained fewer false EST joins during scaffolding; further assembly of the 454 sequenced individuals might therefore use these assemblies. For the Illumina genomes, the *de Bruijn* assembler Velvet was the clear choice over the greedy string assembler QSRA, producing both longer contigs and N_{50} values. Choosing between Illumina and 454 data depends upon the required outcome; 454 assemblies were likely to produce longer contigs, but Illumina assemblies created high N_{50} values. It might also be worth noting that no false EST joins were detected during scaffolding of the Illumina data, unlike the 454 data, therefore the best Illumina scaffolds were used for further assembly and mapping of the different species genomes.

The single genome assemblies demonstrated that creating reasonably sized assemblies from the *Microbotryum* data is almost impossible without the use of paired end data, therefore to assemble the genomes without paired end data a mapping approach would be recommended. Scaffolding using paired end data was predominantly more effective when the assembly software contained an internal scaffolding module, rather than using a stand-alone scaffolder, however separate scaffolders can be particularly useful during hybrid assemblies carried out in stages, or when the assembler is not able to create scaffolds.

The use of the scaffolding program Bambus to create contig links based on transcript data is beneficial for genome finishing, but was found to reduce the length of most assemblies. From the results it was apparent that a large number of additional joins were found using transcriptome data. Unfortunately, Bambus is designed to find all possible paths between contigs, often leading to ambiguous contig joins, which can be useful when manually examining results but not for automated assembly. To resolve this problem and use only unambiguous scaffolds to avoid false joins, Bambus performs an untangling process, which removes ambiguities by choosing the longest path through the graph that does not overlap itself (Pop, et al., 2004), which often resulted in the shortening of scaffold lengths.

Assembling multiple sequencing technologies together proved to be a difficult task. A large part of this difficulty stems from the algorithms used; OLC assembly methods are usually too memory and time intensive to use on short read data (i.e. < 200 bp), due to the volume of reads produced. *De Bruijn* assemblers can theoretically work on any type of data, but can struggle to assemble long reads effectively and were disrupted to a greater degree by sequence errors and SNPs. Our results do not provide a clear indication of whether it is more efficient to assemble the

Microbotryum data in stages or as a fully hybrid assembly, however, it is likely that with slightly longer Illumina read lengths of around 100 bp, such as those now being produced, the OLC method Celera would be one of the most efficient programs for hybrid genome assembly. The *de Bruijn* based assembler Ray was found to be the most sensitive program to SNPs and sequence error and performed the worst of all programs tested. This program extends seeds rather than searching for Eularian walks to find a path through the data; the path is then broken at ambiguous read clusters (Boisvert, et al., 2010). It is therefore possible that the algorithm would have been intrinsically sensitive to the heterozygous reads introduced through the different individuals, which would create large ambiguous read clusters with high coverage.

Combining genomes from different organisms was not expected to cause many problems for the assembly software, given that the individuals were from the same species of *Microbotryum*. In reality, the hybrid assemblies were all hindered to varying degrees by SNPs between the genomes and sequence error. Population genetics analysis on the individuals of *M. violaceum* var. *latifolia* from different regions might help to explain whether there is higher than expected divergence between populations, or if a similar result would be found with different individuals of any species. The nature of any graph-based assembly method makes it fundamentally difficult to assemble heterozygous regions, and although it might be possible to use coverage to detect genuine polymorphic regions, it was not possible with the programs available to resolve these regions, despite parameter optimization. Overall, the *de Bruijn* based methods performed the least well in terms of N_{50} and longest contig length on the hybrid assemblies, however, these assemblies could have a greater number of complete ESTs aligned to them, suggesting that the *de Bruijn* methods produced the most accurate hybrid assemblies. The most efficient solution to this problem was the iterative mapping of paired reads onto a single genome. A different approach

might be to attempt to combine the genomes using assembly combiner software such as the MatLab package MAIA (Nijkamp, et al., 2010) or Minimus (Sommer, et al., 2007).

Creating assemblies by mapping is obviously more effective at recreating long contigs and producing assemblies with high N_{50} values than assembling *de novo*. However, as genomes diverge, rearrangements and regions that will not align to the reference become more abundant, and can be clearly seen in our mapping results, as the total genome base counts decrease with divergence. The use of paired end information for scaffolding can help to resolve this difficulty, but there will always be regions of the genome that must be assembled *de novo*. Fortunately, *de novo* genome assembly is still developing and more efficient and sophisticated algorithms in combination with decreasing sequencing cost and increasing read lengths, data output, reliability, speed and computer resources will aid the genomic assembly process. It is likely that many short read assembly methods will soon become obsolete due to increased read length, but as genome assembly algorithms fall into the mathematical complexity class of NP hard problems, assembly solutions will continue to take a heuristic approach, and therefore an assembly is never certain to produce the best result. Despite the improvements in heuristics and computational advances, it is unlikely that genomes will be finished without human checking. Fortunately, every new genome that is finished provides a potential template for other genome sequencing projects, and so it is likely that mapping assemblies will become more prolific as the volume of finished data increases. This will greatly increase the speed with which genomes can be produced.

Here, we have successfully assembled the genomes of seven *Microbotryum* species, including six genomes of a single species. The genomes of the same species analysed here proved to be too divergent for most genome assembly methods to reliably assemble together, although iterative

scaffolding of a single genome using paired-end reads proved to be an effective method for reducing assembly gaps. For the purpose of creating long scaffolds, it would also be possible to manually resolve repeat regions using the stand-alone scaffolding module Bambus. The best single genome assembly provided a reasonable template upon which to map the additional species genomes. With an initial goal of the evolutionary analysis of coding content, it is sufficient to have genomes mapped onto the assembly with the highest number of complete EST alignments and unambiguous scaffolds.

Overall, a number of points have been learned from these assemblies. When using 454 data for *de novo* assembly, both Newbler and Celera are recommended, as both software are easily obtainable, designed with 454 reads in mind and produce extremely similar results. Additionally, both programs were fast, and memory effective, although Newbler required the least Bioinformatics knowledge to use. When *de novo* assembling Illumina reads, the program Velvet is recommended as it produces accurate and long assemblies; the main drawback with Velvet is that it is extremely memory intensive, as required memory increases with the amount of read data. Caution should be taken when using new programs, and trialling such a large number of programs here did not improve results. Indeed, with large sequencing centres continuously updating their software (e.g. Newbler, Celera and Velvet), these programs are likely to be the most reliable. Older programs such as QSRA and EULER-SR are also not recommended over these well-used programs. If using paired-end data, programs with internal scaffolding modules should be used in preference over external scaffolders. Mapping assemblies were found to be more efficient than *de novo* assemblies, though rearrangements between species could create mis-assemblies. This problem could in part be resolved by using paired-end data to search for these mis-assemblies.

Unless a reference sequence is available for mapping, it is not recommended to sequence without paired-end information, as results in most cases will be extremely poor. In our assemblies, scaffolding of paired end data produced results with longest contigs nearly an order of magnitude longer than from the un-paired data. Interestingly, increasing coverage does not necessarily improve sequence assemblies; the additional data was found to cause problems due to sequence errors. It is therefore more important to focus on the quality of the reads rather than quantity. Indeed, further reads were sequenced for a number of *M. violaceum* var. *latifolia* species increasing the coverage over that of the Illumina paired end data; the resultant contigs/scaffolds were of similar lengths to the original assemblies (data not shown). Coverage is even less important with longer reads; the paired 454 assemblies produced here were of similar lengths to the paired Illumina assemblies but with lower coverage.

Caution should be taken when combining data from multiple individuals within a species, as SNPs between individuals are likely to disrupt assembly graphs, resulting in chimeric data, duplicated contigs or loss of the SNP in question. Indeed, if assembling multiple organisms together in order to get a consensus reference (as was the rationale here), the most effective method, if paired end data is available, appears to be to assemble the highest coverage genome and to map additional reads to it for use in scaffolding. If paired end data is not available, extremely thorough filtering of reads is suggested before a combined assembly, to minimise the effects of SNPs and sequence errors.

Although the programs used for hybrid assembly struggled, Celera, Velvet and CLC were all able to successfully assemble the 454 and Illumina reads together. Therefore, if performing hybrid assemblies with sequences from the same organism it is likely that any of these programs could be effective. As before, when increasing read data (particularly short read data) increased filtering

is also suggested. There was no large difference found between hybrid assemblies using a single program over multiple programs; therefore time might be saved by using a single program only.

2.5 REFERENCES

Almaraz, T., *et al.* (2002) Phylogenetic relationships among smut fungi parasitizing dicotyledons based on ITS sequence analysis, *Mycological Research*, **106**, 541-548.

Bao, S., *et al.* (2011) Evaluation of next-generation sequencing software in mapping and assembly, *J Hum Genet*.

Bennett, S. (2004) Solexa Ltd, *Pharmacogenomics*, **5**, 433-438.

Boetzer, M., *et al.* (2010) Scaffolding pre-assembled contigs using SSPACE, *Bioinformatics*.

Boisvert, S., Laviolette, F. and Corbeil, J. (2010) Ray: Simultaneous assembly of reads from a mix of high-throughput sequencing technologies, *Journal of Computational Biology*, **17**, 1519-1533.

Bryant, D., Wong, W.-K. and Mockler, T. (2009) QSRA — a quality-value guided de novo short read assembler, *BMC Bioinformatics*, **10**, 69.

Chevreux, B., Wetter, T. and Suhai, S. (1999) Genome sequence assembly using trace signals and additional sequence information, *Computer Science and Biology: Proceedings of the German Conference on Bioinformatics (GCB)*, **99**, 45-56.

Delcher, A.L., *et al.* (2002) Fast algorithms for large-scale genome alignment and comparison, *Nucleic Acids Research*, **30**, 2478-2483.

Denisov, G., *et al.* (2008) Consensus generation and variant detection by Celera Assembler, *Bioinformatics*, **24**, 1035-1040.

Devier, B., *et al.* (2009) Ancient trans-specific polymorphism at pheromone receptor genes in Basidiomycetes, *Genetics*, **181**, 209-223.

DiGuistini, S., *et al.* (2009) De novo genome sequence assembly of a filamentous fungus using Sanger, 454 and Illumina sequence data, *Genome Biology*, **10**, R94.

Dohm, J.C., *et al.* (2008) Substantial biases in ultra-short read data sets from high-throughput DNA sequencing, *Nucleic Acids Research*, **36**, e105.

Giraud, T., *et al.* (2008) Mating system of the anther smut fungus Microbotryum violaceum: selfing under heterothallism, *Eukaryotic Cell*, **7**, 765-775.

Gnerre, S., *et al.* (2009) Assisted assembly: How to improve a de novo genome assembly by using related species, *Genome Biology*, **10**, R88.

Harismendy, O., *et al.* (2009) Evaluation of next generation sequencing platforms for population targeted sequencing studies, *Genome Biology*, **10**, R32.

Hood, M.E., Katawczik, M. and Giraud, T. (2005) Repeat-induced point mutation and the population structure of transposable elements in Microbotryum violaceum, *Genetics*, **170**, 1081-1089.

Hunt, P., *et al.* (2010) Experimental evolution, genetic analysis and genome re-sequencing reveal the mutation conferring artemisinin resistance in an isogenic lineage of malaria parasites, *BMC Genomics*, **11**, 499.

Huse, S., *et al.* (2007) Accuracy and quality of massively parallel DNA pyrosequencing, *Genome Biology*, **8**, R143.

Jeck, W.R., *et al.* (2007) Extending assembly of short DNA sequences to handle error, *Bioinformatics*, **23**, 2942-2944.

Kamper, J., *et al.* (2006) Insights from the genome of the biotrophic fungal plant pathogen Ustilago maydis, *Nature*, **444**, 97-101.

Le Gac, M., Hood, M.E. and Giraud, T. (2007) Evolution of reproductive isolation within a parasitic fungal species complex, *Evolution*, **61**, 1781-1787.

MacLean, D., Jones, J.D.G. and Studholme, D.J. (2009) Application of 'next-generation' sequencing technologies to microbial genetics, *Nat Rev Micro*, **7**, 287-296.

Margulies, M., *et al.* (2005) Genome sequencing in microfabricated high-density picolitre reactors, *Nature*, **437**, 376-380.

Miller, J.R., *et al.* (2008) Aggressive assembly of pyrosequencing reads with mates, *Bioinformatics*, **24**, 2818-2824.

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

Myers, E.W., *et al.* (2000) A whole-genome sssembly of Drosophila, *Science*, **287**, 2196-2204.

Nagarajan, N., Read, T.D. and Pop, M. (2008) Scaffolding and validation of bacterial genome assemblies using optical restriction maps, *Bioinformatics*, **24**, 1229-1235.

Ng, P., *et al.* (2006) Multiplex sequencing of paired-end ditags (MS-PET): a strategy for the ultra-high-throughput analysis of transcriptomes and genomes, *Nucleic Acids Research*, **34**, e84.

Nijkamp, J., *et al.* (2010) Integrating genome assemblies with MAIA, *Bioinformatics*, **26**, i433-i439.

Nowrousian, M., *et al.* (2010) De novo assembly of a 40 Mb Eukaryotic genome from short sequence reads: Sordaria macrospora, a model organism for fungal morphogenesis, *PLoS Genet*, **6**, e1000891.

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**, 9748-9753.

Pop, M. (2009) Genome assembly reborn: Recent computational challenges, *Briefings in Bioinformatics*, **10**, 354-366.

Pop, M., Kosack, D.S. and Salzberg, S.L. (2004) Hierarchical scaffolding with Bambus, *Genome Research*, **14**, 149-159.

Refregier, G., *et al.* (2008) Cophylogeny of the anther smut fungi and their caryophyllaceous hosts: Prevalence of host shifts and importance of delimiting parasite species for inferring cospeciation, *BMC Evolutionary Biology*, **8**, 100.

Sanger, F., *et al.* (1982) Nucleotide sequence of bacteriophage [lambda] DNA, *Journal of Molecular Biology*, **162**, 729-773.

Schatz, M.C., Delcher, A.L. and Salzberg, S.L. (2010) Assembly of large genomes using second-generation sequencing, *Genome Research*, **20**, 1165-1173.

Siegel, A.F., *et al.* (2000) Modeling the feasibility of whole-genome shotgun sequencing using a pairwise end strategy, *Genomics*, **68**, 237-246.

Simpson, J.T., *et al.* (2009) ABySS: A parallel assembler for short read sequence data, *Genome Research*, **19**, 1117-1123.

Sommer, D., *et al.* (2007) Minimus: A fast, lightweight genome assembler, *BMC Bioinformatics*, **8**, 64.

Steuernagel, B., *et al.* (2009) De novo 454 sequencing of barcoded BAC pools for comprehensive gene survey and genome analysis in the complex genome of barley, *BMC Genomics*, **10**, 547.

Thrall, P.H., Biere, A. and Antonovics, J. (1993) Plant life-history and disease susceptibility — the occurrence of Ustilago violacea on different species within the Caryophyllaceae, *Journal of Ecology*, **81**, 489-498.

Tong, P., *et al.* (2010) Sequencing and analysis of an Irish human genome, *Genome Biology*, **11**, R91.

Turner, T.L., *et al.* (2010) Population resequencing reveals local adaptation of Arabidopsis lyrata to serpentine soils, *Nat Genet*, **42**, 260-263.

Vercken, E., *et al.* (2010) Glacial refugia in pathogens: European genetic structure of anther smut pathogens on Silene latifolia and Silene dioica, *PLoS Pathog*, **6**, e1001229.

Votintseva, A.A. and Filatov, D.A. (2009) Evolutionary strata in a small mating-type-specific region of the smut fungus Microbotryum violaceum, *Genetics*, **182**, 1391-1396.

Votintseva, A.A. and Filatov, D.A. (2011) DNA polymorphism in recombining and non-recombing mating-type-specific loci of the smut fungus Microbotryum, *Heredity*, **106**, 936-944.

Wang, J., *et al.* (2008) The diploid genome sequence of an Asian individual, *Nature*, **456**, 60-65.

Wu, T.D. and Watanabe, C.K. (2005) GMAP: A genomic mapping and alignment program for mRNA and EST sequences, *Bioinformatics*, **21**, 1859-1875.

Yockteng, R., *et al.* (2007) Expressed sequences tags of the anther smut fungus, Microbotryum violaceum, identify mating and pathogenicity genes, *BMC Genomics*, **8**, 272.

Zerbino, D.R. and Birney, E. (2008) Velvet: Algorithms for *de novo* short read assembly using de Bruijn graphs, *Genome Research*, **18**, 821-829.

Zerbino, D.R., *et al.* (2009) Pebble and Rock Band: Heuristic resolution of repeats and scaffolding in the Velvet short-read *de novo* assembler, *PLoS ONE*, **4**, e8407.

Zhang, W., *et al.* (2011) A practical comparison of *de novo* genome assembly software tools for Next-Generation Sequencing technologies, *PLoS ONE*, **6**, e17915.

3. FRANC: A SIMPLE, POWERFUL, FLEXIBLE GENOME ANNOTATION RESOURCE

KATE E. RIDOUT

3.1 INTRODUCTION

The identification and classification of structural and functional elements is one of the most important goals of modern sequencing projects. The difficulty of this task depends on how well a genome has been characterised, the presence of sequenced close relatives, and the complexity of the genome. With whole genome sequencing becoming accessible to individual labs, *de novo* sequencing projects can arise where no close relatives of the organism in question have been sequenced from which annotation information might be obtained. In cases such as these, it is not feasible to characterise the genomic content from existing information or experimental procedures. Thus, structural and functional elements must be predicted. With the high demand for simple and efficient gene prediction software, many researchers and consortia predicting their own genes make genomic annotation software freely available (Korf, et al., 2001; Schiex, et al., 2001; Stanke and Waack, 2003). The elements of the genome that can be predicted include gene and transcript prediction, repetitive element detection and identification, both small (microsatellites, tandem repeats; Benson, (1999)) and large (retrotransposons; Ellinghaus, et al., (2008)), splice sites (Hebsgaard, et al., 1996) and regulatory sites (Newberg, et al., 2007).

Current gene prediction and identification methods can be broadly divided into two categories; homology-based and *ab initio*, each with distinct benefits. In order to predict a gene, software must have an idea of the features a gene is supposed to possess, and will search for DNA sequences that are likely to be similar. Homology-based prediction software takes homologous genes – and potentially expressed sequence tags (ESTs) – from closely related species (inter- and intra-specific) and searches for conserved exons. Additional signals such as transcription start sites, common splice sites and translation start sites can then be used to refine gene boundaries. This approach relies upon gene homologues existing in databases that are sufficiently similar to

be detected. *Ab initio* gene prediction combines the signals previously mentioned with other gene compositional information obtained from related species. These include the hexamer frequency spectrum of the coding regions, nucleotide content and codon content. This information is then used by modelling algorithms; for more information see Sleator, (2010). Additional *ab initio* gene prediction methods can be combined with a homology approach to include the use of multiple closely related genomes to search for conserved regions of DNA that might represent genes (Korf, et al., 2001). The reliability of *ab initio* gene prediction generally depends on the presence of a training set from a closely related organism. The more closely related the comparison species and the more genes that can be used for training, the more reliable the predictions. Though these methods do not rely on the genes to be predicted being already known in other species, they can be less reliable and will tend to have an increased sensitivity (more elements can be found) but a lower specificity (fewer of the elements that actually exist are correctly predicted).

Repetitive elements can be predicted with varying degrees of success, depending on their complexity. Care must be taken when identifying and masking regions identified by prediction software; genes and gene elements can be identified and masked as repeat structures. Repetitive elements generally fall into two classes: tandem arrays and interspersed repeats. Tandem arrays comprise adjacent sequence duplicates (or approximate duplicates), such as minisatellites, microsatellites and other low complexity sequences, and can be identified through various pattern-matching algorithms such as the Smith-Waterman approach taken by the program Crossmatch (http://www.phrap.org/phredphrapconsed.html). Interspersed repeats include features such as transposable elements that are able to copy themselves across the genome and contain coding regions. These mobile elements can be divided into three groups: LTR retotransposons, non-LTR retrotransposons (such as LINEs and SINEs) and DNA transposons. The identification of interspersed repeats is a task well suited to homology-based methods, as key proteins required

for duplication of the elements are often conserved, but programs such as LTRharvest are also able to identify LTR retrotransposons without homology information, as the structure of these elements lends itself to modelling approaches (Ellinghaus, et al., 2008).

Though prediction programs are easily obtainable, they are often designed for model organisms (Reese, et al., 2000; Salamov and Solovyev, 2000), and in many cases the software available will be only a portion of a larger pipeline which is not accessible, such as those used by the Venter Institute, Broad Institute and Vector Base. In many cases, the gene prediction procedure will also require a large amount of training data (Blanco, et al., 2002; Majoros, et al., 2004; Stanke and Waack, 2003). Often, the nearest sequenced relative or training set to a novel genome will be too divergent to be used to predict genes reliably – a challenge that is greatly increased in intron-rich eukaryotic genomes. When genomes are large and rich in introns and repeats, annotation becomes difficult for both homology-based and *ab initio* methods. It is tempting to simply align transcriptome information onto a genome in order to detect genes; unfortunately, alignments of ESTs provide an accurate representation of a partial exon only and not the whole gene, and protein families may make it difficult to find the correct EST location. For a review of gene prediction approaches see Mathé, et al., (2002) and Sleator, (2010). These difficulties mean that automated and reliable gene predictions are extremely challenging to produce. To increase the reliability of genome annotations, a number of combiner algorithms have been produced (Allen and Salzberg, 2005; Brejová, et al., 2005; Elsik, et al., 2007; Haas, et al., 2008), which have been demonstrated to generate the most accurate gene sets of any of the gene finding approaches (Coghlan, et al., 2008). Unfortunately, these programs can require a long time and considerable bioinformatic support to run the various programs and gather the outputs into a usable format.

The pipeline presented here, named FRANC (Functional and Repetitive ANnotation Combiner), has been designed to address this problem. It provides a flexible annotation resource that automatically formats all outputs and pipes them into the combiner program EvidenceModeler (Haas, et al., 2008). FRANC is designed to integrate repeat, EST and known protein evidence, and to use BLAST searches in order to include information on protein homology. These results are formatted to be used as training information for five major gene prediction programs, making the pipeline ideal for novel genomes with no annotated close relative. The pipeline can be iteratively trained on the most reliable output to further refine the results. A weighting system, as provided by EvidenceModeler allows the user to identify the sources of data that are most reliable; example weighting files are provided with the pipeline that provide methods of producing results that are either more sensitive or specific. The programs available through the pipeline include LTRharvest, RepeatMasker, Tandem Repeats Finder, Crossmatch, BLAST, Exonerate, GeneWise, GMAP, GeneMark, AUGUSTUS, GlimmerHMM, GeneID, EuGene and the combiner EvidenceModeler. FRANC's modular design means that further programs, developed in the future, can also be easily accommodated. All of these programs can be run individually, or part of the same process, depending on the user's preference, time and computing resources. FRANC has been designed for the use of labs with any genome, ranging from those that are well characterised, to vastly different from the closest sequenced relative with no training set.

3.2 METHODS

Functional and Repetitive, ANnotation Combiner (FRANC) is divided into four sections, two of which are provided as separate (independent) perl modules that are run together by a perl wrapper. These modules comprise a repeat masking step and a gene prediction step. The self-

training segment is included as part of the wrapper script, as is the final stage of combining all predictions. Results from external programs are provided individually, as well as the combined prediction, and all stages and programs can be run independently. All program results are provided in GFF3 (generic feature format 3) files for compliance with current annotation standards. Training data can be incorporated into various stages of the pipeline, or alternatively, many of the gene prediction programs are distributed with specific pre-trained organism options. Owing to the thoroughness of the complete pipeline, the initial gene set may be small, therefore re-training either the pipeline or programs within the pipeline on this gene set is recommended.

There are five major steps to the annotation pipeline FRANC, which should be run in order. The first is the repeat masking stage: genomes should be masked for transposable elements before gene prediction is carried out, otherwise these elements may be predicted as genes. The second stage is the production of a training set, which is designed to train other gene prediction programs. This combines a gene predictor requiring no training data with NCBI BLAST (Basic Local Alignment Search Tool) searches and filtering to identify predicted genes that can be verified by known proteins. Without this stage, the gene prediction programs must be run with the most appropriate training set provided with each program. The third stage is the integration of homology evidence by mapping known proteins and ESTs onto the sequences to be annotated. This stage should be carried out before the gene predictions, as they can be entered into a number of programs to aid prediction. The fourth stage is the gene prediction step, which included sequences masked in stage 1, programs trained using data from stage 2, and homology information gathered in stage 3. The final stage of the pipeline in the combining stage, which will collect evidence from all previous pipeline stages. An overview of the pipeline is displayed in Figure 3.1. The programs and stages are discussed in more detail below.

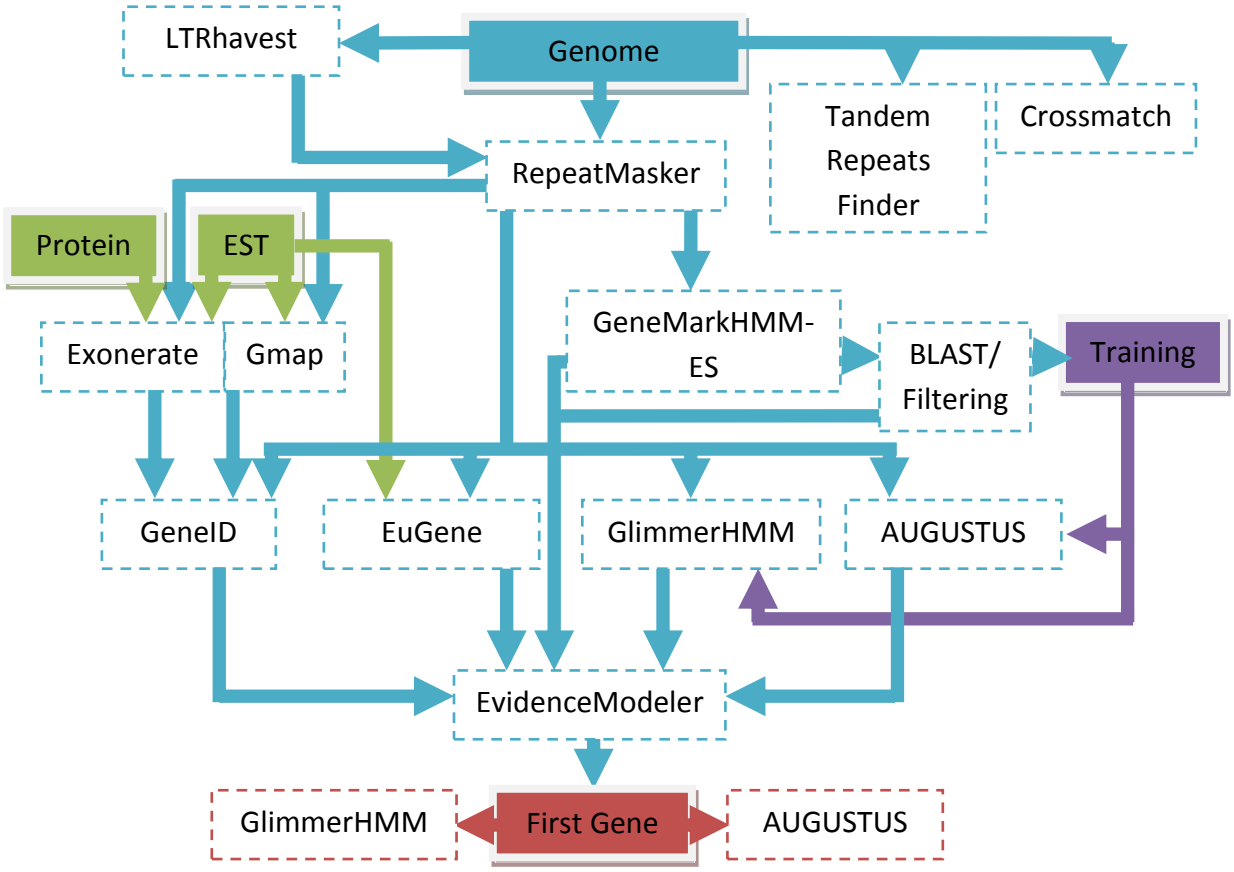

Figure 3.1. The gene prediction process: Sequence files are represented by block colours and programs with dashed lines. The blue arrows represent the initial pipeline run. Additional sources of evidence (known proteins and ESTs) are green. The training data created by the pipeline and used to train AUGUSTUS and GlimmerHMM is coloured purple. The second training iteration is represented by the red arrows.

3.2.1 REPEAT MASKING

The repeat masking module of FRANC comprises four programs: LTRharvest (Ellinghaus, et al., 2008), Repeat Masker (Smit, et al., 1996-2010), Tandem Repeats Finder (TRF) (Benson, 1999)

and Crossmatch (http://www.phrap.org/phredphrapconsed.html). These four programs are included to provide both homology-based and *de novo* identification of repetitive elements, and potential genome masking. LTRharvest and RepeatMasker are designed for eukaryotic organisms only. By contrast, RepeatMasker performs homology searching on a given library, set in the configuration file, and it is therefore theoretically possible to perform homology searching with DNA from any organism. TRF and Crossmatch are not restricted to eukaryotes.

LTRharvest is a tool that identifies full length LTR retrotransposons – a type of transposable element flanked by long terminal repeats and consistently containing several key genes (for more information, see Havecker, et al., (2004). LTR retrotransposons are highly abundant in eukaryotic genomes (Ellinghaus, et al., 2008), making up 8% of the human genome, 10% of the mouse genome, and are responsible for considerable variation in genome sizes in plants (Vitte and Panaud, 2005). Identification and masking of LTR retrotransposons is recommended in order to remove these from gene sets produced by the program, and thus removing the possibility of including these repeats when training prediction software. Alternatively, results from LTRharvest can be included into EvidenceModeler under the repeat identification section.

RepeatMasker is one of the best known software tools for identifying repetitive elements in DNA. It searches for low complexity regions of DNA such as homopolymer runs or simple tandem repeats, and interspersed repeats. Interspersed repeats include LTR retrotransposons, such as those masked by LTRharvest, non-LTR retrotransposons and DNA transposons. These elements are identified through homology-based comparisons with those elements already characterised from various species. The effectiveness of RepeatMasker will in part depend on the availability of a related genome to the study organism. Most homology will be due to conserved proteins

encoded within the interspersed repeat. Caution should be taken with RepeatMasker, as genes that have evolved from interspersed repeats can be masked. The low complexity searching is automatically switched off by FRANC, as low-complexity regions may be valid portions of certain genes.

Tandem Repeats Finder (TRF) was designed to identify and potentially mask stretches of DNA containing contiguous approximate copies of two or more nucleotide patterns. These stretches of DNA are known as tandem repeats, and can over time suffer mutations, hence why copies need only be similar, and not always exact. It is possible for genes to contain tandem repeats, thus data should not be masked prior to gene prediction, however, identification of tandem repeats in the genome and in genes can be valuable (Benson, 1999). TRF is presented here to provide additional information about the sequences to be annotated, and to mask the genome if necessary. It is not intended to be used as a part of gene prediction.

Crossmatch (http://www.phrap.org/phredphrapconsed.html) uses an enhanced version of the Smith-Waterman algorithm (Smith, et al., 1981) to compare DNA sequences. Here, we use it for comparing the DNA to itself, in order to detect repetitive regions of the genome. Performing gene predictions on data masked with Crossmatch is not recommended. As with TRF, Crossmatch is a valuable tool, and is intended to provide additional information on sequence data, rather than be used as part of the gene prediction process.

3.2.2 BLASTP TRAINING (GeneMark → BLASTP → Filtering → Formatting)

Producing a full training set requires a large number of accurate genes, including start and stop positions of all exons and introns. The hidden Markov Models produced during training are used to model gene predictions, and the algorithms require thousands of parameters. In order to produce the required number of parameters around 1000 genes must be used for training (Ter-Hovhannisyan, et al., 2008). Gene prediction programs that require a reliable training set produce far more accurate predictions than distantly trained or un-trained methods, however the experimental validation required to be certain of gene form and function is time consuming, expensive and not always practical. Here we create a proxy for a full training set using GeneMark and BLASTp (Basic Local Alignment Search Tool - protein).

GeneMark is a suite of gene prediction programs that do not require training information and are suitable for both prokaryotes and eukaryotes. GeneMark.hmm-ES (eukaryotic) or GeneMark.hmm-PS (prokaryotic) are suggested for this section of the pipeline, although any *GeneMark* program could be used. Both programs are self-training, and iteratively train complex intron models without the need for known genes (Ter-Hovhannisyan, et al., 2008).

GeneMark gene predictions are collected into translated fasta format by the pipeline and sent via an internet connection to the NCBI's BLASTp tool. The proteins are then queried against the NCBI's non-redundant protein database (nr) using multiple sequence alignments and the top hits are collected, returning them to the pipeline.

Filtering is then performed on the BLAST results. The first stage is for the pipeline to separate all results containing hits; those that did not return BLAST hits are moved to a 'no hits' file. The top 10 hits for each gene are then considered. For a gene to be included it must have a shared identity of at least 50% to at least one of the top 10 genes. Additionally, at least 90% of the gene must be included in at least one of the matches. This approach was chosen to account for differences in splicing and domain content across species.

Formatting is the final stage of the training section. Those genes that passed the filtering stages are output into GFF3 format for eventual incorporation into EvidenceModeler, as well an exon file suitable for using to train the gene prediction program GlimmerHMM.

3.2.3 INTEGRATING EST/CDNA AND PROTEIN HOMOLOGY EVIDENCE

As a proxy for a full training set, a number of reliable known proteins or ESTs for the species can be aligned to the genome to aid in gene boundary discovery, intron detection and gene identification. Currently, FRANC can use three programs for mapping external evidence to a given set of sequences: GeneWise (Birney, et al., 2004), Exonerate (Slater and Birney, 2005) and GMAP (Wu and Watanabe, 2005). The output of these programs can be used later in the pipeline by the gene prediction program GeneID, as well as by EvidenceModeler for creating the final data set. These programs are called via the perl module GenePred.pm, and can be run independently of the pipeline.

GeneWise was designed as a tool for gene prediction by protein homology, and has been used in the Ensembl annotation pipeline (Curwen, et al., 2004). Gene predictions are calculated using hidden Markov models by mapping known proteins to the genome and modelling introns. *Exonerate* can be used to map either proteins or cDNAs onto the genome. *GMAP* can be used to map cDNAs/ESTs to the genome using a word indexing strategy. Matching oligomers of 8bp long are found between genome and ESTs and the optimal global chain between them is then found. The methods used by GMAP result in a much faster run time than Exonerate.

3.2.4 GENE PREDICTIONS

Five gene prediction programs are available though FRANC thus far. These are GeneMark, AUGUSTUS, GlimmerHMM, GeneID and EuGene. The gene prediction section forms a separate perl module that is called by the pipeline (GenePred.pm), and can therefore be run independently of the pipeline framework. Any or all of the gene prediction programs can be run within this section. These five programs represent a combination of *ab initio* and homology-based prediction methods and can all be re-trained using gene information. Only GeneMark is currently able to predict genes within prokaryotic organisms.

GeneMark is presented in the BLAST training methods. It can be run as a stand-alone program within the gene prediction module, and can be applied to both prokaryotic and eukaryotic organisms.

AUGUSTUS uses hidden Markov models to predict genes in eukaryotic organisms using training data from related species (Stanke and Waack, 2003). A benefit of AUGUSTUS is that it can be iteratively trained using genes identified by the pipeline; this process can be improved by filtering for likely genes, such as those with EST support, or genes with homology to a known protein. AUGUSTUS is able to include a hints file, which can include cDNA evidence and known proteins. It is possible to train AUGUSTUS based on genes predicted by the pipeline using the AutoAUG.pl script provided by AUGUSTUS.

Glimmer-HMM is a eukaryotic gene finder which is based on generalised hidden Markov Models (Majoros, et al., 2004). The program requires a training set in order to produce parameters for the model, and provides a number of pre-trained parameter sets. As with AUGUSTUS, the less divergent the training set from the genes to be predicted, the more reliable the result will be. Glimmer-HMM does not incorporate additional gene support such as ESTs or proteins, but can be easily trained using exon information from known genes.

GeneID uses position weight arrays and a Markov model to predict coding regions in eukaryotic organisms (Guigó, et al., 1992). A large number of species parameter files (used to train the program models) are provided based on training of known genes, many of which are useful for multiple species. For example, it has been suggested that the human training set is acceptable for all mammalian species (Blanco, et al., 2002). Additional information from ESTs or known proteins can be added to the program in GFF3 format. GeneID has been developed for sensitivity over specificity; this means that a very large number of genes will be predicted, some of which might be broken at the intron boundaries. Additionally, genes begin from as little as one amino

acid in length. To account for this problem, a minimum gene length can be set in the configuration file, and genes below this length will be discarded.

EuGene is one of the most complex eukaryotic gene prediction programs, and uses a graph-based probabilistic model for gene prediction, incorporating a number of different evidence sources for prediction using small software components called plugins (Schiex, et al., 2001). An internet connection is required to run EuGene, as the program will query the Netgene2 (Brunak, et al., 1991; Hebsgaard, et al., 1996), SPlicePredictor and NetStart (Pedersen and Nielsen, 1997) web servers for sequence information such as splice sites. EuGene is the most sensitive of the gene prediction programs, and will produce the highest base output. To avoid splicing confusion, EuGene is likely to break genes at intron boundaries; therefore it is not the best program for creating a training set.

3.2.5 COMBINING RESULTS WITH EVIDENCEMODELER

The program EvidenceModeler (Haas, et al., 2008) is designed to combine gene predictions, repeat predictions and homology information from a variety of different sources using a maximum likelihood approach. FRANC uses EvidenceModeler to combine all transcript, protein homology, repeat region and gene prediction information generated in the pipeline with the exceptions of Crossmatch and Tandem Repeats Finder, which are intended for gathering statistics and potentially masking the genome. GFF3 files from the pipeline are automatically formatted for processing with EvidenceModeler, and output is converted back into GFF3 from the default output format. The most important factor of EvidenceModeler is its ability to weight the various inputs depending on the user's confidence in the data and desired output. Results can be weighted

for sensitivity or specificity, depending on user preference, and towards programs with more supporting evidence. Example weights files are included with the pipeline.

3.3 TESTING

The main goal of FRANC is to provide a framework for gene prediction in organisms with very little training data. As the programs used in the pipeline have been individually tested by the makers, they are not independently tested here. In order to assess the sensitivity and specificity of the pipeline, we performed pipeline analyses on a well-characterised genome using either distantly related or no training data, and following the suggested procedures. Using arbitrary training data in this way more closely resembles reality, where often the closest training data available is what must be used.

Chromosome 1 of *Arabidopsis thaliana,* known proteins and EST sequences were downloaded from The Arabidopsis Information Resource (TAIR ftp://ftp.arabidopsis.org/home/tair/Sequences/whole_chromosomes/) release 10 of the genome. This chromosome was chosen because it was the longest.

Tandem Repeats Finder and Crossmatch are included in the pipeline to collect information on the genomes rather than for masking, and are not required for gene prediction. RepeatMasker was used to mask interspersed repeats only, using predicted LTRs and the Arabidopsis repeat libraries (Smit, et al., 1996-2010). LTRharvest was used to mask the *Arabidopsis* genome for predicted LTRs – the masked genome was then used for all further predictions.

Training was carried out using GeneMark and BLAST, which do not themselves require training data. The results of these programs were then used to train AUGUSTUS and GlimmerHMM. Known proteins were mapped to the genome using Exonerate, and ESTs using GMAP.

Gene predictions were carried out using trained AUGUSTUS and GlimmerHMM. In AUGUSTUS, genes are predicted for specificity rather than sensitivity, and are some of the longest genes of any of the programs in the pipeline. In the results of the GlimmerHMM predictions, the number of predicted bases is second highest of all of the gene prediction programs tested (where EuGene produces the greatest number of bases). Genes are predicted for sensitivity and specificity, and are usually shorter than those produced by AUGUSTUS and sometimes broken at the intron boundaries. EuGene trained on the genome of the fungal plant pathogen *Microbotryum violaceum* was provided with the *Arabidopsis* EST data to aid predictions. Finally, GeneID trained on wheat was run, which also incorporated the Exonerate and GMAP EST and protein alignments. Weights for EvidenceModeler were set intuitively, as recommended (Haas, et al., 2008) and are shown below:

```
ABINITIO_PREDICTION   GeneMarkHMM-ES_BLAST  2
ABINITIO_PREDICTION   GeneMarkHMM-ES        1
ABINITIO_PREDICTION   AUGUSTUS              3
ABINITIO_PREDICTION   GlimmerHMM            3
ABINITIO_PREDICTION   geneid_v1.3                2
ABINITIO_PREDICTION   EuGene                    2
PROTEIN         exonerate_protein_alignments     5
TRANSCRIPT      gmap_transcript_alignments       5
```

EST and protein alignments were weighted highest, followed by the gene prediction programs trained within the pipeline, the additional trained programs and the test set (GeneMark/BLAST) and finally self-trained GeneMark. EvidenceModeler was used to combine all predictions. The results of these gene predictions are presented in Table 3.1.

Table 3.1. Prediction accuracy of the initial run of FRANC. AUGUSTUS and GlimmerHMM were trained on the GeneMarkHMM-ES/BLAST training data set. EuGene was trained on the fungal plant pathogen *Microbotryum violaceum* and GeneID was trained on wheat. SN: sensitivity, SP: specificity. SN and SP calculated using Eval (Keibler and Brent, 2003). Gene and exon genomic overlaps represent where a predicted gene/exon overlaps a real gene/exon. The full exact gene column refers to genes with all exons and start/stop positions matching exactly. The exact exon column represents those real exons that are matched exactly by the predicted exons. Finally, Start/Stop refers to the exact start and stop positions of a gene within the genome.

Program	Genomic overlap: gene		Genomic overlap: single exon		Full exact gene		Exact exon		Start/Stop	
	SN	SP	SN	SP	SN	SP	SN	SP	SN	SP
GeneMarkHMM-ES	45.55	61.62	40.23	56.52	2.12	1.80	27.47	22.80	19.00	16.11
GeneMark with BLAST filtering	66.80	76.61	63.05	100.0	0.00	0.00	10.27	4.55	15.88	25.83
AUGUSTUS	88.58	96.76	86.84	94.73	6.93	8.02	80.57	88.36	58.09	67.21
GlimmerHMM	92.22	85.77	92.88	70.83	7.48	5.96	83.08	71.32	57.58	45.93
EuGene	80.78	96.96	79.49	91.76	4.25	4.73	72.88	84.86	31.05	34.52
GeneID	55.16	98.34	57.68	70.00	0.03	0.41	17.08	89.44	0.05	0.83
EvidenceModeler	82.04	97.06	86.50	93.46	7.14	7.37	74.08	82.99	51.16	52.78

BLAST filtering of the GeneMark genes greatly improved the proportion of true and accurate overlaps (Table 3.1). Reducing the data set in this way resulted in a reduction of full exact genes and exact exons, as only matching segments were considered, therefore genes were shortened. Sensitivity was consistently decreased due to the reduction in the amount of data. The gene prediction programs AUGUSTUS and GlimmerHMM, trained by the pipeline consistently produced more sensitive results in all tests than the more distantly trained EuGene and GeneID. Specificity was usually higher in genes produced by AUGUSTUS and of a similar level to other programs in genes produced by GlimmerHMM. Both programs trained by the pipeline were more sensitive and specific when generating full exact genes than any other programs. Combining

results with EvidenceModeler produced a gene set with a more even sensitivity and specificity. Weighting the results towards programs with either a higher sensitivity or specificity would decrease the other value, as seen in the other program results. The weighting method used in this test resulted in one of the highest full exact gene sensitivities and specificities, beaten by AUGUSTUS on specificity not sensitivity and by GlimmerHMM on sensitivity not specificity.

AUGUSTUS and GlimmerHMM were trained on the results of EvidenceModeler and re-run. The weighting of these programs was then increased by 1 to reflect the improved accuracy of the training data. EvidenceModeler was run a second time replacing the original AUGUSTUS and GlimmerHMM predictions with the new predictions (Table 3.2).

Table 3.2. Prediction accuracy for the second run of the pipeline FRANC. AUGUSTUS and GlimmerHMM were trained on the results from the previous EvidenceModeler predictions. SN: sensitivity, SP: specificity. SN and SP calculated using Eval (Keibler and Brent, 2003). Column headings are the same as in Table 3.1.

Program	**Genomic overlap: gene**		**Genomic overlap: single exon**		**Full exact gene**		**Exact exon**		**Start/Stop**	
	SN	**SP**	**SN**	**SP**	**SN**	**SP**	**SN**	**SP**	**SN**	**SP**
AUGUSTUS	88.57	96.36	86.61	92.49	8.09	8.73	82.56	88.27	67.10	72.45
GlimmerHMM	93.36	88.99	93.16	78.92	5.93	4.96	78.08	67.97	41.19	34.43
EvidenceModeler	80.08	97.19	85.25	93.74	6.94	7.15	71.08	79.65	47.07	48.49

Re-training AUGUSTUS on the pipeline output resulted in an increase in the proportion of full exact genes correctly identified and start and stop sites correctly identified (Table 3.2). There was some slight variation in the other categories, but these remained largely similar. GlimmerHMM

showed the opposite trend, where the percentage of real overlapping genes and exons was increased but with a slight decrease in the proportion of correct full exact genes, exons and start and stop sites. Re-running EvidenceModeler performed in a similar way to the previous run of the pipeline; by evening out the results. This run produced a gene set with slightly higher specificity of genomic and exonic overlaps than the previous run, but with slightly lower values in the previous categories.

Overall, iterative training of AUGUSTUS and GlimmerHMM on the various program output was affective at increasing the accuracy of these programs. Running EvidenceModeler produced similar results both times, and though the trained programs might have produced more accurate gene sets, this was offset by the less accurate programs. This therefore demonstrates that combining evidence can only go so far to mitigate the worst predictions, therefore it is recommended that weights of the programs expected to be the least accurate should be decreased after the first pipeline run, or alternatively, using these programs for the initial stages only.

For a comparison using untrained programs only, AUGUSTUS and GlimmerHMM trained on more distant species were used to predict genes in the *Arabidopsis* genome. AUGUSTUS was trained on maize and GlimmerHMM on rice. Training organisms were chosen from those available within the same kingdom. Once again, EvidenceModeler was used to combine predictions. Program results were set intuitively as in the first run; however, AUGUSTUS and GlimmerHMM were given the same weight as the other distantly trained programs. The results of these predictions are provided in Table 3.3.

Table 3.3. Prediction accuracy for the third run of the pipeline FRANC. AUGUSTUS was trained on maize and GlimmerHMM on rice. SN: sensitivity, SP: specificity. SN and SP calculated using Eval (Keibler and Brent, 2003). Column headings are the same as in Table 3.1.

Program	Genomic overlap: gene		Genomic overlap: single exon		Full exact gene		Exact exon		Start/Stop	
	SN	SP	SN	SP	SN	SP	SN	SP	SN	SP
AUGUSTUS	47.31	98.60	31.92	94.56	0.21	0.70	34.13	83.40	5.77	18.97
GlimmerHMM	89.26	90.19	84.63	85.99	3.11	3.32	70.20	71.19	31.24	33.36
EvidenceModeler	42.59	98.65	51.02	97.36	2.47	5.41	34.55	78.65	19.11	41.94

The use of training data from distant relatives proved to be far less effective than the self-training pipeline. Although the specificity of overlapping genes and exons was at a similar level in AUGUSTUS and GlimmerHMM as when they were trained by the pipeline there was an extremely large reduction in the sensitivity across all categories tested. AUGUSTUS in particular performed extremely poorly on identifying full exact genes and start and stop sites. Results combined by EvidenceModeler were as expected: sensitivity was extremely low across all categories, and specificity of overlapping sections was acceptable, but low for full exact genes, exons and start and stop positions.

3.4 DISCUSSION AND CONCLUSIONS

Testing the pipeline on Chromosome 1 of the *Arabidopsis* genome revealed that the self-training method put forward here is far more effective than using distantly related training data for gene predictions. Testing also revealed that using the pipeline rather than individual programs generally produced a consensus gene set that was similar to the most effective program in the pipeline, often with a higher specificity. This is extremely important, as when annotating a novel genome with little known data, it will not be possible to determine which program might be producing the best results. Despite this though, results were hampered by programs that produced less accurate results, suggesting that adding more data to the predictions did not necessarily improve results. EvidenceModeler was effective in the initial pipeline run, but for subsequent runs, the numbers of programs combined should be restricted to the most reliable programs; unless there is reason to believe that all programs should perform similarly well. One of the main benefits of FRANC will be the ease of which new software and software updates can be incorporated and the number of gene prediction programs that can be run. Given this, and the results from the testing, it is recommended that the programs most appropriate, or with the most closely related training data for the organism to be annotated should be used for the second run of the pipeline, and the first run to be performed as in the testing. It is also recommended to use the pipeline output to train the gene prediction programs wherever this is possible. Where known protein or EST data in available, the accuracy of prediction programs can be approximated by comparing the numbers of proteins or ESTs that match the genes identified (using the single best match only). This method is similar to identifying known gene and exon overlaps, as performed in the testing stage performed in this chapter.

FRANC is a Linux-based pipeline that was designed predominately for novel genomes without training data and with minimal gene finding hints (known proteins, ESTs). The individual programs within the pipeline have all been tested and verified by the makers, both with and without training data. The pipeline is intended for projects with minimal bioinformatics support, and requires no complex back end databases or programming skills. The pipeline presented is an extremely flexible tool, and can run with any number and combination of the programs available. A suggested way to use FRANC (in the absence of training data) is to mask the sequences using LTRharvest and perform a GeneMark run on the data. These predictions can then be run through BLAST to find protein homology. Known proteins and ESTs can be aligned onto the genome and the resultant output used to train the prediction programs. The final output of EvidenceModeler after running the pipeline can also be used to train the prediction software. The programs can be weighted by the combiner, for sensitivity, for specificity or for a balance of the two. Generally, specificity is increased with reliable training data, and sensitivity with closely-related training data. MAKER is a gene prediction pipeline similar to FRANC, combining RepeatMasker, SNAP/GMAP, BLAST and Exonerate to collect information on the most likely gene predictions. MAKER was designed for emerging model organisms, and becomes less reliable in organisms where training information is absent (Cantarel, et al., 2008). FRANC was created from the need to integrate a larger number of gene prediction programs, combined with a prediction combiner, for use on non-model organisms, where training data is scarce.

3.5 REFERENCES

Allen, J.E. and Salzberg, S.L. (2005) JIGSAW: integration of multiple sources of evidence for gene prediction, *Bioinformatics*, **21**, 3596-3603.

Benson, G. (1999) Tandem repeats finder: a program to analyze DNA sequences, *Nucleic Acids Research*, **27**, 573-580.

Birney, E., Clamp, M. and Durbin, R. (2004) GeneWise and GenomeWise, *Genome Research*, **14**, 988-995.

Blanco, E., Parra, G. and Guigó, R. (2002) Using GeneID to identify genes. In, *Current Protocols in Bioinformatics*. John Wiley & Sons, Inc.

Brejová, B., *et al.* (2005) ExonHunter: A comprehensive approach to gene finding, *Bioinformatics*, **21**, i57-i65.

Brunak, S., Engelbrecht, J. and Knudsen, S. (1991) Prediction of human mRNA donor and acceptor sites from the DNA sequence, *Journal of Molecular Biology*, **220**, 49-65.

Cantarel, B.L., *et al.* (2008) MAKER: An easy-to-use annotation pipeline designed for emerging model organism genomes, *Genome Research*, **18**, 188-196.

Coghlan, A., *et al.* (2008) nGASP — the nematode genome annotation assessment project, *BMC Bioinformatics*, **9**, 549.

Curwen, V., *et al.* (2004) The Ensembl automatic gene annotation system, *Genome Research*, **14**, 942-950.

Ellinghaus, D., Kurtz, S. and Willhoeft, U. (2008) LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons, *BMC Bioinformatics*, **9**, 18.

Elsik, C., *et al.* (2007) Creating a honey bee consensus gene set, *Genome Biology*, **8**, R13.

Guigó, R., *et al.* (1992) Prediction of gene structure, *Journal of Molecular Biology*, **226**, 141-157.

Haas, B., *et al.* (2008) Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments, *Genome Biology*, **9**, R7.

Havecker, E., Gao, X. and Voytas, D. (2004) The diversity of LTR retrotransposons, *Genome Biology*, **5**, 225.

Hebsgaard, S.M., *et al.* (1996) Splice site prediction in Arabidopsis thaliana pre-mRNA by combining local and global sequence information, *Nucleic Acids Research*, **24**, 3439-3452.

Keibler, E. and Brent, M. (2003) Eval: A software package for analysis of genome annotations, *BMC Bioinformatics*, **4**, 50.

Korf, I., *et al.* (2001) Integrating genomic homology into gene structure prediction, *Bioinformatics*, **17**, S140-S148.

Majoros, W.H., Pertea, M. and Salzberg, S.L. (2004) TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders, *Bioinformatics*, **20**, 2878-2879.

Mathé, C., *et al.* (2002) Current methods of gene prediction, their strengths and weaknesses, *Nucleic Acids Research*, **30**, 4103-4117.

Newberg, L.A., *et al.* (2007) A phylogenetic Gibbs sampler that yields centroid solutions for cis-regulatory site prediction, *Bioinformatics*, **23**, 1718-1727.

Pedersen, A.G. and Nielsen, H. (1997) Neural network prediction of translation initiation sites in eukaryotes: Perspectives for EST and genome analysis, *Proc Int Conf Intell Syst Mol Biol*, **5**, 226-233.

Reese, M.G., *et al.* (2000) Genie — gene finding in Drosophila melanogaster, *Genome Research*, **10**, 529-538.

Salamov, A.A. and Solovyev, V.V. (2000) Ab initio gene finding in Drosophila genomic DNA, *Genome Research*, **10**, 516-522.

Schiex, T., Moisan, A. and Rouz, P. (2001) EuGene: An Eucaryotic gene finder that combines several sources of evidence. , *Computational Biology*, 111-125.

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

Sleator, R.D. (2010) An overview of the current status of eukaryote gene prediction strategies, *Gene*, **461**, 1-4.

Smit, A.F.A., Hubley, R. and Green, P. (1996-2010) RepeatMasker Open-3.0, *www.repeatmasker.org*.

Smith, T.F., Waterman, M.S. and Fitch, W.M. (1981) Comparative biosequence metrics, *Journal of Molecular Evolution*, **18**, 33-46.

Stanke, M. and Waack, S. (2003) Gene prediction with a hidden Markov model and a new intron submodel, *Bioinformatics*, **19**, ii215-ii225.

Ter-Hovhannisyan, V., *et al.* (2008) Gene prediction in novel fungal genomes using an ab initio algorithm with unsupervised training, *Genome Research*, **18**, 1979-1990.

Vitte, C. and Panaud, O. (2005) LTR retrotransposons and flowering plant genome size: Emergence of the increase/decrease model, *Cytogenetic and Genome Research*, **110**, 91-107.

Wu, T.D. and Watanabe, C.K. (2005) GMAP: A genomic mapping and alignment program for mRNA and EST sequences, *Bioinformatics*, **21**, 1859-1875.

4. ANNOTATING SEVEN GENOMES OF *M. VIOLACEUM*

KATE E. RIDOUT

4.1 INTRODUCTION

Annotating a genome is a complex task, requiring the identification of gene structures, both protein-coding and not, repetitive elements, regulatory regions and other aspects of coding and non-coding. Here, I discuss eukaryotic protein coding gene identification and repetitive element prediction. The simplified structure of a eukaryotic gene is presented in Figure 4.1; prediction methods determine the exons and introns of a gene as standard. Some software will additionally detect the untranslated terminal regions (UTRs) at either end of the gene.

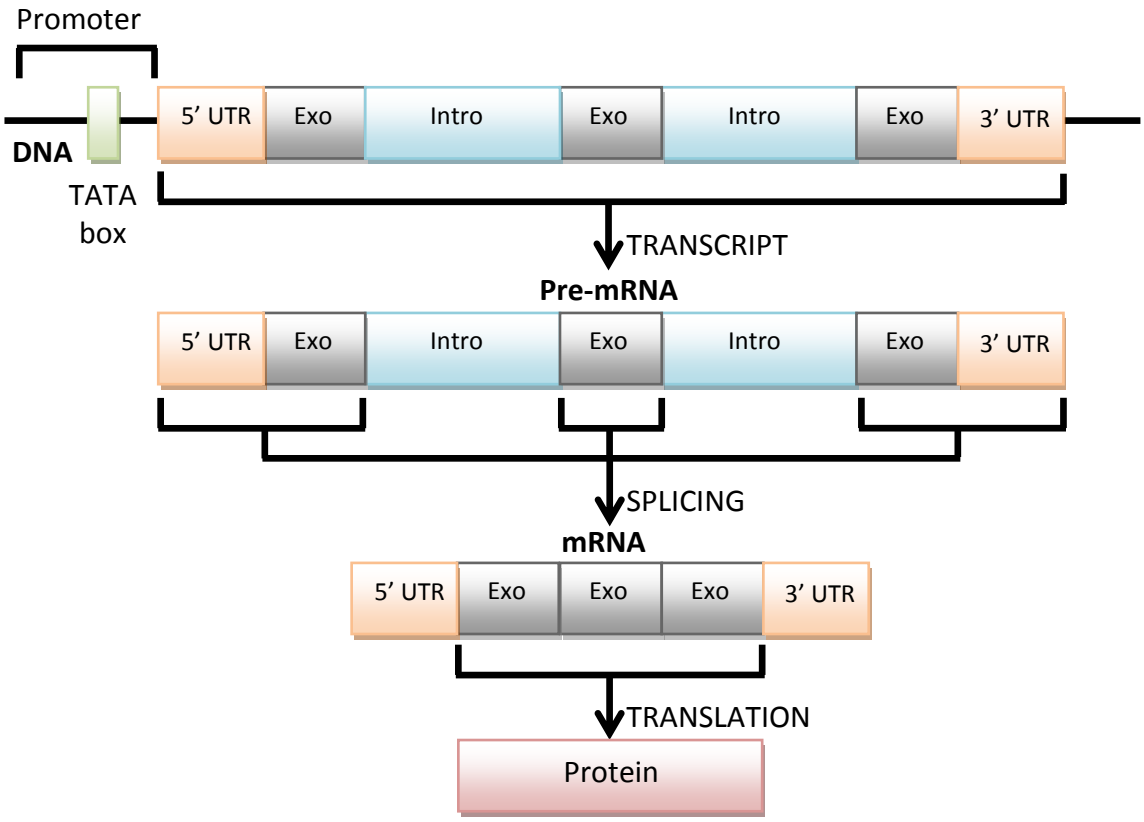

Figure 4.1. The simplified structure of a gene from DNA to protein.

Traditionally, gene prediction software are concerned with predicting transcribed regions only, and will not detect *cis*-regulatory elements (Mathé, et al., 2002). Gene prediction can be used to identify RNA genes, which are not translated into proteins but still may have an important biological role. RNA gene prediction is not carried out here, and is therefore not discussed; for a review see Meyer, (2007).

There are currently two approaches to predicting gene structure, depending on the volume and type of data available for a given species; these methods are *ab initio* and homology–based prediction. *Ab initio* prediction methods can be used when a number of known genes with a high confidence are available; either for the species in question or a closely related species. The positions of these genes within the genome and their exact gene component locations, usually including UTRs, introns and exons, must be known in detail and with a high certainty. This gene set is known as a training set, and is used to train the probabilistic models required by *ab initio* programs for gene prediction. Training sets can range from a few hundred to a few thousand genes. Generally, the more genes used for training, the better the result. Homology or empirical prediction methods can be used when a training set does not exist for an organism. This approach relies upon homologous known proteins or genomes existing in the available databases. New proteins are predicted based on protein homology and potentially genome conservation. For a review of gene prediction methods see (Sleator, 2010).

Gene prediction methods are largely based on two types of sensor: signal sensors and content sensors. Signal sensors detect functional regions of a gene, such as sites relating to transcription, translation and splicing. Transcription and translation signals include start and stop codons, the TATA box (see Figure 4.1) and transcription factor binding sites. Splicing signals are extremely

important for gene prediction, particularly in eukaryotic organisms where the number of introns and exons comprising a gene is large. Splicing signals include donor sites, acceptor sites and branch points. Content sensors are more specific to genes within a given species, and can be extrinsic or intrinsic. Extrinsic content sensors perform sequence alignments of homologous genes, both from different organisms and within the same organism to find protein families. These homology searches are based on the pattern matching problem, and usually involve local sequence alignments using approaches such as the Smith-Waterman method used by Crossmatch (http://www.phrap.org/phredphrapconsed.html). Intrinsic content sensors directly examine the properties of the DNA in coding and non coding regions, and identify coding regions based on base content, codon content and hexamer frequency (Sleator, 2010). For example, intronic regions tend to be more A/T rich than exons (Mathé, et al., 2002), protein coding regions display codon bias and hexamer frequency differs between introns and exons (Fickett and Tung, 1992).

Both homology–based and *ab initio* gene prediction methods have the potential to incorporate homology evidence from alternative sources, such as EST/cDNA information and whole genome sequences. Programs such as Twinscan (Korf, et al., 2001) are able to use related genomes to identify regions of conserved sequence that might be coding, based on the assumption that coding regions should be more conserved (Mathé, et al., 2002). *Ab initio* gene prediction methods rely upon both signal and content sensors, therefore often including the homology approach into the method. For this reason the *ab initio* gene prediction methods might be considered more reliable than simple homology approaches. Homology–based gene prediction methods combine putative protein identification from homologues with signal sensors to better refine the boundaries of the genes. The more homologues in the database and the more closely related they are, the better the predictions will be. Despite these two distinct types of predictors, in reality most modern programs will attempt to combine as many sources of information as possible.

Once information on coding sequences has been collected by signal and content sensors, it can be used to create types of models of the expected appearance of a protein coding sequence. The most commonly used type of model for this process is the Markov model. Here, parameters gained from training sequences provide information on the expected content of a coding sequence. The probability that a nucleotide appears at a given location will then depend on the previous *n* nucleotide states. The coding sequence is evaluated by the model and the probability that the sequence fits the model i.e. is coding, can be calculated. Any number of Markov models can be used to evaluate a sequence, however, coding sequences might usually require three; one for each position in the codon (Mathé, et al., 2002). The higher the value of *n* the more complex the Markov model and the more nucleotide dependencies they can represent. Increasing *n* therefore can create Markov models that better represent the data, but require a greater volume of training data (Mathé, et al., 2002). Gene predictors may use these models in a simpler context; representing coding and non-coding only. Alternatively, some software, for example EuGene (Schiex, et al., 2001) will produce different models for different types of non-coding sequence. Alternatively, position weight matrixes and neural networks can utilise sequence alignments to calculate the probability of a base appearing at a given site (Mathé, et al., 2002). Most gene prediction software will use dynamic programming to combine the various components of the gene into a single gene model.

Gene predictions will never be 100% reliable, and are generally evaluated on their sensitivity – the number of correct genes that they are able to detect, and specificity – how often the gene boundaries, and intron/exon sites are correctly predicted. Prediction of genes through protein homology has a number of drawbacks; the most obvious being that predictions are limited to those genes already in the database. Poor, incorrect and absent gene annotations can provide misleading results (Salzberg, 2007). Homology approaches are limited to finding highly

conserved regions only, and are likely to struggle with gene and exon/intron boundary discovery, particularly as the genetic distance between target and training species increases. EST and cDNA homology can help with these issues; UTRs (untranslated terminal regions at both ends of a gene (Figure 4.1)) are included in the transcript and information can be collected on alternative splicing (Birney, et al., 2004). Unfortunately, correctly identifying the true genomic sequence of an EST with the existence of protein homologues is not trivial; particularly as pattern matching software tend to be heuristic. Using genome homology for gene prediction can also be a difficult task, as conserved regions may extend beyond the exons, rapidly evolving genes can be too divergent (depending on threshold levels for the identification of 'conserved' regions) and little indication is given of the exact gene and exon/intron boundaries (Birney, et al., 2004; Mathé, et al., 2002). *Ab initio* prediction methods present two key problems. The first is the need for training data, effectively ruling out this method of prediction for novel genomes with no close relative. Several ways around this problem have been suggested, such as self iterative training, based on common gene signatures (Ter-Hovhannisyan, et al., 2008) and using protein homology with highly conserved genes to create the training set (Parra, et al., 2007). The second problem is that only genes that follow the expected composition will be detected. Fortunately, this problem will diminish as more proteins are identified.

Prediction software can also be used to identify repetitive genomic elements, which are commonly divided into two types; interspersed repeats and tandem arrays. The tandem array elements comprise adjacent duplicate sequences, or approximate sequences throughout the genome. They are usually identified using a variation of the pattern matching problem, and programs such as Crossmatch, which uses the Smith-Waterman approach. Interspersed repeats are more complex genomic components which have been duplicated throughout the genome and include transposable elements, which are able to duplicate themselves. Transposable elements

(TEs) usually fall into one of three classes: LINE/SINE elements, LTR-retrotransposons and DNA-transposons. TEs can be identified through their homology to known TEs using programs such as RepeatMasker (Smit, et al., 1996-2010), in a similar way to the homology–based gene finders. This approach will also identify partial TEs, although caution must be taken when predicting genes and repeat elements as some genes might have evolved from historic duplications that could be masked by this type of software. It is also possible to identify TEs *de novo*, as many have regular structures that can be modelled in a similar way to the gene prediction software (Ellinghaus, et al., 2008). Unfortunately as with gene predictions, the reliability of repeat prediction software will depend upon TEs that have already been characterised, thus restricting the approach, as even modelling software will base the search on an expected structure.

Unfortunately, prediction software for genome annotations do not produce perfect results, particularly in complex intron–containing eukaryotic genes, where splicing control is not fully understood. One way of increasing the number of reliable usable predictions is to combine the different analysis methods, by using both homology based and *de novo* methods to predict genes and repetitive elements, and a number of programs have been developed to perform this task (Allen and Salzberg, 2005; Cantarel, et al., 2008; Haas, et al., 2008). The reliability of these methods will depend on individual predictions from the programs to be combined, and therefore the more supporting information (cDNAs, homologous proteins, homologous genomes) and training data available, the more reliable the predictions.

4.1.1 ANNOTATING THE GENOME OF *MICROBOTRYUM*

In Chapter 2, seven different species of *Microbotryum* were assembled. In annotating these genomes, it becomes possible to study the genes and repeat content of *Microbotryum* and compare it to other sequenced fungi. Annotating all seven genomes will also help to resolve the *Microbotryum violaceum* phylogeny, and to perform evolutionary genetics analysis. The closest annotated relative to *Microbotryum violaceum* with a full training set at the time of writing is *Ustilago maydis*. Although *Microbotryum* was once thought to be a member of the genus *Ustilago*, it has been re-classified into the class *Microbotryomycetes*, which are more closely related to rust fungi (as demonstrated below using BLAST results). Additionally, the *Ustilago* training set has only been prepared for a small number of prediction programs. Thus, in order to predict genes in the *Microbotryum* genome the sequence annotation framework FRANC was developed, as discussed in the previous Chapter (Chapter 3). FRANC provides a range of repeat and gene prediction methods that can be combined using a maximum likelihood approach taken by the program EvidenceModeler (Haas, et al., 2008). Prediction methods can be iteratively trained on FRANC output, and combined with known protein and repeat information, protein and repeat homology and EST evidence. This has enabled the predictions of genes in *Microbotryum* from extremely divergent training data, and in some cases no training data by combining multiple sources of information.

Here, I describe the optimisation of FRANC to predict genes in the *M. violaceum* var. *latifolia* genome. Genes in other *Microbotryum* species have been predicted using a comparative analysis with the FRANC gene set as well as independently predicted using the gene prediction program AUGUSTUS (Stanke and Waack, 2003) with the training set produced by FRANC. Genes have been computationally verified wherever possible. In total, 9,919 genes have been identified in *M.*

violaceum var. *latifolia*, in which 91% of bases could be validated using EST data, and 2,004 genes were detected across seven *Microbotryum* species. The genes identified appear to have experienced a pattern of intron loss similar to the smut fungus *Ustilago maydis* (Kamper, et al., 2006). The transposable element content of the genome was estimated at 1.65% of the genome, which is low for both fungal and parasitic species. When compared to itself, over 8% of the genome was found to be repeated, therefore it is possible that these areas might contain unclassified, novel transposable elements. The expected number of genes in *Microbotryum* has been estimated at ~10,000 genes. Of the 9,919 genes identified, 68% could be annotated using homology searching, and over 90% of all genes matched a known sequence motif, suggesting that novel genes and transposable elements might be contained in the data set and would warrant further investigation.

4.2 METHODS

The purpose of the methods described below is to produce a set of genes for evolutionary analysis. With this in mind, all methods used to collect the gene set are extremely conservative, and could be extended in a number of ways to expand the final data set. Here, predicted exons are preferentially selected over predicted whole genes if the gene boundaries are not reliable, resulting in potentially shortened but highly accurate genes.

4.2.1 DATA

The genome from a single *M. violaceum* var. *latifolia* individual detailed in Chapter 2 was used for the initial gene predictions. The genome was selected on the basis of highest Illumina

coverage (44 million paired end reads, while for other genomes only 18M to 20M reads were available) and better quality de novo assembly (563a2 Illumina genome assembled using Velvet). All other *Microbotryum* genomes were mapped to this reference genome. EST data was obtained for *M. violaceum* var. *latifolia*, and is detailed in Chapter 2. All unique proteins (68 in total) experimentally determined in *Microbotryum violaceum* at the time of annotation were downloaded from NCBI GenBank (http://www.ncbi.nlm.nih.gov/). Six additional *Microbotryum* genomes were assembled using the mapping approach presented in Chapter 2. These species were collected from different plant species: *Dianthus carthusianorum, Lychnis flos-cuculi, Saponaria officinalis, Silene caroliniana, Silene dioica* and *Silene vulgaris*. All strains used in the analysis belonged to the same mating type (A2).

4.2.2 FRANC

The gene prediction/annotation pipeline FRANC is able to run and combine the results of the software LTRharvest (Ellinghaus, et al., 2008), RepeatMasker (Smit, et al., 1996-2010), Tandem Repeats Finder (Benson, 1999), Crossmatch, BLAST, Exonerate (Slater and Birney, 2005), Genewise (Birney, et al., 2004), GMAP (Wu and Watanabe, 2005), GeneMark (Ter-Hovhannisyan, et al., 2008), AUGUSTUS (Stanke and Waack, 2003), GlimmerHMM (Majoros, et al., 2004), GeneID (Blanco, et al., 2002), EuGene (Schiex, et al., 2001) and the combiner EvidenceModeler (Haas, et al., 2008).

Pipeline methodology

A diagram of the workflow used for the predictions is shown in Figure 4.2. In all programs both the forward and reverse strands of the genome were considered. Only complete genes were predicted to reduce the number of false positives.

Repeat sequences were predicted in the *Microbotryum violaceum* genome using FRANC to run LTRharvest, RepeatMasker, TRF and Crossmatch. For the program LTRharvest, LTRs were considered to be longer than 100bp in length and shorter than 800bp, consistent with suggested fungal LTR lengths (Goodwin and Poulter, 2000; Lauermann, et al., 1997). RepeatMasker was given three libraries for repeat homology identification; the LTRs predicted using LTRharvest and repeats characterised as 'fungi' and 'microbotryum' from the RepeatMasker database (a total of 624 sequences and 866,625bp of repeat data). The genome was masked by LTRharvest only, to avoid masking simple and low complexity repeats, or historic transposable elements that might comprise a gene. Unless otherwise specified, all further steps were performed on the masked genome.

The self-training program GeneMarkHMM-ES, designed for fungi (Ter-Hovhannisyan, et al., 2008) predicted genes, which were then sent to BLASTp to be compared to the NCBI's (http://www.ncbi.nlm.nih.gov/) 'nr' non-redundant protein database. FRANC classified a stretch of DNA as a gene if it fulfilled certain criteria. The top 10 BLAST hits were considered, and for FRANC to identify a gene it must have an identity of ≥50% in at least one of the top 10 genes. Additionally, ≥90% of the gene must be included in at least one of the matches. Genes that satisfied these criteria were added to an initial training set.

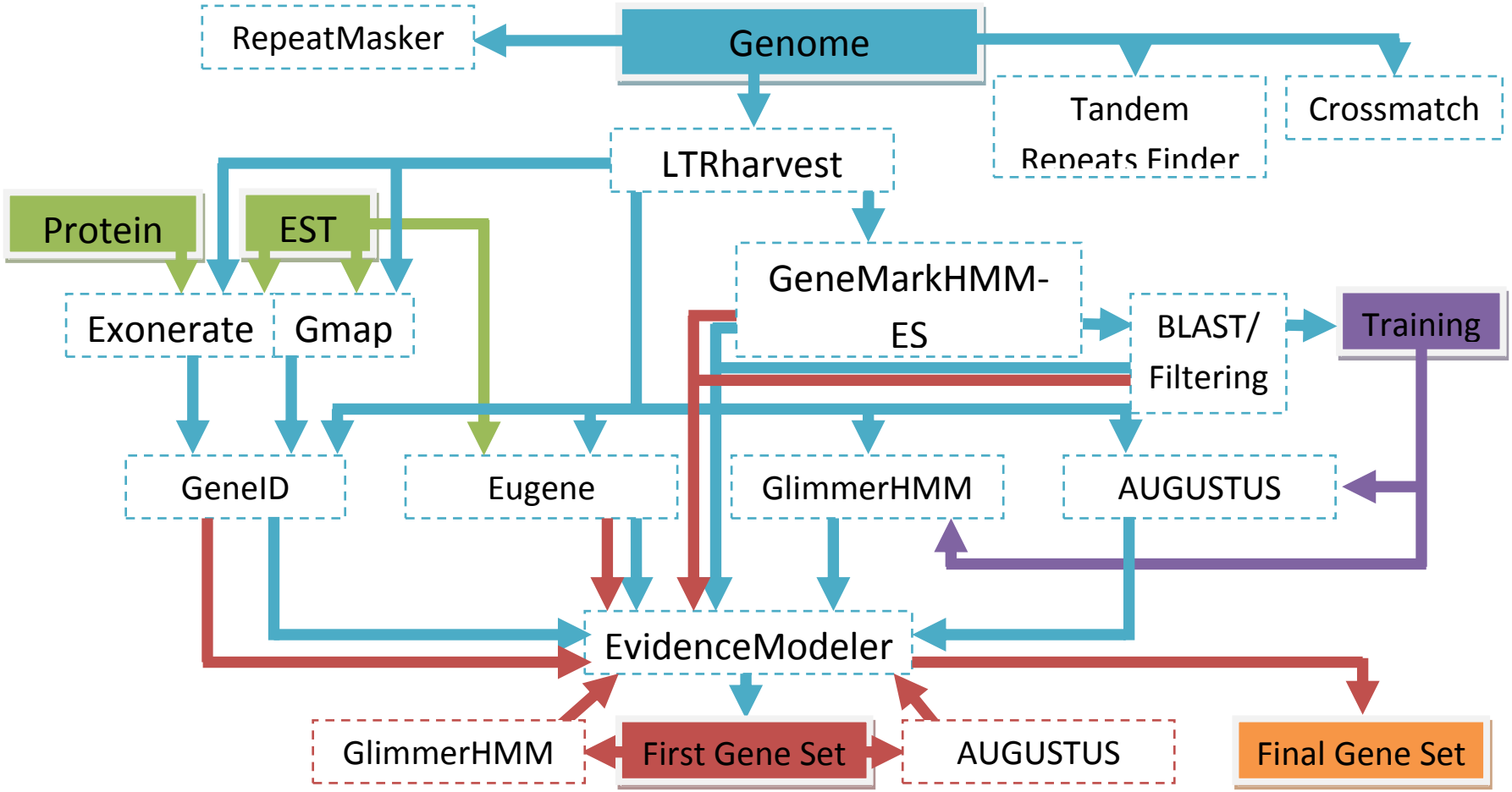

Figure 4.2. The gene prediction process: Sequence files are represented by block colours and programs with dashed lines. The blue arrows represent the initial pipeline run. The second pipeline run is coloured in red. Additional sources of evidence (known proteins and ESTs) are green. The training data created by the pipeline and used to train AUGUSTUS and GlimmerHMM is coloured purple.

EST data and known proteins from *Microbotryum violaceum* were aligned to the genome. ESTs were aligned using the programs Exonerate and GMAP, and proteins using Exonerate only. For Exonerate, both proteins and ESTs were recorded if the match was over 90% homologous. Introns were considered to be between 20 bp and 3000 bp in length, the expected range of fungal introns (Kupfer, et al., 2004). For the same reason, the maximum intron size of GMAP was also set to 3000. These alignments were added into the training set.

All data from the training set were used to train GlimmerHMM, following the procedure set out in the program manual. The same data was also used to train AUGUSTUS, using the automated perl training script AutoAug.pl. Once trained, both gene prediction programs were run on the masked genome.

EuGene and GeneID, could not be trained using the *M. violaceum* var. *latifolia* data, as both require complex and detailed training information that was not available. Both programs accepted EST data to aide predictions. EuGene was used with the default parameter set trained on *Arabidopsis thaliana*, and GeneID was trained on wheat. Both programs require complex training, and so existing training sets were used. Results from the previous chapter demonstrated that including results from these programs when distantly trained does not decrease the overall sensitivity and specificity of the pipeline. GeneID also included the known protein data; all supporting data was provided as alignments from GMAP and Exonerate. Alternative splicing was not predicted due to the expected small intron content of *Microbotrym* and the complications that this approach would introduce.

EvidenceModeler was run on the unmasked genome, combining predictions from LTRharvest, GeneMarkHMM-ES, GeneMark/BLAST, Exonerate, GMAP, GlimmerHMM, AUGUSTUS, GeneID and EuGene. The weights file was set as follows: All predictions from the above steps were aligned to the EST data using tBLASTN. Results were assessed based on the number of genes containing in-frame stop codons, the number of genes completely supported by ESTs, the number of predicted bases supported by EST evidence, and the number of broken genes, chimeric genes and partially supported genes. Where a single EST comprised more than one predicted gene in non-overlapping sections, the genes were considered to be 'broken' i.e. split into sections. Alternative splicing might also cause this result, however *Microbotryum* does not have a large concentration of introns (see results), and so this is not expected to create a large problem. Genes that comprised several ESTs joined together were considered chimeric, and genes where EST evidence supported less than the whole gene were labelled as partially supported. It is worth noting that the ESTs might not all represent complete genes, however results are compared to one another using the same EST set, and thus will not cause a bias. The weights file for EvidenceModeler was set to reflect the results of this assessment and is shown below:

```
ABINITIO_PREDICTION   LTRharvest             1
ABINITIO_PREDICTION   GeneMarkHMM-ES_BLAST   1
ABINITIO_PREDICTION   GeneMarkHMM-ES         1
ABINITIO_PREDICTION   AUGUSTUS              2
ABINITIO_PREDICTION   GlimmerHMM            2
ABINITIO_PREDICTION   geneid_v1.3                 1
ABINITIO_PREDICTION   EuGene                      2
PROTEIN         exonerate_protein_alignments     5
TRANSCRIPT      gmap_transcript_alignments       4
TRANSCRIPT      exonerate_transcript_alignments  4
```

The results of the pipeline were used to train EuGene, based on the directions specified in the manual, GlimmerHMM and AUGUSTUS as before. These three programs were then re-run, and the output files assessed as before (using EST comparisons). The weights file was altered to

reflect the new information and EvidenceModeler was run a second time. The new weights file is shown below:

```
ABINITIO_PREDICTION   LTRharvest             1
ABINITIO_PREDICTION   GeneMarkHMM-ES_BLAST   1
ABINITIO_PREDICTION   GeneMarkHMM-ES         1
ABINITIO_PREDICTION   AUGUSTUS               3
ABINITIO_PREDICTION   GlimmerHMM            3
ABINITIO_PREDICTION   geneid_v1.3                 1
ABINITIO_PREDICTION   EuGene                      3
PROTEIN         exonerate_protein_alignments     5
TRANSCRIPT      gmap_transcript_alignments       4
TRANSCRIPT      exonerate_transcript_alignments  4
```

Genes were spliced as specified by the output of EvidenceModeler (alternative splicing was not considered) and those containing internal stop codons in the exons were removed from the data set, as these genes might represent non coding RNAs or pseudogenes.

4.2.3 EST ANALYSIS

As a proxy for the effectiveness of the gene predictions, the EST data were aligned to the final gene set using tBLASTN. Genes were examined for complete and partial EST support, and as before, broken and chimeric genes were identified.

4.2.4 GENE CHARACTERISATION

The complete set of genes identified through the pipeline above (excluding those containing in-frame stop codons) was entered into the software BLAST2GO (Conesa, et al., 2005; Götz, et al., 2008) for gene characterisation. A combination of BLAST homology searching, Gene Ontology

annotations, Inter Pro annotations and KEGG pathway mapping are used by this software for gene identification and annotation.

Mating-type-linked regions of the *Microbotryum* genome were identified and provided for this analysis (Filatov, unpublished). The identification of putatively mating-type-linked genes was carried out using tBLASTN to map the complete gene set onto areas of the genome identified as being mating-type-linked. Genes that mapped onto the mating-type-linked regions with a top-hit threshold value of ≥95% in the A2 linked regions and ≥90% in the A1 linked regions were considered to be sex mating-type-linked. The A1 linked threshold value was lowered because the *Microbotryum* genome examined was A2, thus allowing for divergence.

Transposable elements are known to contribute to both gene and genome evolution in eukaryotes (Jiang, et al., 2004; Wessler, 2006). Due to the small genome size of *Microbotryum* (~25 mb) the transposable element content is not expected to be large, but it is still possible that the genes predicted within *M. violaceum* var. *latifolia* might have evolved from transposable elements, for example, miRNA genes in pants and animals (Li, et al., 2011), and protein encoding RAG genes in humans (Fugmann, 2010). It is also possible that genes might actually be transposable elements that have not yet been characterised. To investigate these possibilities, RepeatMasker was used on the gene set using repeat libraries characterised from LTRharvest (*M. violaceum* var. *latifolia* specific), *Microbotryum* and fungi.

4.2.5 GENES IN OTHER SPECIES

Identifying the predicted genes in multiple species

One way of independently validating genes that have been computationally predicted is to identify these genes in closely related genomes. Genes where the purported coding regions contain stop codons in alternative species or cannot be identified are less likely to represent real genes. An incorrect gene in this instance may represent a pseudogene, a region that has simply appeared to be coding by chance, a gene predicted with exons in an incorrect frame, or a gene with misplaced intron, exon, start and/or end boundaries. Gene finding was intended to be conservative.

The six additional *Microbotryum* genomes were masked using LTRharvest. Masked genomes were converted into BLAST databases and each genome was searched with tBLASTn against the translated spliced proteins predicted using the FRANC pipeline from *M. violaceum* var. *latifolia*. BLAST extension options were optimised to encourage gaps and penalise low quality matches. This forced BLAST to break an alignment rather than try to extend it over an intron, which would result in the loss of the end of the sequence from the subject. BLAST searching was carried out using the BLOSUM90 matrix to discourage matches between regions of low homology that might arise at intron boundaries. The protein database was used as the query to preserve the reading frame. The top hit to each of the subject sequences was collected from each genome, taking the very start of the matching segments and the very end to retain introns. A relatively low e-value of 1×10^{-5} was chosen for the initial BLAST run, as introns often decreased the e-value where short regions of poor matches occurred at the boundaries. In all cases, softmasking was turned on in BLAST; this option masks repetitive sequence in lowercase rather than with “N” characters, so

that they may still be searched. This was necessary as some genes contain simple sequence repeats.

The program Exonerate was used to splice the resultant gene matches using the protein2genome model, where proteins were the original *M. violaceum* spliced proteins and the genomes were the gene matches from the multiple species identified using BLAST. This model identified introns through protein homology and intron modelling (based on expected minimum and maximum lengths plus an intron penalty). For genes from all genomes, softmasking (as used in the BLAST run) was turned on. Minimum and maximum intron lengths were set as seen previously (20bp to 3000bp). Genes were included in the output if the protein and "genome" had a match of at least 90%. Only matches for the same genes were considered for identifying intron positions.

Due to the heuristic nature of BLAST and closely related protein families, it is possible that the genes identified in the different species genomes might not be the best match of all the genes. To check for this possibility, the previous BLAST search was extended into a reciprocal best BLAST. Spliced genes identified through the initial BLAST run and spliced using Exonerate were aligned to a database of the original *M. violaceum* spliced genes using BLAST. Genes that had a top hit other than the expected original gene were removed from the analysis.

Pairwise gene alignments were carried out between each species and the FRANC gene set using BioPerl. Global sequence alignments were performed using the dpAlign module, which is based on the alignment program SSEARCH (Pearson, 1996), an implementation of the Smith-

Waterman algorithm. Alignments were carried out in order to translate the genes and find those containing in-frame stop codons. Sequencing errors causing frame-shifts are minimised this way.

Additional gene predictions in multiple species

In the previous stages, genes were not predicted independently in all species, thus rejecting genes that might be species–specific, or were not detected in the first gene prediction set. Genes that could be detected with a high certainty in all *Microbotryum* genomes analysed were considered to be extremely reliable. In total 2,004 of these genes were identified. These genes were considered to be more accurate than the total gene set from the FRANC gene predictions, therefore these data were used to re-train the gene prediction program AUGUSTUS as discussed previously. A hints file containing known locations of ESTs on the genome was created using the alignment tool BLAT (Kent, 2002), following the methods suggested in the AUGUSTUS documentation. Complete genes were then predicted in all species using AUGUSTUS with the new training parameters and the hints files generated.

4.2.6 ALL SPECIES GENE ALIGNMENTS

Genes that could be identified in all seven *Microbotryum* species were collected. In order to be sure that the genes were likely to be true homologues, only those genes identified in all species through the original FRANC predictions and refined using reciprocal best BLAST (as detailed above) were used. Multiple sequence alignments were carried out on all genes using ClustalW (Larkin, et al., 2007). Genes were translated *in silico*. Nucleotide sequences containing stop codons in one or more of the sequences were removed, as were sequences that were no longer

multiples of three base pairs; these sequences were likely to represent pseudogenes or assembly errors.

4.2.7 EVOLUTIONARY ANALYSIS

The 2,004 genes were concatenated in all species. A Bayesian approach was taken using MrBayes 3.1.2 (Huelsenbeck and Ronquist, 2001; Ronquist and Huelsenbeck, 2003) to reconstruct the phylogeny for the concatenated gene set, using the GTR+Γ model of DNA evolution. Codeml from the PAML package was used to identify genes under positive selection using a likelihood ratio test (LRT) based on non-synonymous and synonymous divergence (dN/dS = ω). Codeml nested models M7 and M8 were selected for comparison; model M8 differs from M7 only in allowing a class for $\omega > 1$, i.e. positive selection. The LRT is then used to test whether the data better fits the model with or without positive selection. Genes with a positive LRT are considered to be under positive selection. Determining positive selection in this way is only effective if the divergence of the species is high enough to ensure that there is enough power and low enough that dS does not become saturated. Anisimova and Yang (2007) examined the reliability of tests for positive selection at different divergences. Average pairwise divergence of the genes was calculated in MEGA version 5 (Tamura, et al., 2011), with a Poisson model using the concatenated supergene.

4.3 RESULTS

4.3.1 FRANC IDENTIFICATIONS

Repetitive elements

LTRharvest identified a total of 18 regions of the genome that were predicted as full-length LTR-retrotransposons, spanning 17 scaffolds. The lengths ranged from 1,310 bp to 12,354 bp long with a total length of 74,143 bp (0.306% of the genome).

RepeatMasker was able to mask 513,536 bases of various repeat sequences – a total of 2.12% of the genome. The results are presented in Table 4.1. Interspersed repeats included partial LTR-retrotransposons identified through homology searches with those sequences identified using LTRharvest. The 2 DNA elements in Table 4.1 were identified through homology searches and represent only a very small fraction of each element. The first is 85 bp of the *Crypt1* transposon (3,563 bp) from the *hAT* family of transposable elements (Linder-Basso, et al., 2001). The second is 42 bp of the *Mariner-9* DNA transposon (2,125 bp) from the *TcMar-Fot1* subfamily (Galagan, et al., 2005).

Table 4.1. RepeatMasker sequences identified in the *M. violaceum* var. *latifolia* genome.

Element	Number of elements	Total length of elements (bp)	Percentage of the genome
LTR elements	110	22,149	0.09
DNA elements	2	127	0.00
Unclassified interspersed repeats	1,705	317,218	1.31
Small RNA	28	6,138	0.03
Simple Repeat	2,222	95,420	0.39
Low complexity	1,148	72,484	0.30
Totals	**5,215**	**513,536**	**2.12**

Tandem Repeat Finder identified and masked 7,337 simple repeats, for a total of 304,714 bp, which were detected in 1,477/20,557 scaffolds. The masked repeats represented a total of 1.26% of the genome, in addition to the 2.12% identified using RepeatMasker. Crossmatch was the final program to be used and masked 8.80% of the genome. In total, between all of the programs 3,027,119 bp was masked, representing 12.48% of the genome.

Creation of program weights through EST mapping

Predicted genes were extracted from the various program stages of the pipeline runs. For the initial run, results from GeneMarkHMM-ES, GeneMarkHMM-ES BLAST, GlimmerHMM, GeneID, EuGene, AUGUSTUS, GMAP, Exonerate EST and Exonerate protein were examined. A total of 9,131 unique ESTs were mapped to the genes using tBLASTN. Results are shown in Table 4.2. The weighting for the first run of EvidenceModeler was set intuitively, as suggested in the EvidenceModeler documentation. GlimmerHMM, EuGene and AUGUSTUS were given a higher weight than the other prediction software due to their high number of whole gene matches. Transcript alignments were set higher than the gene predictions and the small set of protein alignments were given the highest weight.

After the initial run, AUGUSTUS and GlimmerHMM were re-trained on the genes called using EvidenceModeler; both programs were then run a second time. The EST analyses were performed again on the predictions from these two programs and are presented in Table 4.3. The weighting of these two programs was increased by 1, to reflect the increased accuracy.

Table 4.2. EST support for genes predicted during the initial pipeline run. AUGUSTUS and GlimmerHMM were trained using the training set created through protein and EST alignments, GeneMarkHMM-ES and BLAST. If a single EST comprised more than one predicted gene in non-overlapping sections, the genes were considered to be 'broken' i.e. split into sections. Genes that comprised several ESTs joined together were considered chimaeric, and genes where EST evidence supported less than the whole gene were labelled as partially supported.

Program	Total number of genes predicted	Genes without in-frame stops					Bases supported by ESTs
		Total Genes	Whole genes matching ESTs	Broken genes	Chimaeric genes	Partial genes	
GeneMark HMM-ES	13,455	5,630	430	226	253	1,584	1,267,185
GeneMark HMM-ES BLAST	5,561	662	163	185	103	449	602,124
Glimmer HMM	8,123	4,757	1,558	446	819	3,085	4,943,205
GeneID	6,342	4,949	83	275	317	1,668	922,905
EuGene	21,689	19,199	3,097	5,983	1,125	15,431	9,256,374
AUGUSTUS	6,043	3,143	1,128	119	774	1,990	4,786,170
GMAP EST	12,840	8,282	1,507	67	114	1,102	1,630,899
Exonerate EST	583	245	15	4	2	39	7,653
Exonerate protein	36 (of 68)	36	5	13	4	24	13,497

Table 4.3. EST support for genes predicted after re-training AUGUSTUS and GlimmerHMM using the genes predicted by EvidenceModeler. If a single EST comprised more than one predicted gene in non-overlapping sections, the genes were considered to be 'broken' i.e. split into sections. Genes that comprised several ESTs joined together were considered chimaeric, and genes where EST evidence supported less than the whole gene were labelled as partially supported.

Program	Total number of genes predicted	Genes without in-frame stops					Bases supported by ESTs
		Total Genes	Whole genes matching ESTs	Broken genes	Chimaeric genes	Partial genes	
AUGUSTUS	6,250	3,316	1,177	136	751	2,109	4,674,270
GlimmerHMM	8,598	5,115	1,625	528	797	3,308	4,987,308

Genes identified by FRANC

The final output of the FRANC pipeline contained 10,175 genes, of which 218 contained in-frame stop codons and 38 fell within regions identified as encoding LTRs, leaving 9,919 expected protein–coding genes. Gene lengths (protein coding) ranged from 150 bp to10,369 bp in length, with an average gene length of 1,007 bp in the unspliced data, and from 147 bp to 9,588 bp long with an average length of 930 bp in the spliced data. The distributions of spliced and unspliced gene lengths are shown in Figure 4.3. In total, a predicted 38.04% of the genome was exonic and 3.16% intronic. A brief summary of the predicted exon and intron content of these genes is given in Table 4.4. The shortest exon predicted in the data set was found to be 3 bp long. This length is far too short for a real exon, and demonstrates some of the problems with using prediction software. Fortunately, from the results above it is apparent that these short exons do not lead to excessively short genes.

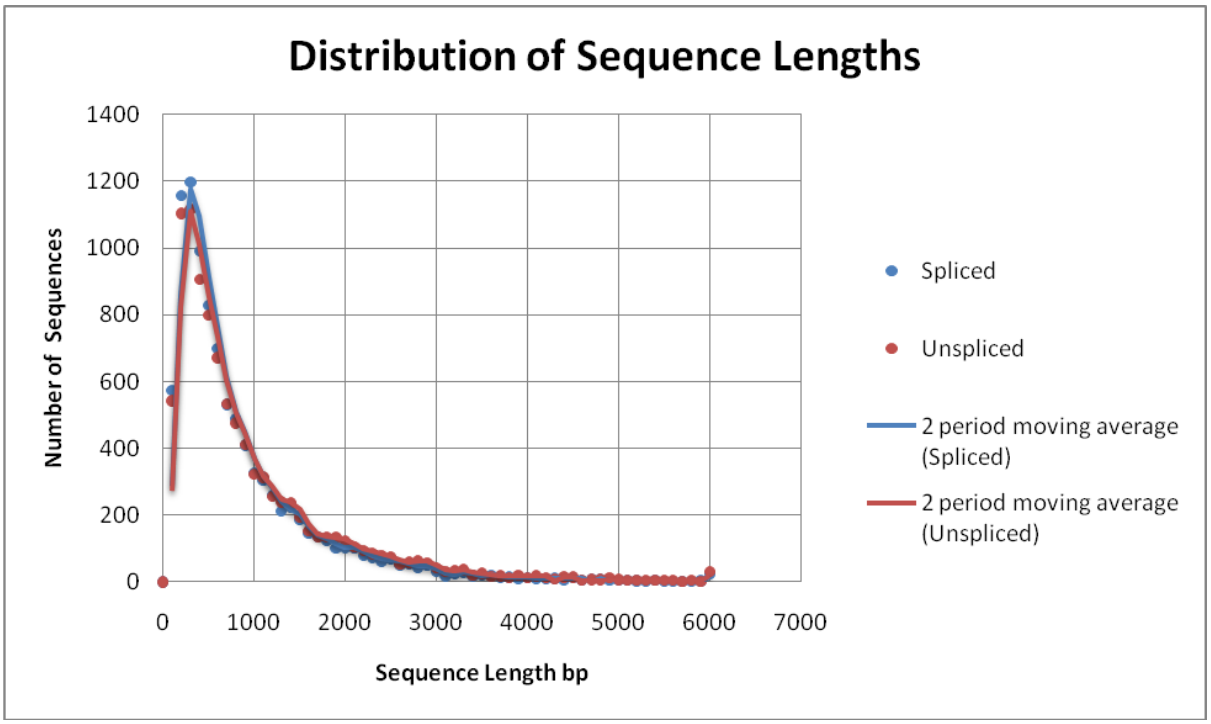

Figure 4.3. The distribution of predicted spliced and unspliced gene lengths using FRANC.

Table 4.4. Exon and intron content of the gene set predicted by FRANC, not including genes containing in frame stop codons.

	Exons	**Introns**
Total bases	9,224,638 bp	766,854
Total exon/intron count	17,973	8,054
Average length	513 bp	95 bp
Shortest exon/intron	3bp	22 bp
Longest exon/intron	4,779 bp	790 bp
Average number per gene	1.812	0.812
Minimum number in a gene	1	0
Maximum number in a gene	15	14
Average %GC content	57.40%	52.04%

4.3.2 GENES WITH EST SUPPORT

The *M. violaceum* EST library introduced in Chapter 2 and used here as supporting information for gene prediction was mapped onto the gene set using tBLASTN. The final data set contained a greater number of genes with complete EST support than any of the individual programs (Table 4.5). Only 160 (1.61%) predicted genes did not match an EST.

Table 4.5. Mapping of EST data onto the predicted genes using tBLASTN.

Total predicted genes	9,919
Total predicted bases (exonic)	9,224,638
Total genes with complete EST support	4,594
Total genes with partial EST support	5,165
Total bases supported by ESTs	8,398,182 (91%)
Chimaeric genes	932
Broken genes	2,644

4.3.3 BLAST2GO ANALYSIS

Of the 9,919 genes predicted by the pipeline and entered into BLAST2GO, 6,783 were assigned BLAST hits. Figure 4.4A shows the species distribution for the BLAST results. At least the top 10 of the species recorded are basidiomycetes. The organism producing the greatest number of hits is the dry rot fungus *Serpula lacrymans*, which produces more hits than the plant pathogens *Melampsora* and *Puccinia* (rust fungi), which were also found in the top five results here. Figure 4.4B is the distribution of the species producing best hits; results are similar to Figure 4.4A, but in this instance the species with the greatest number of highest scoring hits is *Rhodotorula glutinis*, a yeast. Gene Ontology (GO) mapping was performed on the data; 5,186 genes were assigned 1 or more GO terms. Performing protein domain analysis using InterProScan resulted in the InterPro

annotation of nearly 9,000 genes, suggesting that this many genes could either form part of a protein or have been derived from a protein. Of the genes annotated, most contain ontology identifications relating to cellular and metabolic processes Figure 4.5.

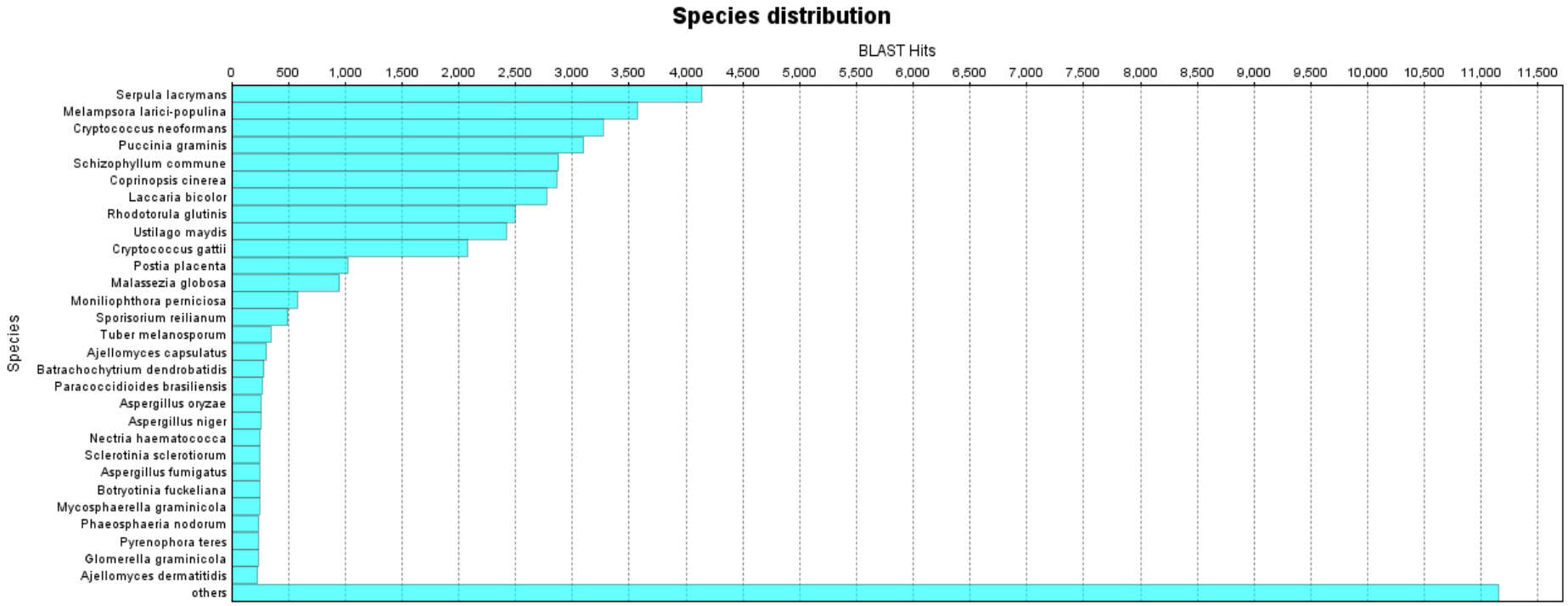

Figure 4.4A. Distribution of species hits from the BLAST2GO BLAST results against the 9,919 *Microbotryum* genes.

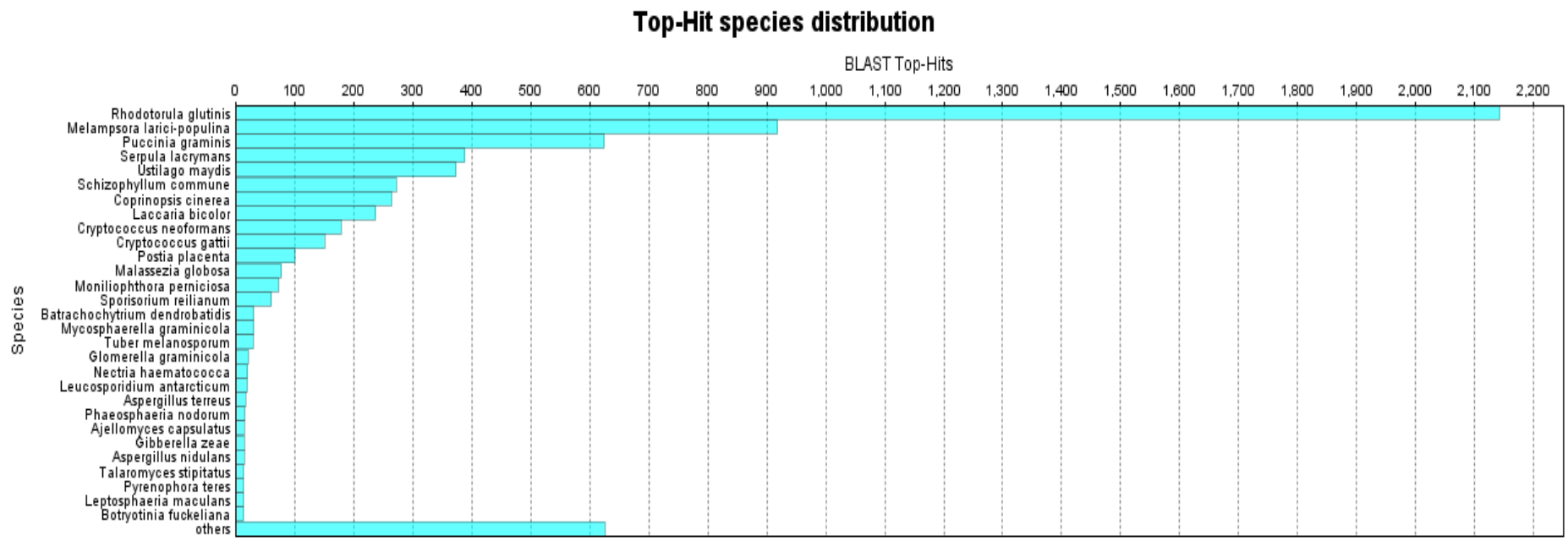

Figure 4.4B. Distribution of species with top hits from the BLAST2GO BLAST results against the 9,919 *Microbotryum* genes.

4.3.4 MATING-TYPE-LINKED GENES

The genome of *M. violaceum* was aligned to the A1 and A2 mating-type-linked regions. A total of 192 genes successfully matched the mating-type-linked regions with a match ≥ 95% in the A2 mating-type-linked regions and an additional 70 genes were identified with a match ≥ 90% in the A1 mating-type-linked regions. From the previous BLAST2GO analysis, 177 (67.56%) genes were annotated using BLAST and 227 (86.64%) could be annotated through InterPro. The difference between the BLAST and InterPro results shows that at least 50 of the genes that are supported by InterPro have no BLAST homologues in the databases. GO term enrichment analysis was carried out using Fisher's Exact Test within BLAST2GO, comparing the 262 mating-type-linked genes to the 9,919 full gene set. The gene set was tested for over and under representation. After correction for false discovery rate (Blüthgen, et al., 2005), and filtering for significantly enriched terms (P>0.5) the mating-type-linked genes were found to be significantly enriched in a number of GO terms, such as nucleotide and DNA binding and control, transcription and expression (Figure 4.6). A single gene was identified as a LTR retrotransposon using RepeatMasker.

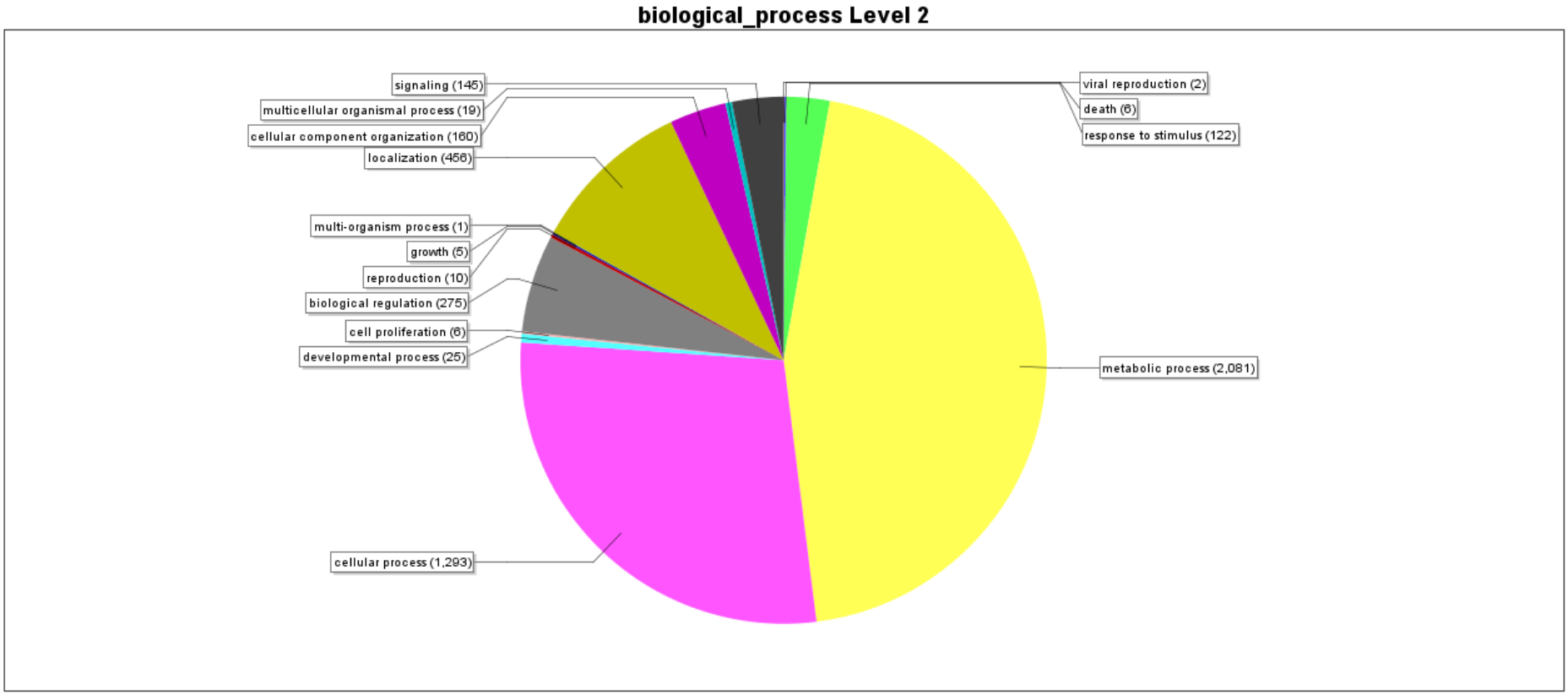

Figure 4.5. Biological process graph produced by BLAST2GO for the 9,919 FRANC gene set.

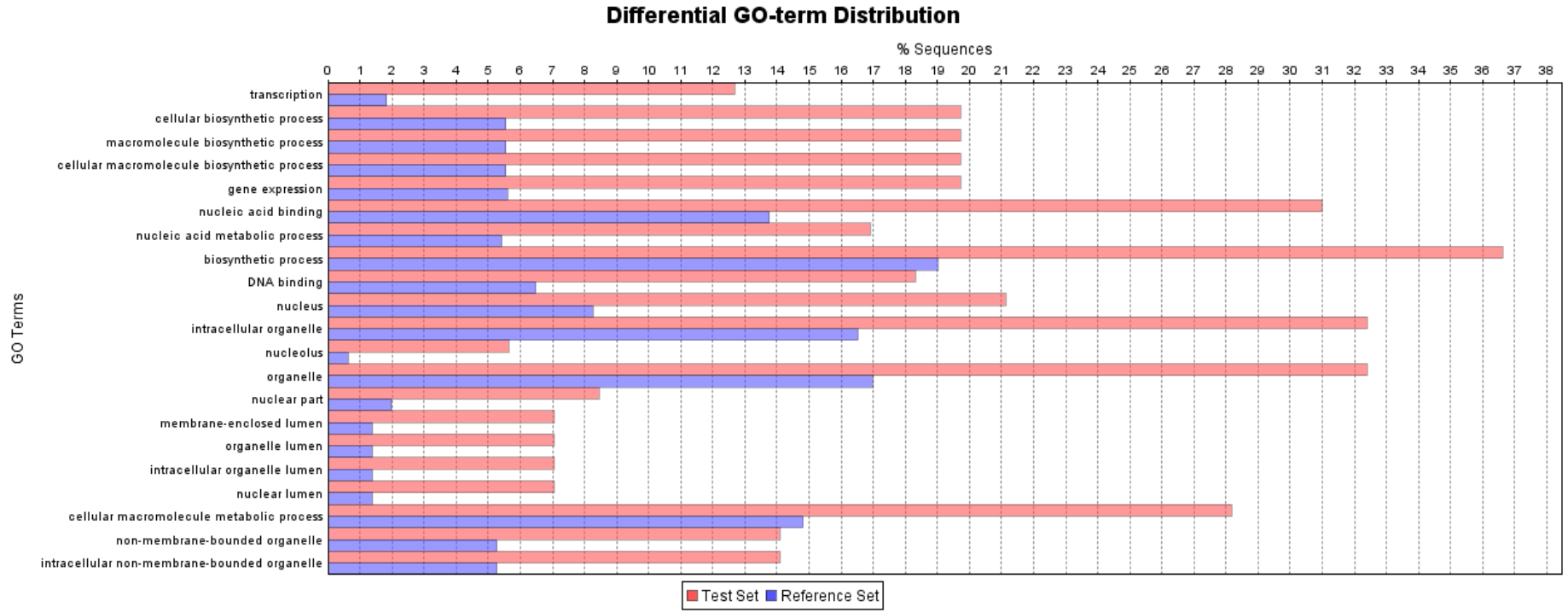

Figure 4.6. GO term enrichment graph produced using Fisher's Exact Test within BLAST2GO, comparing the 262 mating-type-linked genes to the 9,919 full gene set.

4.3.5 REPEAT LINKED GENES

The RepeatMasker analysis revealed that 1,042/9,919 genes contained repetitive elements (Table 4.6), with only 64 (0.64%) of these representing transposable elements (LTR elements and unknown interspersed repeats). Upon examination, the 48 unknown repeats were matches to other portions of the genome, and might represent closely related gene families, duplicated/pseudo genes, or new transposable element genes. No matches to known transposable elements were identified and no DNA elements or small RNAs were found in the data set. Simple repeats were identified as runs of a single amino acid or a small number of repetitive amino acids which can often encode parts of a gene (Katti, et al., 2001).

Table 4.6. Repetative elements of the FRANC genes, identified using RepeatMasker.

Element	Number of elements	Length of elements (bp)	Number of genes effected
LTR elements	15	5,130	16
DNA elements	0	0	0
Unclassified interspersed repeats	52	17,094	48
Small RNA	0	0	0
Simple Repeat	788	41,100	680
Low complexity	423	35,990	372
Total	**1,278**	**99,314 (1.07%)**	**1,042 (10.51%)**

4.3.6 IDENTIFYING GENES IN DIFFERENT SPECIES

Genes from the initial data

The methods used for obtaining genes from the different species were extremely conservative, and designed for accuracy rather than maximising the gene output. Because of this, genes will

have been missed that might have contained alternative exons, or lost/gained exons. It would be possible to search for conserved exons rather than complete genes, and expand the data set in this way. With the exception of the *M. violaceum* var. *caroliniana* and *M. violaceum* var. *dianthus* genes, numbers of genes found in the different species followed the same pattern as the base count of each sequenced genome (Table 4.7). The assembled genomes of *M. violaceum* var. *caroliniana* and *M. violaceum* var. *dianthus* comprised the fewest reads, perhaps accounting for some of the missing genes in these species.

Table 4.7. Genes identified in different *Microbotryum* species. Genomes were aligned to the genes predicted with FRANC using BLAST. FRANC genes were mapped onto the resultant sequences using Exonerate. These genes were reciprocally aligned back onto the FRANC genes using BLAST. *M. violaceum* from *S. latifolia* represents the original FRANC gene set.

Species	Number of genes	Genes without in-frame stop codons
M. violaceum* var. *latifolia	10,137	9,919 (97.85%)
M. violaceum* var. *dioica	9,067	7,802 (86.05%)
M. violaceum* var. *caroliniana	2,904	2,461 (84.75%)
M. violaceum* var. *vulgaris	4,616	4,014 (86.96%)
M. violaceum* var. *saponaria	5,099	4,423 (86.74%)
M. violaceum* var. *dianthus	3,856	3,343 (86.70%)
M. violaceum* var. *flos-cuculi	7,317	6,548 (89.49%)

Genes from trained AUGUSTUS

Genes predicted using AUGUSTUS for each species are presented in Table 4.8, and for comparison, genes in *M. violaceum* var. *latifolia* are also presented. As with genes located through homology searching, numbers of genes detected in each species followed the pattern of assembly length, where the genomes with a higher base content contained more genes. The

average length of the spliced genes increased with genetic distance from *M. violaceum* var. *latifolia*, as did the average number of introns/exons per gene (Table 4.8). The average intron length decreased slightly over this distance. It is possible that predictions become less reliable with distance from the training genes; in this case *M. violaceum* var. *latifolia*, and this possibility should be tested before further examinations into the differences in intron contents between species are performed. Mapping of EST data to the genome from *M. violaceum* var. *latifolia* (Table 4.9) reveals an increase in supported sequence data over the previous training runs of AUGUSTUS, but a decrease when compared to the FRANC genes. Interestingly, the average gene length of the predicted genes decreases with distance from *M. violaceum* var. *dioica*, and the average number of introns per gene increases with distance from *M. violaceum* var. *latifolia*. The average intron numbers are higher than the genes from the FRANC pipeline (1.3–1.5 compared to 0.8). However, from the EST analysis of the genes, we can see that there are many broken genes in the FRANC gene set (2,644), far more than in the trained AUGUSTUS gene set (136). The number of chimaeric genes produced by AUGUSTUS is also usually lower; therefore it is likely that the true intron count is closer to the 1.3–1.5 predicted by AUGUSTUS.

Table 4.8. Genes from 7 *Microbotryum* species predicted using AUGUSTUS trained on the final pipeline output. All measures are for genes without stop codons.

Species	Number of genes	Genes without stop codons	Total spliced gene base count	Average length of spliced genes	Average intron length	Average exon length	Average introns per gene	Average exons per gene	Maximum introns per gene
Silene latifolia	8,976	6,086 (67.80%)	7,110,150	1,168	138	494	1.36	2.36	17
Silene dioica	8,889	5,988 (67.36%)	6,922,287	1,156	134	483	1.39	2.39	20
Silene caroliniana	8,202	5,494 (66.98%)	6,464,601	1,176	125	476	1.47	2.47	22
Silene vulgaris	8,498	5,725 (67.37%)	6,781,500	1,184	126	478	1.48	2.48	18
Saponaria officinalis	8,601	5,746 (66.81%)	6,866,802	1,195	127	481	1.48	2.48	17
Dianthus carthusianorum	8,188	5,430 (66.32%)	6,545,223	1,205	122	482	1.50	2.50	23
Lychnis flos-cuculi	8,728	5,809 (66.56%)	6,826,077	1,175	133	482	1.44	2.44	22

Table 4.9. Mapping of EST data onto the predicted *M. violaceum* var. *latifolia* genes (table 4.8) using tBLASTN. Only genes not containing stop codons were considered. If a single EST comprised more than one predicted gene in non-overlapping sections, the genes were considered to be 'broken' i.e. split into sections. Genes that comprised several ESTs joined together were considered chimeric, and genes where EST evidence supported less than the whole gene were labelled as partially supported.

Total predicted genes	6,086
Total predicted bases (exonic)	12,827,400
Total genes with complete EST support	1,807
Total genes with partial EST support	4,159
Total bases supported by ESTs	6,267,027
Chimaeric genes	950
Broken genes	989

4.3.7 ALL SPECIES GENE ALIGNMENTS

A total of 2,383 genes were found in all seven *Microbotryum* species. After multiple species alignments, 379 genes were found to contain stop codons or were no longer a multiple of three base pairs, leaving 2,004 genes for the PAML analysis. It is likely that many of these stop codons are the result of sequencing error in one or more species that could be corrected, but to avoid including pseudogenes wherever possible, these genes were removed from the analysis.

4.3.8 EVOLUTIONARY ANALYSIS

The consensus gene tree (Figure 4.7A) demonstrates that the phylogenetic relations of *Microbotryum violaceum* species examined mirror the phylogeny of their plant host Caryophyllacae species, with the *Silene* species falling into a monophyletic clade (*Lychnis flos-cuculi* also referred to as *Silene flos-cuculi*). This correlation has been previously demonstrated (Refregier, et al., 2008). Both analyses suggest that *M. violaceum* var. *latifolia* and *M. violaceum* var. *dioica* are the most closely related, consistent with the *Silene* phylogeny. When looking at the plant hosts themselves, Refregier et al., (2008) demonstrated that for the genes used in their analysis, *Lychnis flos-cuculi* fell into a separate clade from the other *Silene* species examined. Results of the *Microbotryum* phylogenies were more similar; both this study and Refregier et al., (2008) suggesting that *M. violaceum* var. *flos-cuculi* is more closely related to *M. violaceum* var. *latifolia* than *M. violaceum* var. *caroliniana*. The placement of *M. violaceum* var. *vulgaris* is the main difference between the concatenated gene tree produced in this analysis and the consensus gene tree created by Refregier et al., (2008) The phylogeny reconstructed in this analysis suggests that *M. violaceum* var. *flos-cululi* is more closely related to *M. violaceum* var. *dioica* and *M. violaceum* var. *latifolia* than *M. violaceum* var. *caroliniana* is, whereas in the Refregier et al., (2008) gene trees, *M. violaceum* var. *vulgaris* has diverged at a similar time to *M. violaceum* var. *flos-cuculi*, and after *M. violaceum* var. *caroliniana*. It is possible that this difference in the placement of *M. violaceum* var. *vulgaris* reflects a genuine difference between different isolates of this species, as two separate invasions of *M. violaceum* var. *vulgaris* have been reported (Le Gac, et al., 2007).

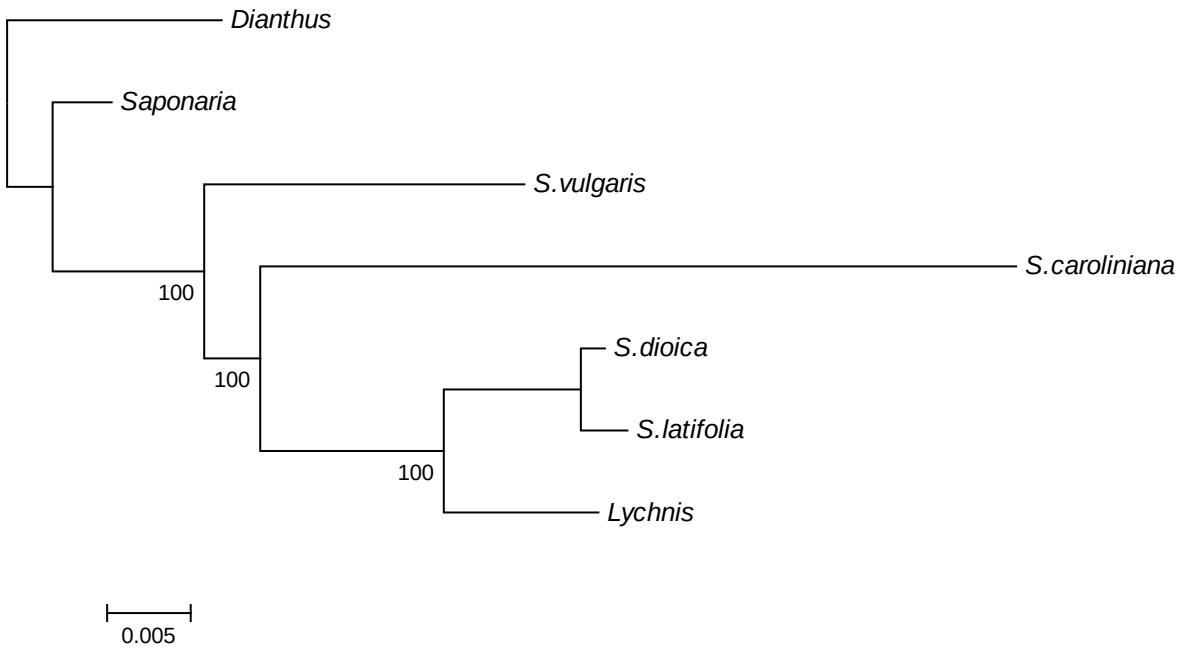

Figure 4.7. The consensus gene trees for seven *Microbotryum* species, artificially rooted with *Dianthus* and computed in MrBayes using a Bayesian approach. *Microbotryum* species are listed under the seven host species names: *Dianthus carthusianorum, Lychnis flos-cuculi, Saponaria officinalis, Silene caroliniana, Silene dioica, Silene latifolia* and *Silene vulgaris*.

Codon-based pairwise divergence at synonymous sites was calculated using MEGA5 between the *Microbotryum* species using the numbers of synonymous substitutions at synonymous sites (Nei and Gojobori, 1986). The highest synonymous divergence across all sites was 0.194 (19%) between *M. violaceum* var. *dianthus* and *M. violaceum* var. *caroliniana,* and the lowest was 0.010 (1%) between *M. violaceum* var. *latifolia* and *M. violaceum* var. *dioica*.

The PAML analysis revealed that 70 of the 2,004 genes (3.49%) were considered to be under selection using the likelihood ratio test (LRT) on models M7 and M8. It is likely that the overall

proportion of selected genes in the genome would be slightly higher, as the determination of protein coding genes was extremely conservative, potentially removing rapidly evolving genes. In Chapter 6, selected genes are examined in 6 mammal species; human, chimpanzee, macaque, mouse, rat and dog. Of the genes investigated, 4.82% were found to be under positive selection – slightly more than the *Microbotryum* genes.

The 70 selected genes ranged from 58 to 2,581 amino acids in length. The average intron and exon counts (1.95/2.95) per gene (Table 4.10) were not significantly higher (one-tailed t-test $P>0.1$) than those in the 1,934 unselected genes present in all species (0.95/1.95).

Table 4.10. Intron and exon content for the 70 selected genes.

	Exons	**Introns**
Total bases	103,450	12,544
Total exon/intron count	195	129
Average length	530	97
Shortest exon/intron	12	43
Longest exon/intron	4,261	690
Average number per gene	2.95	1.95
Minimum number in a gene	1	0
Maximum number in a gene	13	12

Selected genes were examined for their repeat content, using the previous analysis of repeats identified within all genes. Three of the selected genes contain simple trinucleotide repeats, corresponding with short amino acid repeats and one of the genes was found to be GA-rich, as determined by RepeatMasker. A single gene was found to be repeated in another area of the

genome; this gene might represent a gene duplication or shared domain. These results suggest that the selected genes are true genes and not transposable elements.

When aligned to the mating-type-linked regions of the genome using tBLASTn, 2 of the 70 selected genes matched the A1-linked regions with accuracies of 96% and 100% of the complete gene and 1 gene matched the A2-linked region with an accuracy of 99% (the 96% match in the A1 genome). Selected genes were examined using Fisher's Exact Test using BLAST2GO for enrichment of GO terms when compared to the full species gene set. The gene set was tested for over and under representation. After correction for false discovery rate (Blüthgen, et al., 2005), there were no significantly enriched ontology terms in the genes identified as being under selection. Only three of the selected genes did not have BLAST hits, one of which had previously been identified as mating-type-linked. The second mating-type-linked gene matched a hypothetical DNA-binding protein from *Serpula lacrymans*.

4.4 DISCUSSION

It is well established that fungi have small genomes, and *Microbotryum violaceum* is no exception, with an estimated genome size of 25 mb (see Chapter 2). Organisms with small genomes will generally have minimal repetitive element content (Novikova, et al., 2009); for example the basidiomycete *Phanerochaete chrysosporium* has a 30 mb genome (Martinez, et al., 2004) with an LTR retrotransposon content of around 3% of the genome (Novikova, 2010). Given this information, the LTR-retrotransposon content of 0.306% predicted for *M. violaceum* var. *latifolia* by LTRharvest appears low. *Cryptococcus neoformans* is a basidiomycete pathogen that infects humans, and is a model organism for fungal pathogenicity. The 19 mb genome comprises

~5% transposons (Loftus, et al., 2005), compared to 1.65% in *M. violaceum* var. *latifolia*. The lack of repetitive elements in *M. violaceum* var. *latifolia* compared to the smaller genome of *C. neoformans* could be due to the loss of these elements in the *Microbotryum* genome, but considering the percentage of the genome matched to itself using Crossmatch (8.8%) it might be reasonable to assume that the genome contains transposable elements that could not be characterised using the current homology-based and structural modelling approaches. It is also possible that the repetitive element content of the *M. violaceum* var. *latifolia* genome is underestimated due to the difficulties of sequence assembly of these regions. Fortunately this will have more of an effect on simple/tandem repeats and low complexity sequence than it will on transposable elements. Alternatively, the related basidiomycete pathogen *Ustilago maydis* contains only 1.1% transposable element-derived sequences (Kamper, et al., 2006), and it is possible that a similar method of transposable element repression is present in *Microbotryum*.

The initial gene count produced by FRANC was nearly 10,000 genes, with trained AUGUSTUS run on different species constantly coming up with values around 6,000 and only around 2,000 genes could be found to match sufficiently in all species. Computationally predicting genes is an uncertain task, particularly in an organism that lacks empirically verified genes and has no closely related sequenced relative. In this Chapter, steps have been taken to minimise the effect of false positives when comparing genes from different genomes. However, false negatives have been largely ignored in this conservative approach. It might therefore be possible to expand the gene set by running multiple gene predictions and relaxing gene filtering. Genes will also be missed from the data sets which appear in regions of the genome not aligned by the mapping software, and regions that might contain misassemblies. Deeper sequencing of *M. violaceum* var. *dianthus* and *M. violaceum* var. *caroliniana* would increase the number of genes found between species, demonstrating that the number of genes collected is only as strong as the weakest link. It is likely that the true gene count for the species might be somewhere between 6,000 and 10,000 genes.

When also considering the EST library, containing 9,131 unique genes, it is likely that this number is closer to ~10,000. The genome of *C. neoformans* currently has 6,967 annotated protein products (*Cryptococcus neoformans* var. *grubii* H99 Database), and the genome of *Ustilago maydis* has 6,522 annotated genes. The *Microbotryum* genome has a slightly larger estimated genome size (~25 mb) when compared to *C. neoformans* and *U. maydis* (both ~19 mb) and therefore might contain a larger number of genes. Although the fungi mentioned are phylogenetically distant, both were found in the top 10 BLAST hits to the genes identified using FRANC. The recently sequenced genome of the rust fungus *Melampsora laricis-populina* was the second most common top BLAST hit to the genes of *M. violaceum* var. *latifolia*, this genome contains an estimated 16,268 genes (http://genome.jgi-psf.org/Mellp1/Mellp1.home.html), however the assembled genome size, 101 mb, is much larger than that of *Microbotryum*.

The final count of 2,004 genes found in all species might seem low, given the low divergence rates between the species. Indeed, the initial BLAST runs detected >97% of all 9,919 genes in all species, however, when applying the filtering thresholds using Exonerate this number was greatly reduced, particularly in *M. violaceum* var. *caroliniana*, where just over 3,000 genes remained. This species consistently matched fewest genes and had the highest pairwise divergence between all pairs and the longest branch length of the phylogeny. Removing the *M. violaceum* var. *caroliniana* genes resulted in nearly 1000 extra genes that could be found in all species. The number of selected genes increased from 70/2,004 (3.49%) to 88/2,987 (2.95%) with the extra genes, lowering the proportion of selected genes.

The average intron content of *Microbotryum* genes is estimated to be around 1.4 introns per gene. This number is high compared to *U. maydis*, which is thought to have experienced massive intron

loss, with an average of 0.46 introns per gene (Kamper, et al., 2006) but low compared to *C. neoformans,* which has an average of 5.3 introns per gene (in a 19 mb genome) (Loftus, et al., 2005), and other basidiomycetes. For example, the mushroom *Coprinus cinereus* has an average of 4.5 introns per gene, and the white rot *Phanerochaete chrysosporium* has an average of 2.6 introns per gene. The distribution of average intron counts only approximately follows the phylogeny; *Coprinus* (4.5) is most closely related to *Phanerochaete* (2.6) followed by *Cryptococcus* (5.3) and finally *Ustilago* (0.46) and *Microbotryum* (1.4) (James, et al., 2006). We might therefore infer that *Microbotryum* has also experienced intron loss in its genes similar to that inferred in *Ustilago* (Kamper, et al., 2006).

Of the bases in the final FRANC gene set, 91% were supported by EST data, with only 160 (1.61%) genes not matching any portion of an EST. The high number of genes considered 'broken' is likely due to the prediction program Eugene, which is extremely cautious in predicting intron boundaries and often chooses to split genes. The pipeline was weighted to be a balance of the different program strengths, and the high weight given to EuGene has lead to the shortening of genes, but has increased the total base count predicted as more exons can be included. The accuracy of genes might be increased by weighting in favour of a different program, however, no program will be completely accurate, therefore in this instance the high proportion of broken genes has contributed to the number of genes and bases that could be aligned between all 7 species.

A total of 6,783 (68.38%) of the *M. violaceum* var. *latifolia* genes predicted using FRANC were identified using BLASTp searches in BLAST2GO, with nearly 9,000 genes containing recognised InterPro motifs. This suggests that a high proportion of the genes identified are truly protein

coding, or derived from protein coding genes, and may contain novel genes. Of these genes, 64 were found to be linked to transposable elements. The 15 of these that were linked to LTR elements were incomplete LTR elements, therefore these genes represent either LTR derived genes or broken (likely to be inactive) LTRs. This leaves ~2,100 genes with recognised structural domains but no sufficiently related homologues for annotation. Further investigations into these genes would be beneficial to search for novel genes; comparing these genes with the predicted genes of other fungal genomes such as *C. neoformans* and *U. maydis* might help to classify these predictions.

Results identified here are just a small step towards the characterisation of the genome of *Microbotryum violaceum*. Repetitive element identification, gene identification and evolutionary analysis can all be extended into further analyses. Investigations into pathogenicity in the fungus can be carried out by mapping other fungal pathogenicity genes onto *Microbotryum* genes or the genome, which might be useful in understanding how the fungus attacks its plant hosts. Further identification of mating-type-linked genes will allow investigations into the evolution of sex in heterothallic fungi. With the genes identified in this Chapter, characterisation of the *Microbotryum* genomic content and evolutionary history is still in its infancy.

4.5 REFERENCES

Allen, J.E. and Salzberg, S.L. (2005) JIGSAW: integration of multiple sources of evidence for gene prediction, *Bioinformatics*, **21**, 3596-3603.

Anisimova, M. and Yang, Z. (2007) Multiple hypothesis testing to detect lineages under positive selection that affects only a few sites, *Molecular Biology and Evolution*, **24**, 1219-1228.

Benson, G. (1999) Tandem repeats finder: a program to analyze DNA sequences, *Nucleic Acids Research*, **27**, 573-580.

Birney, E., Clamp, M. and Durbin, R. (2004) GeneWise and GenomeWise, *Genome Research*, **14**, 988-995.

Blanco, E., Parra, G. and Guigó, R. (2002) Using GeneID to identify genes. In, *Current Protocols in Bioinformatics*. John Wiley & Sons, Inc.

Blüthgen, N., *et al.* (2005) Biological profiling of gene groups utilizing Gene Ontology, *Genome Informatics*, **16**, 106-115.

Cantarel, B.L., *et al.* (2008) MAKER: An easy-to-use annotation pipeline designed for emerging model organism genomes, *Genome Research*, **18**, 188-196.

Conesa, A., *et al.* (2005) Blast2GO: A universal tool for annotation, visualization and analysis in functional genomics research, *Bioinformatics*, **21**, 3674-3676.

Ellinghaus, D., Kurtz, S. and Willhoeft, U. (2008) LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons, *BMC Bioinformatics*, **9**, 18.

Fickett, J.W. and Tung, C.-S. (1992) Assessment of protein coding measures, *Nucleic Acids Research*, **20**, 6441-6450.

Fugmann, S.D. (2010) The origins of the Rag genes — from transposition to V(D)J recombination, *Seminars in Immunology*, **22**, 10-16.

Galagan, J.E., *et al.* (2005) Sequencing of Aspergillus nidulans and comparative analysis with A. fumigatus and A. oryzae, *Nature*, **438**, 1105-1115.

Goodwin, T.J.D. and Poulter, R.T.M. (2000) Multiple LTR-retrotransposon families in the asexual yeast Candida albicans, *Genome Research*, **10**, 174-191.

Götz, S., *et al.* (2008) High-throughput functional annotation and data mining with the Blast2GO suite, *Nucleic Acids Research*, **36**, 3420-3435.

Haas, B., *et al.* (2008) Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments, *Genome Biology*, **9**, R7.

Huelsenbeck, J.P. and Ronquist, F. (2001) MRBAYES: Bayesian inference of phylogenetic trees, *Bioinformatics*, **17**, 754-755.

James, T.Y., *et al.* (2006) Reconstructing the early evolution of Fungi using a six-gene phylogeny, *Nature*, **443**, 818-822.

Jiang, N., *et al.* (2004) Pack-MULE transposable elements mediate gene evolution in plants, *Nature*, **431**, 569-573.

Kamper, J., *et al.* (2006) Insights from the genome of the biotrophic fungal plant pathogen Ustilago maydis, *Nature*, **444**, 97-101.

Katti, M.V., Ranjekar, P.K. and Gupta, V.S. (2001) Differential distribution of simple sequence repeats in Eukaryotic genome sequences, *Molecular Biology and Evolution*, **18**, 1161-1167.

Kent, W.J. (2002) BLAT—The BLAST-Like Alignment Tool, *Genome Research*, **12**, 656-664.

Korf, I., *et al.* (2001) Integrating genomic homology into gene structure prediction, *Bioinformatics*, **17**, S140-S148.

Kupfer, D.M., *et al.* (2004) Introns and splicing elements of five diverse Fungi, *Eukaryotic Cell*, **3**, 1088-1100.

Larkin, M.A., *et al.* (2007) Clustal W and Clustal X version 2.0, *Bioinformatics*, **23**, 2947-2948.

Lauermann, V., Hermankova, M. and Boeke, J.D. (1997) Increased length of long terminal repeats inhibits Ty1 transposition and leads to the formation of tandem multimers, *Genetics*, **145**, 911-922.

Le Gac, M., Hood, M.E. and Giraud, T. (2007) Evolution of reproductive isolation within a parasitic fungal species complex, *Evolution*, **61**, 1781-1787.

Li, Y., *et al.* (2011) Domestication of transposable elements into MicroRNA genes in plants, *PLoS ONE*, **6**, e19212.

Linder-Basso, D., *et al.* (2001) Crypt1, an active Ac-like transposon from the chestnut blight fungus, Cryphonectria parasitica, *Molecular Genetics and Genomics*, **265**, 730-738.

Loftus, B.J., *et al.* (2005) The genome of the Basidiomycetous yeast and human pathogen Cryptococcus neoformans, *Science*, **307**, 1321-1324.

Majoros, W.H., Pertea, M. and Salzberg, S.L. (2004) TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders, *Bioinformatics*, **20**, 2878-2879.

Martinez, D., *et al.* (2004) Genome sequence of the lignocellulose degrading fungus Phanerochaete chrysosporium strain RP78, *Nat Biotech*, **22**, 695-700.

Mathé, C., *et al.* (2002) Current methods of gene prediction, their strengths and weaknesses, *Nucleic Acids Research*, **30**, 4103-4117.

Meyer, I.M. (2007) A practical guide to the art of RNA gene prediction, *Briefings in Bioinformatics*, **8**, 396-414.

Nei, M. and Gojobori, T. (1986) Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions, *Molecular Biology and Evolution*, **3**, 418-426.

Novikova, O., Fet, V. and Blinov, A. (2009) Non-LTR retrotransposons in Fungi, *Functional and Integrative Genomics*, **9**, 27-42.

Novikova, O.S. (2010) Diversity and evolution of LTR retrotransposons in the genome of Phanerochaete chrysosporium (Fungi: Basidiomycota), *Genetika*, **46**, 725-733.

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

Pearson, W.R. (1996) Effective protein sequence comparison. In Russell, F.D. (ed), *Methods in Enzymology*. Academic Press, pp. 227-258.

Refregier, G., *et al.* (2008) Cophylogeny of the anther smut fungi and their caryophyllaceous hosts: Prevalence of host shifts and importance of delimiting parasite species for inferring cospeciation, *BMC Evolutionary Biology*, **8**, 100.

Ronquist, F. and Huelsenbeck, J.P. (2003) MrBayes 3: Bayesian phylogenetic inference under mixed models, *Bioinformatics*, **19**, 1572-1574.

Salzberg, S. (2007) Genome re-annotation: A wiki solution?, *Genome Biology*, **8**, 102.

Schiex, T., Moisan, A. and Rouz, P. (2001) EuGene: An Eucaryotic gene finder that combines several sources of evidence. , *Computational Biology*, 111-125.

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

Sleator, R.D. (2010) An overview of the current status of eukaryote gene prediction strategies, *Gene*, **461**, 1-4.

Smit, A.F.A., Hubley, R. and Green, P. (1996-2010) RepeatMasker Open-3.0, *www.repeatmasker.org*.

Stanke, M. and Waack, S. (2003) Gene prediction with a hidden Markov model and a new intron submodel, *Bioinformatics*, **19**, ii215-ii225.

Tamura, K., *et al.* (2011) MEGA5: Molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods, *Molecular Biology and Evolution*.

Ter-Hovhannisyan, V., *et al.* (2008) Gene prediction in novel fungal genomes using an ab initio algorithm with unsupervised training, *Genome Research*, **18**, 1979-1990.

Wessler, S.R. (2006) Transposable elements and the evolution of Eukaryotic genomes, *Proceedings of the National Academy of Sciences*, **103**, 17600-17601.

Wu, T.D. and Watanabe, C.K. (2005) GMAP: A genomic mapping and alignment program for mRNA and EST sequences, *Bioinformatics*, **21**, 1859-1875.

5. POSITIVE SELECTION DIFFERS BETWEEN PROTEIN SECONDARY STRUCTURE ELEMENTS IN *DROSOPHILA*

KATE E. RIDOUT

5.1 INTRODUCTION

Factors affecting the rates of evolution in protein coding regions have long been studied by evolutionary biologists. Rates of evolution vary not only between proteins, but also between different sites within a single protein, and many factors have been proposed to account for this variation, such as distance from functional sites (Dean, et al., 2002), base composition (Bernardi, 2005), codon usage (Bernardi, 2005; Bulmer, 1991; Holloway, et al., 2008; Yang and Nielsen, 2008), and degree of solvent exposure (Benach, et al., 2000; Bishop, et al., 2000; Dean, et al., 2002; Hughes and Nei, 1988; Lin, et al., 2007). Functional residues are often the most conserved regions of the protein (Benach, et al., 2000; Dean, et al., 2002; O'Farrell, et al., 2008), and solvent exposed residues are the most changeable. Regions of the amino acid chain that are buried in the protein do not evolve freely (Lin, et al., 2007) while disordered regions of the protein tend to evolve more rapidly (Brown, et al., 2002). However, the action of positive selection in the protein tends to be more complex. In functional regions, for example those involved in protein-protein interactions, certain residues may be highly conserved, or the region might comprise a patch of residues, in which the surrounding physiochemical properties, rather than the exact residues are critical (Binkowski and Joachimiak, 2008; Bouvier, et al., 2009).

Protein secondary structure, the physical arrangement of the amino acid chain produced mainly by the amino acid sequence, is another factor that may contribute to varying rates of evolution at different amino acid positions. The amino acid order directly affects protein folding, and therefore tertiary structure and function, and is highly conserved between homologous proteins. It is known that different secondary structures have different physical and chemical properties and roles in the protein. While this would suggest that protein secondary structure may be involved in determining rates of evolution, this question has not fully been explored, and existing investigations have been on a small scale (Benach, et al., 2000; Dean, et al., 2002; Hanada, et al.,

2006; Petersen, et al., 2007), where results were specific to a particular protein domain or family. However, it is known that the type of protein secondary structure (i.e. α-helix, β-sheet or coil) affects base composition, amino acid frequency and even substitution rates in mammals (Chiusano, et al., 1999). There is therefore good reason to suspect that protein secondary structure plays a role in determining site–specific rates of evolution. To investigate this possibility, a large scale genomic study is required, using source organisms with well sequenced, mapped and annotated genomes.

The publication of complete genomes from twelve closely related species of fruit fly (*Drosophila* 12 Genomes Consortium 2007) provides a valuable comparative resource in which to study the action of natural selection. Similarly, the wealth of knowledge about these organisms facilitates the biological interpretation of any observed trends. Using this data set, Larracuente et al. (2008) investigated and reviewed the many factors that can affect the variation in rates of evolution between different proteins in *Drosophila*. These included gene expression, essentiality, intron number, intron and protein lengths, protein–protein interactions, recombination and translational selection. These factors were shown to act by either increasing the rate of adaptive evolution, or by imposing evolutionary constraints. Since selection was calculated for whole proteins, secondary structure was not included and has remained largely unexplored.

Secondary structures are traditionally separated into two types — ordered regions and aperiodic/unstructured regions. The ordered regions form two main structures, helices and β-structures, while the aperiodic regions can be divided into random coils — natively unstructured stretches of the amino acid chain — and turns (or loops), which are amino acid chain reversals, usually containing one or more hydrogen bonds (Marcelino and Gierasch, 2008; Shepherd, et al., 1999).

The arrangement of an amino acid chain into a secondary structure is based on both the residues in that chain and the surrounding environment. While particular amino acids are more frequent in different structures, these correlations are weaker than previously thought, and neighboring residues (in sequence or in space) are important in determining secondary structure (Beck, et al., 2008).

The likelihood of positive selection to alter an amino acid at a given site may depend on several factors: the physical and chemical nature of the amino acid (will the replacement interact favorably with the surrounding residues and environment without damaging protein function?), the functional importance of the site (how critical is it that the exact residue or a physiochemically similar residue is maintained?), the surrounding environment (does the residue or comprising structure require a specific range of hydropathy?), the physical properties of the structure (degree of order), and the folding properties of the structure. These restrictions on the occurrence of positive selection are complex and not all of these can be analyzed with the data available.

It has been found that the most variable regions of a protein are on the solvent accessible surfaces (Lin, et al., 2007), and are therefore likely to include a high proportion of hydrophilic residues. The weak correlation between secondary structure and the frequencies of different amino acids, which each have different hydropathies based on their side chain charge, means that the four secondary structure categories have different likelihoods of containing hydrophobic or hydrophilic residues (Chou and Fasman, 1974). In particular, β-turns often contain hydrophilic residues (Marcelino and Gierasch, 2008), and are thought to sit on the outer (solvent exposed) surfaces of the protein where they might play a role in protein folding and protein–protein interactions (Marcelino and Gierasch, 2008; Shepherd, et al., 1999). We might therefore expect to see an increase in both positive selection and purifying selection, as both conservation and

adaptation of these residues is important. β-strands (a common type of β-structure) often contain the most hydrophobic residues, and these hydrophobic interactions are the predominant factor that stabilizes β-sheets (Chou and Fasman, 1974; Koehl and Levitt, 1999), which are therefore often buried in the protein core. β-strands may therefore contain less positively selected sites than the other structures. Helices are amphipathic overall (Chou and Fasman, 1974), and may therefore occur anywhere in the protein, with one side of a helix often being hydrophobic and the other side hydrophilic, although, like β-strands, helices can form hydrophobic bundles in the protein core. The numbers of positively selected sites is therefore likely to be greater in helices than β-strands but less than in β-turns. Helices and β-strands are the most rigidly structured types of secondary structure and should therefore contain fewer positively selected sites than β-turns and coils because a greater proportion of potential mutations would be disruptive to the secondary structure. Indeed, several amino acids are known to break the structure of helices and β-strands in their native state (Beck, et al., 2008; Chou and Fasman, 1974). The other type of β-structure examined here, the β-bridge is not expected to differ significantly from β-strands. Finally, random coils (unstructured regions) are by definition free of the structural interactions necessary for other secondary structures; they are therefore less likely to have constraints on hydropathy, position in the protein or amino acid composition. Differences in rates of selection between secondary structures may have profound effects on protein evolution and therefore on phenotypic change. Understanding the degree to which secondary structure determines the amount of positive selection will help to explain the general patterns of evolution, and uncover a previously neglected level at which natural selection may act, between the amino acid and the protein levels.

Here, we infer positive selection in a phylogenetic framework (using the ratio of non-synonymous to synonymous substitutions, dN/dS; Hughes and Nei 1988) across six species of *Drosophila*, using a data set of c. 8,500 genes published by Larracuente et al. (2008). We also analyze the distribution of secondary structures and selected sites along the length of a gene and investigate

how it affects the degree of positive selection. Finally, we examine the levels of hydropathy for sites and structures undergoing positive selection to build an overall picture of how evolution is influenced by protein secondary structure. We demonstrate that within this diverse range of proteins, residue changes characterized as being positively selected are distributed unevenly among protein secondary structures.

5.2 MATERIALS AND METHODS

5.2.1 DATA ACQUISITION

Aligned nucleotide sequence data were obtained from the published genomes of twelve species of *Drosophila* (Drosophila 12 Genomes Consortium, 2007). Following the methods used by Larracuente et al.(2008), genes that exist as single copy orthologs in *D. melanogaster*, *D. simulans*, *D. sechellia*, *D. yakuba*, *D. erecta* and *D. ananassae* were selected for analysis. Saturation in silent site divergence outside the *melanogaster* species group precludes the use of all 12 genomes (Larracuente, et al., 2008). *Drosophila* sex chromosomes evolve at different rates to autosomes, with lower levels of polymorphism and faster divergence (Begun, et al., 2007), and were therefore excluded. Masked nucleotide alignments (i.e. alignments from which uncertain sections have been removed) from the six species in the *D. melanogaster* group were downloaded from the FlyBase FTP site (ftp://ftp.flybase.net/genomes/12_species_analysis/clark_eisen/alignments/melanogaster_group.guide_tree.longest.cds.masked.tar.gz).

Following Larracuente et al. (2008), all sites in the aligned sequences with gaps or ambiguous sites in more than one of the six sequences were removed. In addition we also re-analyzed the same data after exclusion of all sites with gaps present in any of the six species. This has not affected the conclusions of the paper. Any genes whose length varied between the two data sets were then excluded, leaving a total of 8,492 genes for our analyses. Since different alignments

can produce different outcomes in phylogenetic analyses (Wong, et al., 2008), we re-aligned all the genes with CLUSTAL W (Thompson, et al., 1994), DIALIGN-TX (Subramanian, et al., 2008) and MUSCLE (Edgar, 2004) using the default options, and used the resulting alignments in addition to those obtained from Larracuente et al. (2008). As the results presented below are robust to the choice of alignment software we used the alignments obtained from Larracuente et al. (2008).

5.2.2 DETERMINATION OF SECONDARY STRUCTURE

All protein structure sequences (145,944 at the time of writing) from the RCSB Protein Data Bank (PDB) were downloaded and aligned against the 8,492 *Drosophila* genes using NCBI BLAST (BLASTX) with an expectation E-value cutoff of 10^{-6}. For every match, the top hit from each BLAST run was taken. In total, 1,092,117 experimentally determined structure residues aligned to portions of 3,884 genes.

In addition to the experimentally determined structural data, we used computational methods to predict secondary structures for our data set. *Drosophila melanogaster* has the best characterized genome of any of the 12 *Drosophila* species; we therefore chose this model organism for the secondary structure prediction. Since the other sequences were aligned to the *D. melanogaster* genome, any section of the alignment where the sequence for *D. melanogaster* was unavailable would be unreliable, and was excluded from further analyses.

PSIPRED (Bryson, et al., 2005; Jones, 1999) was used to predict secondary structures. PSIPRED uses neural networking and searches for homologous proteins with known structures to determine

the most likely structure at each residue position. The homology information is collected using PSI-BLAST, and is combined with individual properties of the amino acids for creating or breaking different secondary structures and the likely structure lengths. Local sequence information is incorporated using a sliding window approach. Many of the most reliable secondary structure prediction methods available use neural networking in combination with BLAST or PSI-BLAST searches (Montgomerie, et al., 2006). Results obtained during testing using the CASP3 project (Critical Assessment of Techniques for Protein Structure Prediction experiment) demonstrated that the PSIPRED method was the most accurate at that time, achieving a score of nearly 80%, the highest of all programs tested (Moult, et al., 1997). Since these tests, PSIPRED has continued to be used for further developing structure prediction (Zhang, et al., 2008), and remains a leading secondary structure prediction program (Birzele and Kramer, 2006).

PSIPRED reports the probabilities for each site of falling into each of the three structural categories, based on the DSSP structure definitions (Kabsch and Sander, 1983): helix, which contains both the α-helix (DSSP code H) and the 3_{10} helix (DSSP code G); strand, which contains β-sheets (DSSP code E) and isolated β-bridge residues (DSSP code B); and finally coil (all remaining DSSP codes including β-turns). We used the probabilities of each of these states rather than the single most likely structure in order to incorporate the uncertainty of the structure prediction method.

PSIPRED classifies hydrogen–bonded turns and natively unstructured regions together as "coils". In order to tease apart these two structural classes, we used the probabilities given by PSIPRED in conjunction with the predictions made by the neural networking program BTPRED (Shepherd, et al., 1999). BTPRED takes the secondary structure predictions produced by PSIPRED and can

predict whether or not a residue is in a hydrogen–bonded β-turn with an accuracy of over 70% (Kaur and Raghava, 2002), although it has a tendency to over-predict β-turn residues (Shepherd, et al., 1999). BTPRED predicts whether a site is more likely to be a β-turn or a coil and provides a "reliability index" — the amount by which the predicted structure is more likely than the alternative, in tenths. The probability assigned by PSIPRED to the "coil" class was therefore divided between β-turn and natively unstructured, according to the probability derived from BTPRED. In the few cases where BTPRED's chosen prediction was actually the less likely of the two (reliability index = "*"), both were considered equally likely. The probability was therefore used as a conditional probability of BTPRED's prediction being true, given that the structure was considered a "coil" by PSIPRED.

5.2.3 INFERENCE OF POSITIVE SELECTION

The program codeml from the phylogenetic analysis package PAML 4.0 (Yang, 2007) was used to infer sites which have experienced positive selection, based on the ratio of non-synonymous nucleotide changes per non-synonymous site to synonymous changes per synonymous site ($dN/dS = \omega$) at each codon. Synonymous changes are assumed to be functionally neutral (Kimura, 1968). The program assumes a certain number of classes, to which sites are assigned depending on the calculated value of ω. We used the default parameters, and two pairs of nested models: M1a/M2a and M7/M8. In each case, the more general model differs from the other only in allowing an additional class of sites with $\omega > 1$, i.e. sites under positive selection. Thus, a likelihood ratio test (LRT) between such nested models is explicitly testing whether the gene is under positive selection. Model M1a (Yang, et al., 2000) has only two classes — one where ω is between 0 and 1 (negative selection) and one where $\omega = 1$ (neutral evolution) — while model M7 (Yang, et al., 2000) has 10 classes with the value of ω for each following a β distribution between 0 and 1. The models M2a and M8 are similar to M1a and M7 but both include an extra class of

codons with $\omega > 1$ to accommodate positively selected sites (Yang, et al., 2000). We used the robust Bayes empirical Bayes procedure (Wong, et al., 2004; Yang, et al., 2005) implemented in PAML to detect individual sites under positive selection in genes identified by the LRT. PAML gives a probability of each site belonging to each class, and the probability for the class where $\omega > 1$ is therefore the probability that the site is under positive selection, which we will call P_s. Sites in genes with a positive LRT, $\omega > 1$ and $P_s \geq 0.9$ are considered to be positively selected. In the experimentally determined structure data set, sites with $P_s \geq 0.5$ were also considered to be positively selected. This inclusion will increase the number of false positives in the data, therefore, wherever possible, the $P_s \geq 0.9$ threshold was also used.

The greater complexity of model M8 is likely to better fit the situation in nature, but explicitly including a class where $\omega = 1$ in M2a can allow sites evolving under weak positive selection or neutral evolution to fall into this class instead of the class under positive selection. This conservative approach is particularly appropriate for analysis with few taxa (Anisimova, et al., 2002). By using both models to search for the same underlying trends, we hope to avoid any specific effects of individual models, and thus provide stronger support for any results found (Anisimova, et al., 2002).

Since different genes may follow different gene trees, each gene was analyzed using the most appropriate tree topology for that gene. The tree that provided the best result for each gene is listed at ftp://ftp.flybase.net/genomes/12_species_analysis/clark_eisen/paml. Over the 8,492 genes, three trees were used, differing only in the placement of two species, *D. erecta* and *D. yakuba*, for which there is known discordance between gene trees and species trees (Pollard, et al., 2006).

It was suggested by Lindsay et al. (2008) that codon models used to estimate ω might be affected by sequence composition. More recently, Yap et al. (2010) demonstrated that this is indeed the case for such models, for example, the Goldman and Yang method (GY) used by PAML. The GY model uses continuous time Markov processes to model substitutions (in order to estimate ω), the rates of which are specified by an instantaneous rate matrix, the parameters being based on rates of codon change (in this case in the gene). This matrix is then weighted by the frequency of the codon and not by the frequencies of the individual nucleotides that make up that codon. Thus, if sequence composition varied between secondary structures, the rate assumptions made by the codon model would be violated, making them unsuitable. Lindsay et al. (2008) suggested that models which weight substitutions by nucleotide frequencies, such as the MG model (Muse and Gaut, 1994) are more robust to nucleotide composition than the GY model.

The models used for the PAML analysis, M1a/M2a and M7/M8, all use the GY method. It is therefore possible that ω might vary between structures based on their sequence composition. To gain a better understanding of any effects of this bias in ω on our data, the following simulations were run: Sequences were simulated with PyCogent (Knight, et al., 2007), under the MG nucleotide substitution method. The rate parameters for the substitution matrix (i.e. transition/transversion rates and divergences between species) were taken from the concatenation of all 8,492 genes used in our analyses. One large gene was simulated for each of the four structures, where the nucleotide frequencies used to simulate each gene were taken from the overall proportion of a given nucleotide in a particular structure in the real data set (e.g., the proportion of thymine nucleotide's in all helix structures in the 8,492 genes). Two sets of simulations were run; one with ω equal to 1 in all structures and another with the average ω from the real data, 0.26. Using the distributions of gene lengths and structure lengths found in the 8,492 genes, 1,000 genes for ω =1 and 1000 for ω =0.26 were simulated. In the simulation, different secondary structure elements will evolve equivalently, so long as nucleotide composition varying

between the structures has no effect on the result. Therefore, if the GY method is unbiased in this instance, there should be no difference in the proportion of positively selected sites between the four structure classes (when the simulated data is run through the codeml package of PAML to search for positive selection). Though this analysis gives us a better understanding of whether protein secondary structure over the entire data set varies enough in general sequence composition to confound our results, it is not definitive. Due to nucleotide and codon composition potentially varying between secondary structure elements (Chiusano, et al., 1999), and the different structural compositions of genes, it is possible that this effect may still confound the results.

We also investigated the degree of codon bias in different structures, since certain secondary structures may use rare codons preferentially, in order to slow translation down, and thereby aid protein folding (Komar, 2009). An excess of rare codons in any of the secondary structures could lead to a reduction in the synonymous substitution rate, decreasing dS, which could artificially increase ω. To test whether variation in codon bias across the structures could affect synonymous substitution rate and hence estimates of ω, we compared the effective number of codons (Wright, 1990) in the four secondary structures, where structures were determined by computational prediction.

Amino acid content may vary between structures; it is therefore possible that differing rates of selection in amino acids might lead to the difference between the secondary structures. To examine the link between amino acid content and positive selection in a structure, the amino acid and the predicted structure at each selected position from the *D. melanogaster* lineage was recorded. The secondary structure each amino acid belongs to was also recorded at all sites. The fraction of selected sites was calculated for each amino acid by dividing the number of selected

sites of an amino acid by the total number of sites of that amino acid (regardless of structure). The expected number of sites under selection for a particular amino acid in a structure was then estimated by multiplying the fraction of selected sites for each amino acid by the observed number of that amino acid in each structure. This number was compared to the observed numbers of selected sites for each amino acid in all four structures.

5.2.4 HYDROPATHY IN SELECTED AND NON-SELECTED SITES

Since amino acids with different hydropathies can favor different secondary structures (Chou and Fasman, 1974), a better understanding of how the likelihood of selection varies with secondary structure might be gained by looking at the changes in hydropathy resulting from changes between the current amino acid and its ancestral state at selected sites. Changes in hydropathy, measured with the hydropathy index of Kyte and Doolittle (1982), at selected sites and non-selected sites were calculated for each of the four structures in the predicted structure data set using the amino acids corresponding with the ancestral nucleotide sequence reconstructed by PAML (marginal reconstruction) from the M8 analysis. The mean hydropathy was also calculated from the current amino acids for each of the four structures in all sites and at selected sites. Variation in hydropathy along the length of a single protein could lead to a bias in the amino acids and hence relative proportions of the secondary structures found at different positions in a protein. To test for this possibility each gene was divided into 20 equal segments and the mean hydropathy of the amino acids calculated for each.

The distance of a residue in a protein from the periphery and the core of a protein has an effect on the likelihood of positive selection (Lin, et al., 2007). In addition, the likelihood of a secondary structure to be solvent exposed, and therefore in the exposed peripheral residues of the protein

varies due to the intrinsic amino acid content of each secondary structure (Chou and Fasman, 1974). To explore this link we used experimentally determined structures from the Protein Data Bank to produce an independent estimate of how often different secondary structures are present on the exposed surfaces of proteins. All structures reported for *D. melanogaster* were examined, with duplicates (proteins that displayed over 95% sequence similarity) excluded. In total, 160 proteins were available. The solvent–exposed areas of each structure from this random data set were calculated. Secondary structures were taken directly from the Protein Data Bank and solvent accessibility was calculated using MSMS (Sanner, et al., 1996).

5.2.5 SPACING OF SELECTED SITES

Distances between selected sites were recorded along each gene. To test whether any clustering of selected sites was due to the different proportions of positively selected sites in different secondary structures, rather than directional selection, we ran the following simulation: Data were simulated using the known proportions of selected sites in different secondary structures and the observed length distributions of secondary structures in the data set as a whole. The length distributions of the different structures were recorded from the 8,492 *Drosophila* genes by taking the most likely of the four structures (from the combined PSIPRED and BTPRED structure predictions) to be the absolute structure at each residue. For the simulation, lengths of structures were chosen randomly from the observed distribution, without replacement. Each site was given the appropriate structure–specific probability of being under positive selection. The intervals from one selected site to the next were recorded. Simulations were performed in Java, and repeated 5000 times in order to provide confidence limits on the observed frequencies.

We also tested whether the clustering of sites under positive selection is due to changes on the same or different branches of the phylogeny. The proportion of parallel changes that we might expect to observe on a single branch of the phylogeny depends on the branch lengths in the tree, since the probability of a second substitution occurring on the same branch is equal to the branch's length as a proportion of the entire tree length. The overall proportion of parallel substitutions is therefore $\Sigma(L_i^2)$, where L_i is the length (as proportion of the whole tree) of the i^{th} branch. These lengths were approximated using the PAML ancestral state reconstructions to count the occurrences of an amino acid at a selected site changing along each branch of the tree. Pairs of adjacent selected sites along a given amino acid sequence were categorized by the distance between them, with distances above an arbitrary boundary of 30 amino acids being considered large, and those below 30 amino acids considered small. The proportion of pairs of adjacent selected sites which experienced substitutions on the same branch of the tree was compared against the expectation, for both large and small distances. It is possible that both the observed and expected results are slightly underestimated, as multiple changes in the same position on the same branch cannot be detected. However, both results are calculated using the same method, and therefore, we do not expect this to introduce a bias.

5.2.6 SELECTION IN THE ENDS OF THE GENES

To investigate whether the proportion of sites under selection is influenced by the position within the gene, every gene was arbitrarily split into 20 equal segments, and the number of selected and non-selected sites were counted in each computationally predicted structure and in each segment. Using the M8 data, the ω values of every residue were plotted for each of the 20 gene segments, to see if the distribution of ω varies with position in the gene. A skew of ω values to be closer to 1 at the ends of genes would indicate a relaxation of purifying selection in these regions. In addition, if the ends of the gene experience a relaxation of purifying selection then in the context

of the whole gene these regions would be more likely to be picked up as positively selected sites. To examine this possibility, each gene was manually divided into two parts. One contained the first 15% of the gene, concatenated with the last 5%, as these regions encapsulate the gene segments with the sharpest increase in the proportion of positively selected sites. The second part comprised the remainder (the central part) of the gene. PAML model M2a was run twice for every gene on the two parts separately to determine whether there was a difference in strength of positive selection, number of positively selected sites, strength of purifying selection and number of sites under purifying selection between the gene ends and the gene centre.

5.2.7 INSERTIONS AND DELETIONS WITHIN STRUCTURES

Pascarella and Argos (1992) demonstrated that insertions and deletions (indels) were enriched in reverse turn and coil structures. Indels occurring in one or more species in our data were removed; despite this, the remaining indels and the areas that surround previous indel sites may represent areas of increased alignment ambiguity. This could potentially lead to a perceived increase in substitution rate, and therefore a high false inference of positive selection in these regions, predominantly at turn and coil sites, where we would expect the most indels. To test whether regions surrounding indels bias the proportion of selected sites found in each structure, we examined the distribution of selected sites around each indel within a predicted structure. Each gene was examined individually, for each section of a structure along the length, the distance from every site and every selected site to the nearest indel was recorded. As all sequences were aligned against *D.melanogaster* the gaps in this species of the raw data (as downloaded, before any gaps were removed) was used.

5.3 RESULTS

5.3.1 SELECTION IN SECONDARY STRUCTURES

Among the 3,884 genes for which experimentally determined structure data was available, a total of 1,092,117 residues were included. Selected sites identified using model M8 from genes with a significant LRT were not randomly distributed among the four experimentally determined structures (χ^2 test: $P < 0.00001$, with strands and β-turns containing fewer residues undergoing positive selection than would be expected by chance (0.53 × expectation and × 0.57 respectively). Coil regions contained more positively selected residues than expected by chance (× 1.83) and helix regions slightly less (× 0.95). Similar results were obtained using the M1a/M2a LRT (Figure 5.1A). The results did not qualitatively differ using M7/M8 where $P_s \geq 0.9$; strands, β-turns and helix regions contained fewer selected residues than expected (× 0.24, × 0.43 and × 0.75 respectively), and coils more (× 4.19).

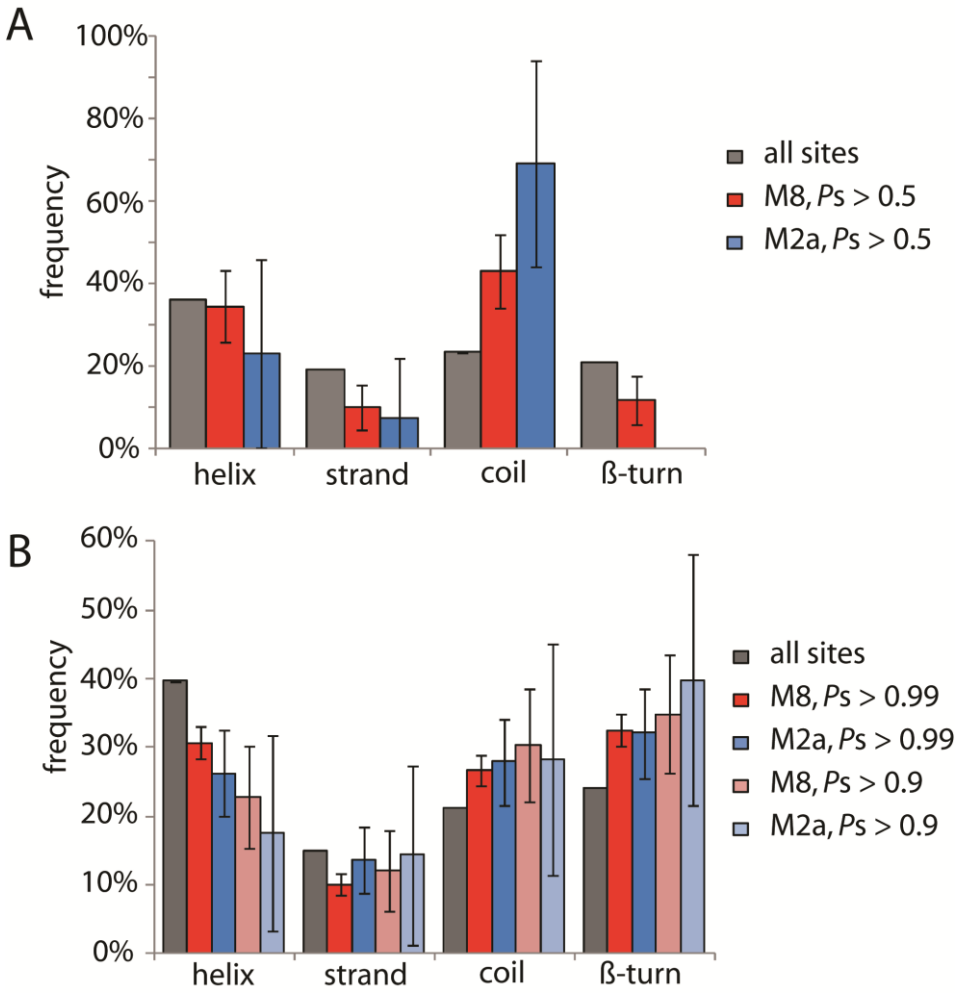

Figure 5.1. Proportions of all sites (grey bars) and positively selected sites (colored bars) according to the M1a/M2a and M7/M8 models in different secondary structures determined experimentally (A) and predicted computationally (B). A threshold probability of $P_s \geq 0.5$ was used in (A) and two thresholds ($P_s \geq 0.9$ or ≥ 0.99) were used in (B). Error bars represent standard deviations.

The data set of computationally predicted secondary structures comprised 8,492 genes, with a total of 4,125,829 aligned residues (Table 5.1). Similarly to the experimentally determined structures, the distribution of selected sites was not random (χ^2 test: $P < 0.00001$ where $P_s \geq 0.9$ and $P = 0.000176$ where $P_s \geq 0.99$ using the M7/M8 comparison). Strands were less likely to

contain positively selected sites than expected (× 0.67); however, β-turns contained more positively selected sites than expected (× 1.35). It was also observed that coil regions contained more selected sites than expected (× 1.26) and helix structures slightly less (× 0.77) (Figure 5.1B). Again, the results were not qualitatively different using genes identified as being positively selected by the M1a/M2a LRT to determine selected sites. For both models, the results were similar at the threshold values $P_s \geq 0.9$ and $P_s \geq 0.99$. When examining only sites with $P_s \geq 0.99$, β-turns (× 1.45) and coils (× 1.43) again have more selected sites than expected, while helices (× 0.57) and strands (× 0.80) are both under-represented.

Table 5.1. Summary statistics for each of the four secondary structures, including total number of sites in each data set, along with the number of sites under selection (at both $P_s \geq 0.9$ and $P_s \geq 0.99$ for predicted structures, but only $P_s \geq 0.5$ for experimentally determined structures using model M8; percentages are expressed as a proportion of all sites in genes with a significant LRT), the effective number of codons (ENC) and the mean value of ω.

	Predicted			**Experimentally determined**		**ENC**	**Mean ω**
	Total sites	**$P_s \geq 0.9$**	**$P_s \geq 0.99$**	**Total sites**	**$P_s \geq 0.5$**		
Helix	1,635,453.8	437.072 0.21%	27.315 0.013%	398,083	41 0.095%	50.47	0.135
Strand	621,585.2	143.728 0.21%	14.466 0.021%	208,798	12 0.053%	50.73	0.129
Coil	873,646.1	380.503 0.54%	36.444 0.052%	227,527	51 0.175%	52.56	0.160
β-Turn	995,143.8	462.893 0.32%	41.911 0.029%	257,769	14 0.054%	52.48	0.158
Total	**4,125,829.0**	**1424.196 0.29%**	**120.136 0.024%**	**1,092,117**	**118 0.098%**	**51.53**	**0.146**

Yap et al. (2010) demonstrated that estimates of ω are affected by nucleotide composition because the codon models assume that genes are homogeneous in terms of sequence composition. If composition varies between the four structures, the assumptions made by the codon models used in this study would be violated, confounding the results. To test whether nucleotide composition heterogeneity could lead to the observed differences in positive selection between secondary structures we analyzed simulated datasets generated in such a way that the only difference between the regions with different structure was their nucleotide composition. There should therefore be no significant difference in the number of positively selected sites between the four structure classes, unless the difference is due to nucleotide composition. Indeed, no such difference was detected (Table 5.2), confirming that nucleotide composition differences between the regions encoding different protein secondary structures are unlikely to cause the observed difference in the number of positively selected sites.

Table 5.2. Overall nucleotide content of the four structures taken from the 8,492 genes and the proportions of selected sites in the four structures taken from the simulated genes. Data was collected using pyCogent, gaps were excluded. CI represents the confidence interval of the proportion of selected sites per structure.

	Amino acids				**Simulated genes**		
	A	**C**	**T**	**G**	**Total sites**	**Selected**	**95% CI**
Helix	0.247	0.261	0.218	0.274	213,252	186	0.00075 – 0.00100
Sheet	0.220	0.253	0.275	0.251	91,437	86	0.00074 – 0.00114
Turn	0.256	0.302	0.164	0.278	169,129	185	0.00094 – 0.00125
Coil	0.262	0.279	0.184	0.276	45,333	45	0.00070 – 0.00128

If the data set contained an excess of rare codons, or strong codon bias, particularly in one structure over the others this could lead to decreased values of dS (the number of synonymous mutations at synonymous sites) when compared to dN (non-synonymous mutations at non-synonymous sites). This decrease in dS relative to dN could cause the artificial inflation of ω (dN/dS) and hence the false inference of positive selection. To investigate this possibility we examined codon bias in the four structures. There was a difference in the effective number of codons (Wright, 1990) used in the different structures. The two ordered structures, helices and strands, had values of 50.47 and 50.73 respectively, while the aperiodic regions showed weaker codon bias (β-turns: 52.48; coils: 52.56). This is the opposite to what we would expect if stronger codon bias was inflating the signal of positive selection in β-turns and coils.

Observed and expected (see Materials and Methods) rates of selection for each amino acid in every structure were compared to test whether biased positive selection of specific amino acids could be responsible for the trend that proportions of positively selected sites vary between structures. The expected proportion of selected sites in each structure was calculated, assuming that neither structure, nor amino acid content had any effect on the number of positively selected sites in each structure. These proportions were compared to the expected number of selected sites if amino acid content alone had an effect on the proportion of selected sites in each structure (data not shown). The analysis revealed that if amino acid content were the cause of the distribution of positively selected sites in secondary structure, we would expect fewer selected turn and coil sites than if the distribution was random, slightly less selected helix sites and a greater number of selected sheet sites. These results are very different from the proportions of selected sites found in the structures, where turns and coils contain more selected sites than expected and sheets less. Helices contain fewer sites under selection than expected at random, though it is significantly less than predicted by the rate of selection in the amino acids. Thus, it appears unlikely that the

difference between the proportions of selected sites in secondary structures is due to different frequencies of amino acids.

5.3.2 ANALYSIS OF HYDROPATHY IN SELECTED AND UNSELECTED SITES

Changes in amino acid hydropathy from the ancestral state to the derived state were measured at selected and unselected sites (Table 5.3). Overall, structures show decreasing hydropathy at both selected ($P_s \geq 0.9$, M7/M8) and not selected sites. This indicates that protein hydropathy is not at equilibrium in these 6 species of *Drosophila*. In β-turns, hydropathy at positively selected sites is more conserved than at all other sites, unlike the other three structures, which show the opposite trend; hydropathy is more conserved at sites that are not undergoing selection. β-turns are expected to extend into the solvent, and therefore might be expected to have different hydropathy characteristics. In terms of overall composition, strand residues were found to be highly hydrophobic on average, β-turns and coils were strongly hydrophilic and helices were weakly hydrophilic.

Table 5.3. Changes in hydropathy from the ancestral amino acid state to the derived state. Mean changes in hydropathy from the PAML reconstructed ancestral state per amino acid substitution for each secondary structure, at selected ($P_s \geq 0.9$) and at not selected ($P_s < 0.9$) sites using the model M8.

	Selected	**Not selected**
Helix	-0.001193	-0.0006577
Strand	-0.001358	-0.0005468
Coil	-0.001322	-0.0012434
β-Turn	-0.000380	-0.0014548
All	-0.000994	-0.0009550

The degree of solvent exposure was calculated for the set of all experimentally determined structures deriving from *D. melanogaster* (regardless of whether we possess a corresponding sequence alignment). The corresponding experimentally determined secondary structure was then taken to determine if any structure was more likely to be solvent exposed or accessible. β-turns are the most likely structure to be in the solvent exposed regions (Figure 5.2). Coils are the next most solvent accessible, followed by helices and finally strands.

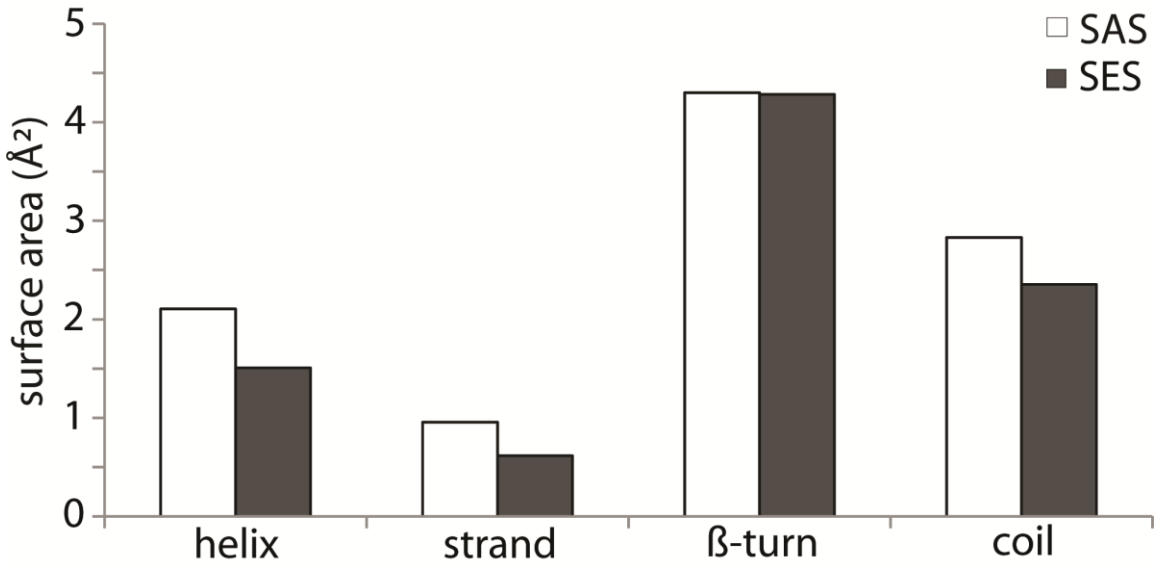

Figure 5.2. Mean solvent exposed (black) and solvent accessible (white) areas, expressed in square ångströms, in each of the four secondary structures.

5.3.3 SPACING OF SELECTED SITES

Under a purely random distribution, the distance from one selected site to the next would be expected to follow a geometric distribution, since the probability of each subsequent site being under positive selection will be equal. We observe a large departure from this expectation, with the likelihood of being under positive selection decreasing with increasing distance from other selected sites. After one selected site, the next selected site was more likely to be encountered within the following 30 amino acids than expected by chance (Figure 5.3). This result was significant ($P < 0.0001$) in genes with a significant LRT under all examined combinations of models and threshold values (M1a/M2a: $P_s \geq 0.9$ and M7/M8: $P_s \geq 0.9$, $P_s \geq 0.99$).

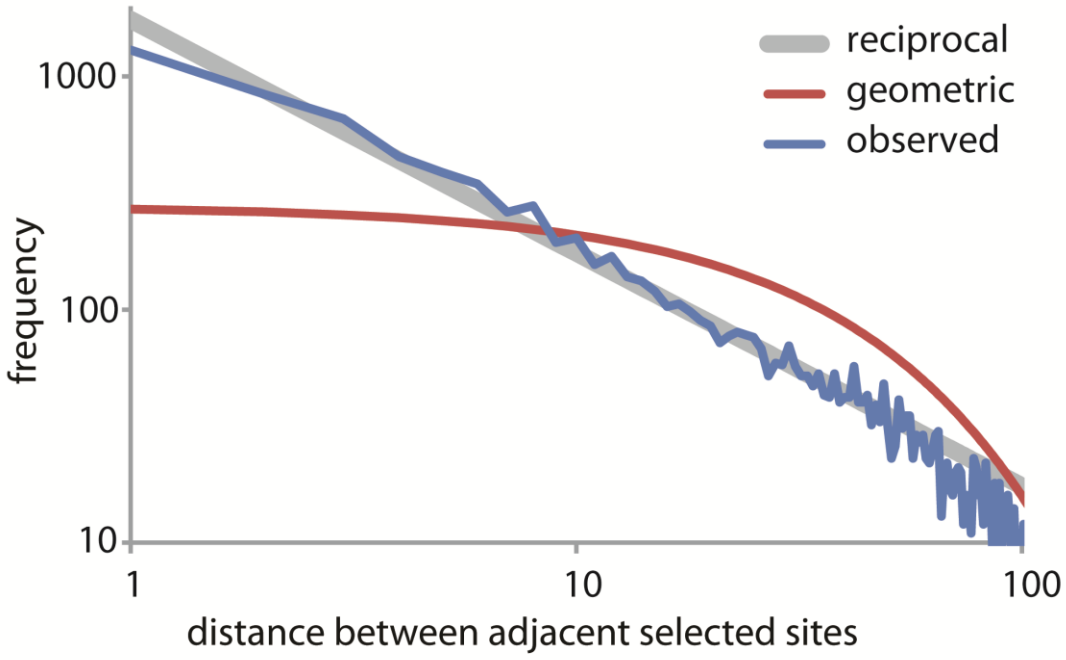

Figure 5.3. The size distribution of intervals between adjacent selected sites on a log–log scale, with the geometric curve expected given no clustering of sites, and a power law fitted to the curve at lower gap sizes.

To test whether this effect was due to the different rates of selection in different secondary structures, data were simulated using the known proportions of selected sites in different secondary structures and the observed length distributions of secondary structures in the data set as a whole. These simulations showed only a slight deviation from the expected geometric distribution, equivalent to a small increase of 15% in the frequencies of positively selected sites one residue apart, compared to the five–fold increase observed in our data. Thus, the observed clustering of positively selected sites cannot be explained by different rates of selection in different secondary structures.

There are at least two possible explanations for the clustering of sites under positive selection. One possibility is that a given gene region may be particularly prone to positive selection forming evolutionary 'hotspots'. On the other hand, selection–driven change at one site may cause an increase in selection at nearby sites, such as compensatory mutations. We can distinguish between these two types of process by observing where on the species phylogeny amino acid changes at selected sites occur. Compensatory mutations should cause adjacent selected site to evolve in concert, on the same branch of the tree. If, on the other hand, selective hotspots are responsible for the pattern, the amino acid changes should be distributed randomly across the phylogeny. We found that the proportion of amino acid changes at adjacent selected sites occurring on the same branch of the tree for smaller intervals (selected residues <30 amino acids apart) and for larger intervals (≥30 residues apart), were significantly greater than the expected values (Table 5.4). Thus, sites under selection within an individual gene are more likely to occur on the same branch of the gene tree than different branches. This result was stronger where sites were closer together (<30 amino acids) and where a more stringent threshold of positive selection was used.

Table 5.4. Proportion of adjacent amino acid changes occurring on the same branch of the phylogeny. 99.15% confidence interval (equivalent to 95% C.I. but after Bonferroni correction for multiple testing). Distances between adjacent selected sites are significantly different from the expected values in all counts. The expected probability of two changes occurring on the same branch of the tree was determined by the sum of squares of the branch lengths.

	Number of observations		**>30 A.A.**	**≤30 A.A.**
	>30 A.A.	**≤30 A.A.**		
Expected			29.0%	29.0%
M2a, $P_S \geq 0.9$	80	468	52.63% (41.96% – 63.30%)*	76.22% (71.69% – 80.75%)*
M8, $P_S \geq 0.9$	488	949	47.42% (43.32% – 51.53%)*	60.56% (57.31% – 63.81%)*
M8, $P_S \geq 0.99$	17	352	58.62% (34.52% – 82.72%)*	83.61% (78.86% – 88.36%)*

5.3.4 SELECTION IN THE ENDS OF THE GENES

When dividing each gene into sections of equal length, secondary structures were found to vary in frequency along the length of a gene (Figure 5.4), with the beginning of a gene and, to a lesser extent, the end, showing a significant decrease in strands ($P < 0.0001$). In addition, there is a significant increase of positively selected sites at both ends of the gene (Figure 5.5). There was not sufficient data to determine whether this variation at the gene ends was due to the increased number of selected residues at β-turn and coil sites, and the increase in β-turns and coils at the ends of genes. However, this is unlikely to account for the entirety of the variation, as the increase in β-turn and coil residues is far smaller than the increase in selected sites at the ends of genes, suggesting that at the ends of the gene there is an additional change to the selective pressures. This result does not differ qualitatively between the two PAML model comparisons, nor between different threshold values. Exclusion of sites with alignment gaps reveals the same pattern: the number of selected sites is still increased at the N and C termini of the genes (data not shown).

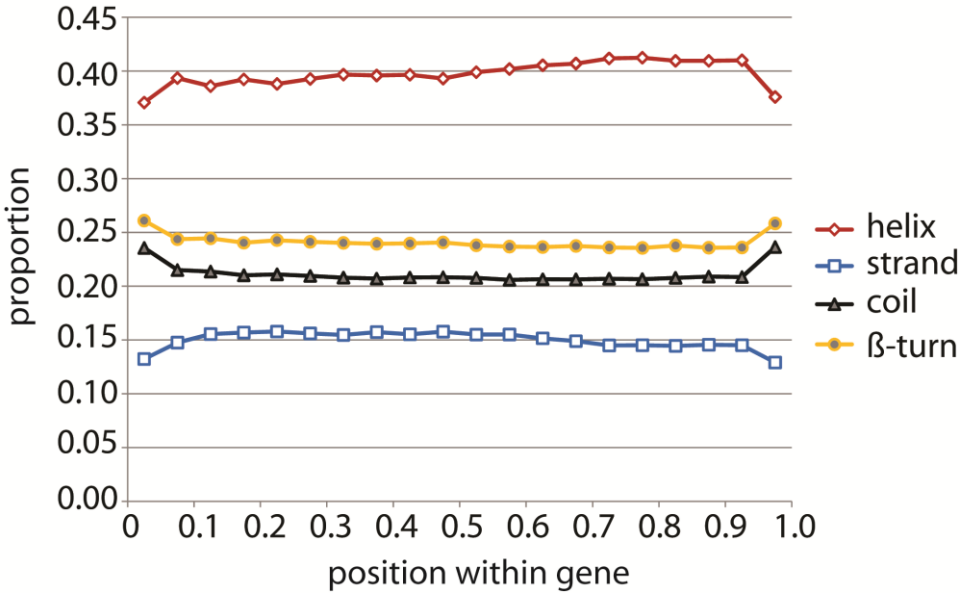

Figure 5.4. Variation in the frequencies of different secondary structures along the length of genes when divided into 20 equal segments at all sites.

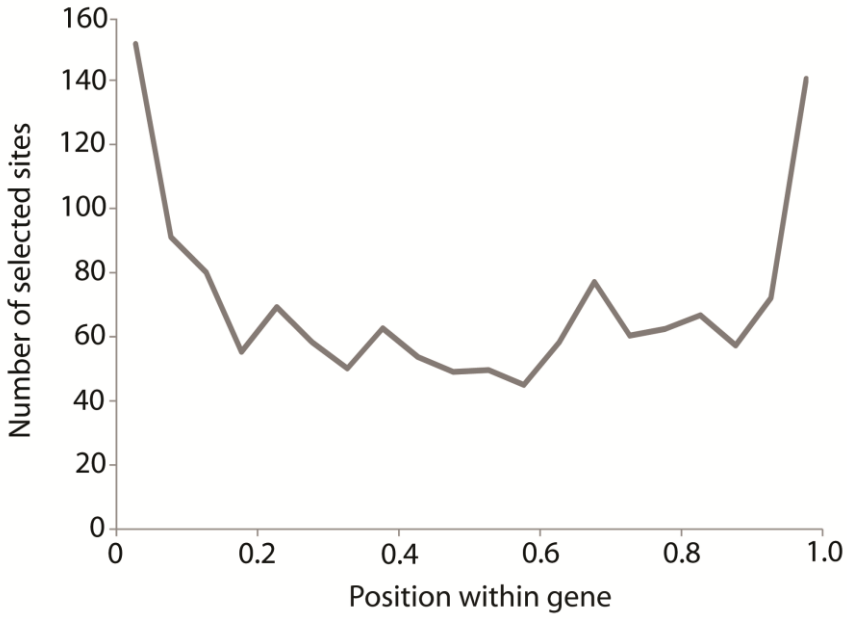

Figure 5.5. Variation in the proportion of sites under selection along the length of a gene, when divided into 20 equal segments. Zero marks the start (N-terminus) and 1 the end (C-terminus).

The distribution of ω along the length of a gene is shown in Figure 5.6. Mean ω is inflated in the first 15% and the last 5% of a gene. Values of ω closer to 1 may be explained by either of two phenomena: a relaxation of purifying selection or an increase of positively selected sites.

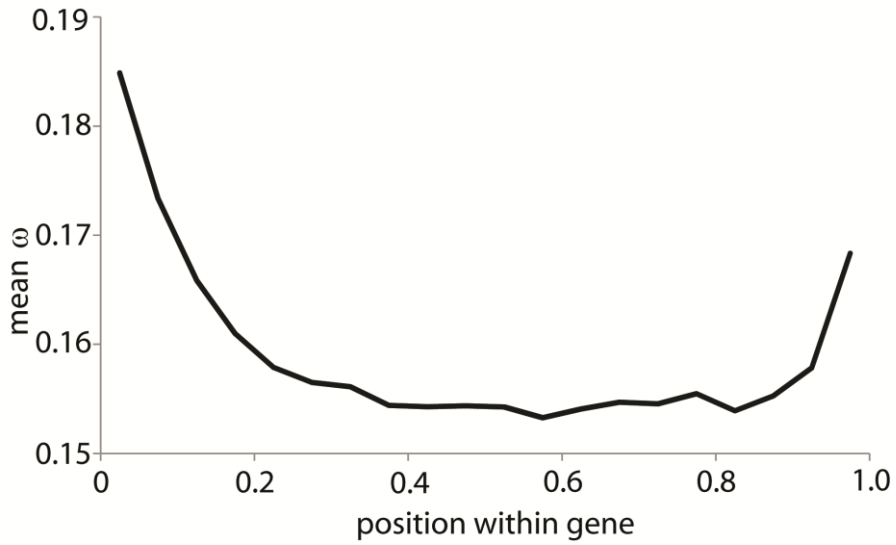

Figure 5.6. Graph of mean ω against position in the gene. Data was binned into 20 equal segments along the gene.

Due to the relative scarcity of positively selected sites compared to the number of sites under purifying selection, the distribution of ω, where ω < 1 in each of the four structures (Figure 5.7) and mean ω values (Table 5.1) provide a broad indication of the strength of purifying selection. These results demonstrate that helices and strands are subject to the strongest purifying selection, with coils experiencing stronger purifying selection than turns, whereas coils were found to contain a higher proportion of selected sites than turns. This therefore suggests that the different proportions of selected sites found between secondary structures should not be wholly due a relaxation of purifying selection. However, the more ordered structures

clearly experience stronger purifying selection than the more disordered structures, and so could potentially contain greater numbers of false positives.

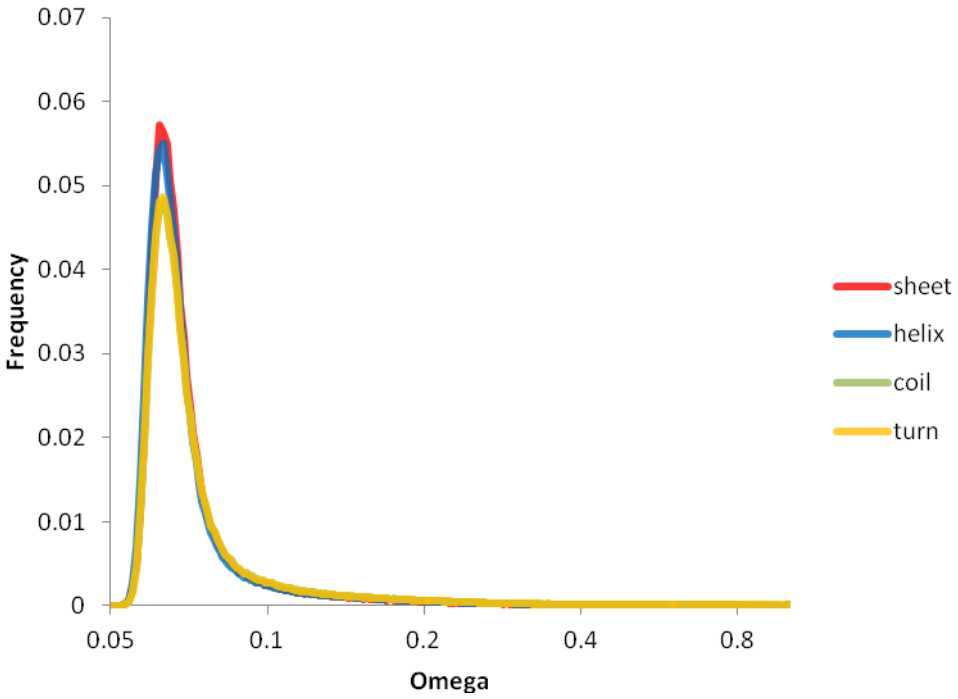

Figure 5.7. Graph of the frequency of ω values where ω < 1 from the M7/M8 data set.

Partitioning the results obtained from running model M2a on all genes into the ends of the genes (the first 15% and the last 5%) and the middle (the remainder), revealed that the distribution of ω > 1 (positive selection) is not significantly different between the middle and the ends of genes ($P = 0.65$, unpaired *t*-test). However, the frequency of positively selected sites at the ends of genes was significantly greater than in the middle ($P < 0.0001$). The number of sites under purifying selection was significantly lower ($P < 0.0001$) at the ends of the genes than in the middle, as more sites were under positive selection. In addition, the distribution of ω < 1, (purifying selection) was significantly skewed towards 1, and therefore weaker at the ends of the genes ($P < 0.0001$).

5.3.5 INSERTIONS AND DELETIONS (INDELS) WITHIN STRUCTURES

Indels may potentially affect the number of positively selected sites identified in a region, as they introduce some uncertainly into alignments. Examining the distribution of indels in different secondary structures reveals that β-turn structures contain less indels than we would expect to see, whereas all other structures contain more indels than the expected value (if indes were equally distributed between structures - data not shown). If the propensity of indels in β-turns were causing the increase in the proportion of selected residues in these regions we would expect to see the opposite result; thus, indels are unlikely to be the cause of the unequal distribution of positively selected sites between secondary structures.

5.4 DISCUSSION

Previous studies of positive selection in secondary structures have examined single genes or domain families (Kosiol, et al., 2008; Mondragón-Palomino, et al., 2002). The results of these analyses each tell us something about the evolution of a specific protein or protein family, but though thorough studies exist to explore many factors affecting the rate of evolution (Larracuente, et al., 2008), no such studies have yet been conducted to examine the relationship between positive selection and secondary structure on the genomic scale. One recent study examined the correlation between SNPs and secondary structure (Liu, et al., 2008). Solvent–exposed regions were the least conserved, while helices and strands were under stronger purifying selection, although the effects of positive selection were not analyzed. Those studies that have discussed the structures in which selection occurs (Alvarez-Valin, et al., 2000), have not had the power to determine differences in selection between secondary structures. This is particularly important where the variability of amino acid residues is used as a proxy to determine sites of functional importance. For example, Thomas et al. (2003) specifically use conserved regions of coding

sequence to infer functionality. However, our results show that different structures (particularly strands) are also likely to produce regions where the amino acids are strongly constrained. In this case, it would be useful to examine the structural composition of the region to determine if this is the case. It would appear from previous studies that regions of functional importance which must adapt quickly (e.g. virus binding regions) contain more positively selected sites (Kosiol, et al., 2008). This demonstration of how positive selection can be spatially limited along a gene demonstrates the importance of understanding why selection varies along a gene.

We demonstrate here that secondary structure has a significant effect on the rate of adaptive evolution in proteins. It appears that of the four predicted secondary structures, β-turns and coils are the most likely to experience positive selection and more periodic strands and helices the least. On the other hand, in the data set with experimentally determined structures, β-turns contained less positively selected sites than expected. This might be due to the difficulty to predict β-turns, however, neural networking methods such as PSIPRED are the most reliable methods of structure prediction currently available (Kaur and Raghava, 2002). Alternatively, it might be due to the difficulty in determining the structure of disordered protein regions. Disordered regions do not have a definite 3D structure, and are therefore difficult to crystallize. Thus, experimentally determined structures may not be a random sample of the *Drosophila* genome. As unstructured regions contain more instances of positive selection, particularly in hydrophilic areas likely to be on the outer surface of the protein, the β-turns and hydrophilic regions of structures (and thus positively selected sites) might be under represented in the experimentally determined data set. Therefore, the β-turns that remain in the experimentally determined structures are likely to be internal to the protein, and therefore behave in a similar fashion to structured regions.

The identification of sites under positive selection using PAML can be hindered by sites experiencing a relaxation of positive selection. In these areas, there will be a higher proportion of sites where ω is close to 1, which can be picked up as being positively selected. This problem is offset partially by the conservative Bayes Empirical Bayes approach taken by PAML (Yang, et al., 2005), and by considering only sites where $P_s \geq 0.9$ or $P_s \geq 0.99$. Additionally, PAML model M2 is known to be particularly conservative when considering weak positive selection, as many of these sites will be rounded into the class ω = 1, or neutral evolution. Despite these controls, relaxed purifying selection will still influence the results. For example, in disordered structures and towards the ends of genes, certain physiochemical properties have lead us to suggest that these sites are more tolerant to mutations, which leads to more variation, a relaxation of purifying selection and therefore a higher proportion of selected sites. This increase in selected sites may include a slight increase in false positives, and almost certainly an increase in weakly selected sites in these regions. Thus, caution should be taken when examining sites in these relaxed variable regions, as selected sites might not be as strongly selected for as a fixed mutation in a more conserved part of a protein.

Changes in hydropathy calculated from the ancestral state of an amino acid to the descendent state at both the $P_s \geq 0.9$ and $P_s \geq 0.99$ threshold levels (M7/M8) revealed that hydropathy is not at equilibrium in the 8,498 genes examined in the 6 species of *Drosophila*. The decreasing hydropathy at sites that were not identified as evolving under positive selection may suggest that additional factors not examined here play a role in shaping the amino acid sequences of proteins. Hydropathy at positively selected sites in coil, helix and strand regions is less conserved than at all other sites, however, the opposite is found in β-turns. This suggests that β-turns might have different hydropathy characteristics to the other three structures examined here. We have also demonstrated for the first time that secondary structures are not evenly distributed along the length of the gene, there being more β-turns and coils towards the ends (Figure 5.4). Positively

selected sites are also more likely to be located at the ends of the gene (Figure 5.5). However, the increase in β-turns and coils at the ends of genes is not sufficient to fully explain the increase of positively selected sites at the ends of genes.

Distributions of ω values for positively selected residues ($\omega > 1$) are not significantly different between the central part and the ends of the gene, although there are significantly more sites under positive selection at the ends of the genes than in the middle. When looking at codons with $\omega < 1$, it was noted that the distribution of ω was closer to 1 at the ends of the genes, and there were fewer sites under purifying selection, indicating an overall relaxation in purifying selection at terminal parts of genes, relative to the middle. This reduction in purifying selection, coupled with more variable sites and less structure (more coils and β-turns), suggests that amino acids at the ends of genes are less constrained than in the middle, and there is therefore more opportunity for mutations to be positively selected.

When observing the variation of selected sites across the length of a gene, the distances between adjacent selected sites deviated from the expected distribution, with a significant excess of sites at shorter distances. It would be reasonable to assume that this clustering of selected sites is because mutations would either be compensatory or in a region of decreased conservation. Purifying selection may tolerate mutations constrained by protein structure only after certain neighboring mutations have occurred. The fact that amino acid changes at neighboring selected sites were more likely to be on the same branch of the reconstructed tree suggests that such mutations are not independent and possibly reflect compensatory evolution. A similar tendency for selection to act on nearby sites along the same branch in a phylogeny has been noted previously for mammals (Bazykin, et al., 2004). It is interesting that parallel changes are detectable over such long timescales as the rat–mouse divergence or speciation within the *D. melanogaster* group. It would

be interesting to study how quickly these parallel changes can occur by carrying out similar comparisons for more closely related taxa. Aris-Brosou (2005), presented the Extended Complexity Hypothesis, discussing the nature of proteins within complex interaction networks to be more conserved by evolution. It may be possible that the observed clustering of selected sites is related to this hypothesis, which might suggest regions of conservation where interactions occur.

Varying strengths in codon bias between the structures could lead to the observed signal of more positive selection in β-turns and coils, compared to other structures. However, although codon bias does differ between the secondary structures, the difference is in the opposite direction to that which would be expected if stronger codon bias were the cause. We have also examined the possibility that a relaxation of purifying selection in certain structures might have been mistaken for positive selection. However, by examining the frequency distribution of ω (where $\omega < 1$) for each structure we have revealed no difference between the four structures (data not shown).

In a recent study of positive selection in *Escherichia coli*, Petersen et al. (2007) pointed out that positive selection was more often found on the outside of the protein, and in proteins on the outer surface of the cell. For example, external loops thought to be responsible for phage binding contain many more positively selected sites than the internal β-barrel region (composed of strands). This is a strong demonstration that regions of proteins that are in contact with external forces are a more likely target for positive selection. Our initial expectation was that the more structured regions (strands and helices) would contain fewer positively selected sites (and polymorphic sites) because they are governed by more strict rules about which residues are physicochemically acceptable than unstructured regions. For example, proline, glycine and valine are known to break helices in their native state (Beck, et al., 2008; O'Neil and DeGrado, 1990).

Therefore, mutations towards these amino acids might not be favorable in helical regions. In addition, the more structured regions — strands in particular — are more likely to contain hydrophobic residues (Chou and Fasman, 1974; Koehl and Levitt, 1999), which is consistent with our results. They are therefore less likely to be in the solvent-exposed regions of the protein, and more likely to be important for protein stability (Dudgeon, et al., 2009) and the prevention of protein aggregation due to hydrophobic interactions. It has also been suggested that internal residues are more important for maintaining the folding of a protein (Alvarez-Valin, et al., 2000; Creighton and Darby, 1989), and that external regions have lower structural constraints. Again, suggesting that external regions should be more susceptible to positive selection and are more robust to both synonymous and non-synonymous polymorphism. In contrast, coil and β-turn regions are more likely to be on the outside of a protein as they do not have the same structurally induced physiochemical constraints (for example necessary hydrophobicity). Thus, these unstructured regions (β-turns in particular) are often hydrophilic (Marcelino and Gierasch, 2008). Helices are known to be amphipathic, and can contain both hydrophobic and hydrophilic residues (Chou and Fasman, 1974; Koehl and Levitt, 1999). Ferrada and Wagner (2008) discuss the correlation between protein robustness and evolution, they suggest that the more 'designable' a protein is (the number of sequence variations that can fold into the correct structure) the greater its ability to evolve. Therefore, proteins that contain structures with more amino acid flexibility (turns and coils) might be expected to have a faster rate of evolution.

Unfortunately, the prevalence or absence of residues in different structures alone is not enough to predict protein secondary structure, and it has recently been contested that the intrinsic tendencies of amino acids for specific conformational preferences is not as strong as previously assumed (Beck, et al., 2008). Our own investigations have determined that in our dataset with predicted structures, β-turns are the most likely to occur in the solvent–exposed regions of the protein, are the most hydrophilic and contain the greatest number of positively selected sites. Strands occurred

on the external solvent–exposed regions of the protein the least out of all of the structures, were the most hydrophobic, and contained the lowest proportion of positively selected sites. Helices contained slightly more positively selected sites than strands, were slightly more hydrophilic, and slightly more likely to occur on the periphery of the protein. Finally, coils were slightly less hydrophilic than β-turns, and were slightly less likely to occur on the outside of the protein. From these results, a pattern begins to emerge where the most structured regions form the complex highly folded, hydrophobic, conserved protein core that experiences more purifying and less positive selection, compared to coils and β-turns.

These results are the first of their kind to demonstrate on a genomic scale that the probability of a residue being under positive selection is dependent on the structure to which the residue belongs. We also determine that other factors, such as position along the gene, hydropathy and distance from the closest selected site have an effect on selection. It will be important for future studies to understand exactly why selection varies along the length of the gene, and to what extent all the results found in this study affect the likelihood of a site to experience positive selection. Knowing how secondary structures are selected will help to disentangle the reasons behind positive selection in a region of a protein, and therefore aid the discovery of positively sites that may be functionally important.

5.5 REFERENCES

Alvarez-Valin, F., Tort, J.F. and Bernardi, G. (2000) Nonrandom Spatial Distribution of Synonymous Substitutions in the GP63 Gene From Leishmania, *Genetics*, **155**, 1683-1692.

Anisimova, M., Bielawski, J.P. and Yang, Z. (2002) Accuracy and Power of Bayes Prediction of Amino Acid Sites Under Positive Selection, *Molecular Biology and Evolution*, **19**, 950-958.

Aris-Brosou, S. (2005) Determinants of adaptive evolution at the molecular level: The extended complexity hypothesis, *Molecular Biology and Evolution*, **22**, 200-209.

Bazykin, G.A., *et al.* (2004) Positive selection at sites of multiple amino acid replacements since rat-mouse divergence, *Nature*, **429**, 558-562.

Beck, D.A.C., *et al.* (2008) The intrinsic conformational propensities of the 20 naturally occurring amino acids and reflection of these propensities in proteins, *Proceedings of the National Academy of Sciences*, **105**, 12259-12264.

Begun, D.J., *et al.* (2007) Population genomics: Whole-genome analysis of polymorphism and divergence in Drosophila simulans, *PLoS Biol*, **5**, e310.

Benach, J., *et al.* (2000) Structure–function relationships in Drosophila melanogaster alcohol dehydrogenase allozymes ADHS, ADHF and ADHUF, and distantly related forms, *European Journal of Biochemistry*, **267**, 3613-3622.

Bernardi, G. (2005) Structural and evolutionary genomics — natural selection in genome evolution (review on nucleotide contents), *Elsevier Science*, 204:434.

Binkowski, T.A. and Joachimiak, A. (2008) Protein functional surfaces: Global shape matching and local spatial alignments of ligand binding sites, *BMC Structural Biology*, **8**, 45.

Birzele, F. and Kramer, S. (2006) A new representation for protein secondary structure prediction based on frequent patterns, *Bioinformatics*, **22**, 2628-2634.

Bishop, J.G., Dean, A.M. and Mitchell-Olds, T. (2000) Rapid evolution in plant chitinases: Molecular targets of selection in plant-pathogen coevolution, *Proceedings of the National Academy of Sciences*, **97**, 5322-5327.

Bouvier, B., *et al.* (2009) Shelling the Voronoi interface of protein–protein complexes reveals patterns of residue conservation, dynamics, and composition, *Proteins: Structure, Function, and Bioinformatics*, **76**, 677-692.

Brown, C.J., *et al.* (2002) Evolutionary rate heterogeneity in proteins with long disordered regions, *Journal of Molecular Evolution*, **55**, 104-110.

Bryson, K., *et al.* (2005) Protein structure prediction servers at University College London, *Nucleic Acids Research*, **33**, W36-W38.

Bulmer, M. (1991) The selection-mutation-drift theory of synonymous codon usage, *Genetics*, **129**, 897-907.

Chiusano, M.L., *et al.* (1999) Correlations of nucleotide substitution rates and base composition of mammalian coding sequences with protein structure, *Gene*, **238**, 23-31.

Chou, P.Y. and Fasman, G.D. (1974) Conformational parameters for amino acids in helical, beta-sheet, and random coil regions calculated from proteins, *Biochemistry*, **13**, 211-222.

Consortium, D.G. (2007) Evolution of genes and genomes in the Drosophila phylogeny, *Nature*, **450**, 203-218.

Creighton, T.E. and Darby, N.J. (1989) Functional evolutionary divergence of proteolytic enzymes and their inhibitors, *Trends in Biochemical Sciences*, **14**, 319-324.

Dean, A.M., *et al.* (2002) The pattern of amino acid replacements in α/β-Barrels, *Molecular Biology and Evolution*, **19**, 1846-1864.

Dudgeon, K., Famm, K. and Christ, D. (2009) Sequence determinants of protein aggregation in human VH domains, *Protein Engineering Design and Selection*, **22**, 217-220.

Edgar, R.C. (2004) MUSCLE: Multiple sequence alignment with high accuracy and high throughput, *Nucleic Acids Research*, **32**, 1792-1797.

Ferrada, E. and Wagner, A. (2008) Protein robustness promotes evolutionary innovations on large evolutionary time-scales, *Proceedings of the Royal Society B: Biological Sciences*, **275**, 1595-1602.

Hanada, K., Gojobori, T. and Li, W.-H. (2006) Radical amino acid change versus positive selection in the evolution of viral envelope proteins, *Gene*, **385**, 83-88.

Holloway, A.K., *et al.* (2008) Accelerated sequence divergence of conserved genomic elements in Drosophila melanogaster, *Genome Research*, **18**, 1592-1601.

Hughes, A.L. and Nei, M. (1988) Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection, *Nature*, **335**, 167-170.

Jones, D.T. (1999) Protein secondary structure prediction based on position-specific scoring matrices, *Journal of Molecular Biology*, **292**, 195-202.

Kabsch, W. and Sander, C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features, *Biopolymers*, **22**, 2577-2637.

Kaur, H. and Raghava, G.P.S. (2002) An evaluation of β-turn prediction methods, *Bioinformatics*, **18**, 1508-1514.

Kimura, M. (1968) Evolutionary rate at the molecular level, *Nature*, **217**, 624-626.

Knight, R., *et al.* (2007) PyCogent: a toolkit for making sense from sequence, *Genome Biology*, **8**, R171.

Koehl, P. and Levitt, M. (1999) Structure-based conformational preferences of amino acids, *Proceedings of the National Academy of Sciences*, **96**, 12524-12529.

Komar, A.A. (2009) A pause for thought along the co-translational folding pathway, *Trends in Biochemical Sciences*, **34**, 16-24.

Kosiol, C., *et al.* (2008) Patterns of positive selection in six Mammalian genomes, *PLoS Genet*, **4**, e1000144.

Kyte, J. and Doolittle, R.F. (1982) A simple method for displaying the hydropathic character of a protein, *Journal of Molecular Biology*, **157**, 105-132

Larracuente, A.M., *et al.* (2008) Evolution of protein-coding genes in Drosophila, *Trends in genetics : TIG*, **24**, 114-123.

Lin, Y.-S., *et al.* (2007) Proportion of solvent-exposed amino acids in a protein and rate of protein evolution, *Molecular Biology and Evolution*, **24**, 1005-1011.

Lindsay, H., *et al.* (2008) Pitfalls of the most commonly used models of context dependent substitution, *Biology Direct*, **3**, 1-17.

Liu, J., *et al.* (2008) Natural selection of protein structural and functional properties: a single nucleotide polymorphism perspective, *Genome Biology*, **9**, R69.

Marcelino, A.M.C. and Gierasch, L.M. (2008) Roles of β-turns in protein folding: From peptide models to protein engineering, *Biopolymers*, **89**, 380-391.

Mondragón-Palomino, M., *et al.* (2002) Patterns of Positive Selection in the Complete NBS-LRR Gene Family of Arabidopsis thaliana, *Genome Research*, **12**, 1305-1315.

Montgomerie, S., *et al.* (2006) Improving the accuracy of protein secondary structure prediction using structural alignment, *BMC Bioinformatics*, **7**, 301.

Moult, J., *et al.* (1997) Critical assessment of methods of protein structure prediction (CASP): Round II, *Proteins*, **1**, 2-6.

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**, 715-724.

O'Farrell, H., *et al.* (2008) Sequence and structural evolution of the KsgA/Dim1 methyltransferase family, *BMC Research Notes*, **1**, 108.

O'Neil, K.T. and DeGrado, W.F. (1990) A thermodynamic scale for the helix-forming tendencies of the commonly occurring amino acids, *Science*, **250**, 646-651.

Pascarella, S. and Argos, P. (1992) Analysis of insertions/deletions in protein structures, *Journal of Molecular Biology*, **224**, 461-471

Petersen, L., *et al.* (2007) Genes under positive selection in Escherichia coli, *Genome Research*, **17**, 1336-1343.

Pollard, D.A., *et al.* (2006) Widespread discordance of gene trees with species tree in Drosophila: Evidence for incomplete lineage sorting, *PLoS Genet*, **2**, e173.

Sanner, M.F., Olson, A.J. and Spehner, J.C. (1996) Reduced surface: an efficient way to compute molecular surfaces, *Biopolymers*, **38**, 305-320.

Shepherd, A.J., Gorse, D. and Thornton, J.M. (1999) Prediction of the location and type of β-turns in proteins using neural networks, *Protein Science*, **8**, 1045-1055.

Subramanian, A., Kaufmann, M. and Morgenstern, B. (2008) DIALIGN-TX: Greedy and progressive approaches for segment-based multiple sequence alignment, *Algorithms for Molecular Biology*, **3**, 6.

Thomas, J.W., *et al.* (2003) Comparative analyses of multi-species sequences from targeted genomic regions, *Nature*, **424**, 788-793.

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**, 4673-4680.

Wong, K.M., Suchard, M.A. and Huelsenbeck, J.P. (2008) Alignment uncertainty and genomic analysis, *Science*, **319**, 473-476.

Wong, W.S.W., *et al.* (2004) Accuracy and power of statistical methods for detecting adaptive evolution in protein coding sequences and for identifying positively selected sites, *Genetics*, **168**, 1041-1051.

Wright, F. (1990) The 'effective number of codons' used in a gene, *Gene*, **87**, 23-29.

Yang, Z. (2007) PAML 4: Phylogenetic Analysis by Maximum Likelihood, *Molecular Biology and Evolution*, **24**, 1586-1591.

Yang, Z. and Nielsen, R. (2008) Mutation-selection models of codon substitution and their use to estimate selective strengths on codon usage, *Molecular Biology and Evolution*, **25**, 568-579.

Yang, Z., *et al.* (2000) Codon-Substitution Models for Heterogeneous Selection Pressure at Amino Acid Sites, *Genetics*, **155**, 431-449.

Yang, Z., Wong, W.S.W. and Nielsen, R. (2005) Bayes empirical bayes inference of amino acid sites under positive selection, *Molecular Biology and Evolution*, **22**, 1107-1118.

Yap, V.B., *et al.* (2010) Estimates of the effect of natural selection on protein-coding content, *Molecular Biology and Evolution*, **27**, 726-734.

Zhang, H., *et al*. (2008) Sequence based residue depth prediction using evolutionary information and predicted secondary structure, *BMC Bioinformatics*, **9**, 388.

6. POSITIVE SELECTION IN PROTEIN SECONDARY STRUCTURES: MAMMALS AND FUNGI

KATE E. RIDOUT

6.1 INTRODUCTION

The work detailed in Chapter 5 concerns the investigation into directional selection within the *Drosophila* phylogeny, particularly with regards to protein secondary structure, spatial organisation of selected sites and position along the gene (Ridout, et al., 2010). It was concluded that the distributions of selected sites in protein secondary structures are likely to vary due to various physiochemical properties of that structure, such as hydropathy, or charge. The clustering of selected sites along the length of a gene was found to be predominately the result of mutations on the protein structure and function, which we referred to as compensatory mutations. Finally, the increase of selected sites at gene ends was thought to be a result of relaxed purifying selection, a higher proportion of β-turn and coil residues at the ends of genes and a greater concentration of hydrophilic residues. These trends and explanations presented in Chapter 5 are all features that could potentially be consistent with other phyla and kingdoms. To test this hypothesis, similar analyses were performed on two data sets: gene alignments from six mammalian genomes and six fungal genomes. In mammals, it is known that proteins with stronger structural constraints experience stronger purifying selection (Liu, et al., 2008), suggesting that at least the structural result detected in *Drosophila* might persist in mammals. In fungi, very little is known about the correlation between protein secondary structure and evolvability.

Microbotryum violaceum is a smut fungus with no closely related species with sequenced genomes at the time of writing. It represents a large and poorly studied group of basidiomycetes, *Puccinomycotina*, which is mainly comprised of rust and smut fungi. As discussed in previous chapters, *M. violaceum* isolates from different hosts have been shown to be different species, and so we adopt the same naming conventions adopted in previous chapters. Using genes from these species is an ideal opportunity to investigate the effects of positive selection in a novel fungal

genome. Although the number of genomes sequenced is rapidly increasing, gene sets that are available from multiple species with a divergence low enough to perform evolutionary analysis on are not common, and are often restricted to model organisms such as the mammalian gene sets discussed below, or the *Drosophila* genes used in the previous chapter.

Genes from seven predictions and alignments for seven different species of *Microbotryum* were described in Chapter 4 and aligned to one another. Upon examining the concatenated gene tree of these species (Figure 4.7) it was apparent that the extremely long branch length of *M. violaceum* var. *caroliniana* might give rise to 'long-branch attraction' (Kolaczkowski and Thornton, 2009) in the Bayesian tree, and hence resulting the misplacement of the branch. It is therefore possible that this species might represent the root of the species tested. Alternatively, various evolutionary and demographic events might have lead to the high divergence of this species with regards to other *Microbotryum* species with *Silene* host species. With this in mind, it was decided that the species *M. violaceum* var. *caroliniana* would be left out of this analysis until the reason for the long-branch length and the exact positioning of the species in the phylogeny could be determined. From the analysis in Chapter 4; genes from six species remained: *M. violaceum* var. *latifolia*, *M. violaceum* var. *dioica*, *M. violaceum* var. *flos-cuculi*, *M. violaceum* var. *vulgaris*, *M. violaceum* var. *Saponaria* and *M. violaceum* var. *Dianthus*. A total of 2,987 genes could be obtained for all of these species. It was determined in Chapter 4 that *Microbotryum* may comprise up to 10,000 genes, however, when collecting genes from multiple species a great deal of these were filtered out of the final data set. Genes were only included in this analysis if they appeared in all six species and if they were at least 90% similar to one another, as calculated by the program Exonerate. Genes were also removed where there was a stop codon in any species. This filtering was extremely conservative, and was carried out to minimise the inclusion of pseudogenes, false exons or false intron positions, which might lead to false inferences of positive selection. Due to

the nature of the filtering, genes that were rapidly evolving, or had become shorter (in relation to genes from *M. violaceum* var. *caroliniana*) were removed from the analysis. Portions of genes that had lengthened in other species were also not included. To ensure that the results were as accurate as possible, genes in the data represented 1:1 orthologues only.

Positive selection was recently analysed in six mammalian genomes by Kosiol et al. (2008). All six genomes (human, chimpanzee, macaque, mouse, rat and dog) were sequenced with high coverage and considered to be high quality. As mentioned by the authors, previous studies have suffered with lack of power when searching for positive selection in mammals. The depth of the phylogeny in this study allowed for greater power over these previous investigations when detecting positively selected genes. As with the *Drosophila* analysis, using model organisms is expected to improve the accuracy of results due to accurate gene annotations and alignments, known protein structures and a well resolved phylogeny.

Previous results from the *Drosophila* analysis suggested that the more disordered secondary structures, coils and β-turns, experienced more instances of positive selection than the more ordered helix and strand regions, due to the physiochemical restraints on different structures. The naming of secondary structures used here is consistent with that used in Chapter 5, and is based on the Dictionary of Protein Secondary Structure (DSSP) naming conventions (Kabsch and Sander, 1983). Because these results were based on the physiochemical properties of the structures themselves, which are not specific to *Drosophila*, the same trend was expected in the mammalian and *Microbotryum* data sets. The clustering of selected sites detected in *Drosophila* (i.e. a given selected site was more likely to be with 30 amino acids of another selected site than expected by chance) was found to be independent of protein secondary structure and thought to

be due to compensatory mutations (Ridout, et al., 2010). Given this, it is expected that this result should persist in the genomes of other species; here we expect positively selected sites to be closer to one-another than expected by the random placement of selected sites within genes. Instances of positively selected sites were found at gene ends in *Drosophila*. Genes from the mammal data set had been truncated at the ends where quality scores were low, however the increase was gradual, and so it should still be possible to detect this result in the data. The stringent filtering of the *Microbotryum* gene set minimises the possibility that genes might have been extended over the gene start and stop codons, however it is possible that a number of genes might have been truncated, or represent single exons only. It is also possible that exons from different genes might have been forced together, and non-coding regions classified as introns.

Here, we examine secondary structure in the genomes of six mammalian species and six isolates from a fungal species complex. Species are examined for broad evolutionary trends such as those found in the previous analysis of the *Drosophila* genome. We find that many of the results are consistent across the species examined. The distribution of selected sites in secondary structures is qualitatively similar between mammals, insects and fungi, as is the clustering of selected sites. The increased proportion of selected sites at the gene ends detected in *Drosophila* is apparent here in *Microbotryum*, and to a lesser extent mammals.

6.2 METHODS

6.2.1 DATA

Nucleotide alignments of 16,549 Mammalian genes were obtained from http://compgen.bscb.cornell.edu/projects/mammal-psg/geneset-masked-nodups-v4.tgz, for the genomes of human, chimpanzee, macaque, mouse, rat and dog, of which 5,588 genes were present in all 6 species. Genomes were considered to be of near-optimal divergence for determining positive selection; balancing gene conservation with a well resolved phylogenetic signal (Anisimova and Yang, 2007; Kosiol, et al., 2008). The Mammalian gene set was predicted by Kosiol et al., (2008) through rigorous comparison of well known annotated human genes to the remaining 5 genomes, therefore reducing the effects of incorrect annotations found in gene databases for other species. Low quality bases were removed from the data set, which was rigorously filtered by Kosiol et al., (2008); therefore results are expected to be conservative. The highly accurate nature of the annotations makes this data set an ideal candidate for testing the distribution of positively selected sights and the structure of a gene. The phylogeny of the genomes used is well resolved; therefore a single tree was used for the analysis.

The genomes used in the fungal analysis belong to the *Microbotryum violaceum* species complex. The six strains used in this study were taken from *Microbotryum violaceum* species from six different plant taxa. The host species were *Silene latifolia*, *Silene dioica*, *Lychnis flos-cuculi* (also called *Silene flos-cuculi*), *Silene vulgaris, Saponaria officianalis* and *Dianthus carthusianorum*. A total of 2,987 genes were identified in all six species and aligned using ClustalW (Larkin, et al., 2007; Thompson, et al., 1994). Due to the uncertainty of the species tree, a concatenated gene tree was produced for all 2,987 genes and used in place of the species tree. The un-rooted

concatenated tree was reconstructed using a Bayesian approach in MrBayes (Huelsenbeck and Ronquist, 2001; Ronquist and Huelsenbeck, 2003) using the GTR+Γ model of DNA evolution. The following tree topology was recreated with 100% branch support and was labelled based on host species:

((*Dianthus*, *Saponaria*), (*S. vulgaris*, (*Lychnis*, (*S. dioica*, *S. latifolia*))))

Genome divergence was well within the acceptable range for investigations into positive section (Chapter 4). Genes were identified in Chapter 4 through conservative gene prediction methods, and as already noted, might be truncated at the termini. GO term enrichment analysis was carried out using Fisher's Exact Test within BLAST2GO (Conesa, et al., 2005; Götz, et al., 2008) on genes identified as being under selection in the following analysis, against the 9,919 genes identified in *M. violaceum* var. *latifolia*. This was carried out to ensure that any results produced from the selected genes were not biased due to protein function.

Genes were examined for insertions and deletions, as it is known that these sites are enriched in reverse turn and coil structures (Pascarella and Argos, 1992). Sites where gaps occurred (caused by insertions and deletions) in more than one species were removed from all species in all data sets, to avoid this bias. Sites containing ambiguous (masked) data were also removed from the analysis, following the method of Larracuente, et al., (2008).

6.2.2 DETERMINATION OF SECONDARY STRUCTURE

All protein sequences were downloaded from the RCSB (Research Collaboratory for Structural Bioinformatics) Protein Data Bank (PDB) and the protein sequences aligned to genes using BlastX with an *E* value cutoff of 10^{-6}. The top hit from each BLAST report was taken. At the time of writing, the PDB consisted of 183,503 sequences, which were aligned to the human genes identified as being positively selected and the *M. violaceum* var. *latifolia* genes also experiencing positive selection. These species were chosen as they represented the most accurately annotated genes in their respective data sets and should therefore have the most reliable structures.

As with previous analysis, secondary structures were computationally predicted using the program PSIPRED to determine sheet, helix and coil residues. BTPRED was then run to separate the coil category into unstructured coils and turn residues. These predictions resulted in 108,568 structural residues identified in the mammalian genes and 37,253 residues in the *Microbotryum* genes.

6.2.3 INFERENCE OF POSITIVE SELECTION

Positively selected genes and sites were identified in the mammalian and *Microbotryum* gene sets. Positive selection was determined using a Likelihood Ratio Test (LRT) on two sets of nested models in the codeml program from the phylogenetic analysis package PAML 4.4 (Yang, 2007). The previous study (Ridout, et al., 2010) was carried out using two different sets of nested models; M1a/M2a and M7/M8. Results did not qualitatively differ between these models, therefore the following analysis contains comparisons between the M7/M8 set only. The LRT

tests whether the model with positive selection (M8) fits the data significantly better than the model that does not allow for positive selection (M7). Sites from genes predicted to be under selection using the LRT were examined in the analysis. Codons with $\omega > 1$, and the probability of selection (P_s) ≥ 0.9 were considered to be undergoing positive selection. The chi squared test was used to compare observed and expected results throughout the data for statistical significance, and should be considered the method used unless otherwise specified. Previously, the concern was raised that the codon models used to calculate ω might be affected by structure; based on the work of (Lindsay, et al., 2008) and (Yap, et al., 2010). This theory was tested for effects on the analysis of positive selection in protein structures (Ridout, et al., 2010) and found to not alter the results, therefore this concern is not tested here.

In the previous study it was noted that secondary structures might have different degrees of codon bias. An increase in the proportion of rare or biased codon usage in a given structure could decrease the number of synonymous substitutions at synonymous sites (dS), relative to non-synonymous mutations at non-synonymous sites (dN). This would lead to an artificial inflation of ω (dN/dS) in that structure and thus a false inference of positive selection. The previous analysis of *Drosophila* (Chapter 5) revealed no correlation between codon bias and protein secondary structure, however, as codon bias varies between organisms we examine the effective number of codons (Wright, 1990) in the four secondary structures in mammals and *Microbotryum*. Mean ω was recorded in each of the four structures.

6.2.4 SPACING OF SELECTED SITES

Positively selected sites in the genomes of *Drosophila* were found to cluster together, i.e. a greater number of selected sites were found within 30 amino acids of one another across a gene than expected by chance (Ridout, et al., 2010). To test whether this result was specific to *Drosophila*, or could be found in these other datasets, the distance between selected sites in both the mammalian and *Microbotryum* genes were recorded and tested against an expected geometric distribution.

6.2.5 UNEQUAL DISTRIBUTION OF SELECTED SITES ALONG GENE LENGTHS

A greater number of positively selected sites were found towards the gene ends (which we consider here to be the start and end of the protein) in *Drosophila* than expected by chance. To test for this same phenomenon in the two data sets examined here, each gene (coding regions only) was divided into 20 equal segments consistent with previous analysis (Ridout, et al., 2010), and the number of positively selected sites and non-selected sites counted in each segment. Proportions of the different secondary structures were also found to vary along the length of the gene in the previous analysis. Therefore, the proportions of each secondary structure within each of the 20 genomic segments were also recorded.

6.2.6 SELECTION AT INTRON BOUNDARIES

Microbotryum genes were predicted, filtered and aligned thoroughly in Chapter 4, providing information on splicing and the position of introns within each selected gene. Parmley, et al.,

(2007) examined the rate of evolution at intron boundaries in mammals, and determined that this rate was lowest around splice sites, which was shown to be in part due to efficient purifying selection maintaining optimal splice enhancers. It is known that various aspects of splicing differ across distantly related species, for example splice sites and the amount of alternative splicing, and between intron-rich and intron-poor species (Irimia, et al., 2009). Mammals are considered intron-rich species with an average of around 5 introns per gene (Csuros, et al., 2011; Logsdon Jr, 1998), and so it is possible that the uneven distribution of structures with distance from the nearest intron might give the suggestion of the proportion of selected sites varying with secondary structure. *Drosophila* have an average of 1.7 introns per gene (Csuros, et al., 2011), and *Microbotryum* around 1.4 introns per gene (Chapter 4), therefore a similar trend might not be so pronounced, but may still be present. To examine the relationship between selected sites, structure and distance from the nearest intron, the distance of each positively selected site from the nearest intron was recorded in *Microbotryum*. Additionally, the distance of every residue in the protein from the nearest intron was recorded, along with the secondary structure of that residue.

6.3 RESULTS

6.3.1 SELECTION IN SECONDARY STRUCTURES

A total of 269 genes were found to be positively selected in mammals, and 88 genes in *Microbotryum*. After correction for false discovery rate (Blüthgen, et al., 2005), no GO terms were found to be significantly over or under represented in the *Microbotryum* selected gene data set. Of the 269 selected mammalian genes, 144 matched portions of experimentally determined structures, resulting in 39,746 structure residues, of which only 113 were considered to be under selection with the extremely low value of $P_s \geq 0.5$ and just 23 residues where $P_s \geq 0.9$ using

model M8 and genes with a significant M7/M8 LRT. Selected sites were not found to differ significantly in their distribution across secondary structures ($P = 0.07$ with M8 $P_s \geq 0.5$). The 23 selected residues meant that there was not sufficient data to perform the chi squared calculation. The *Microbotryum* genes contained even fewer selected sites with a known experimentally determined structure (10,931), resulting in 10 residues identified as being under selection with $P_s \geq 0.5$ and only 3 residues where $P_s \geq 0.9$. Once again, these values were not sufficient for the chi squared test. Results of the experimentally determined structures are therefore of little help here. The previous analysis of *Drosophila* demonstrated that selected sites in the experimentally determined structures followed the same broad trends as with the predicted structures; therefore, it is reasonable to assume that the predicted genes determined here will also provide sufficiently accurate results.

The predicted structure data sets comprised 108,568 residues in the mammalian data set and 37,253 residues of the *Microbotryum* genes (Table 6.1). For both data sets the distribution of selected sites in secondary structures was not random (chi squared test: $P < 0.00001$ where $P_s \geq 0.9$; $P = 0.002$ where $P_s \geq 0.99$ (mammals), and $P < 0.00001$ where $P_s \geq 0.9$; $P < 0.05$ where $P_s \geq 0.99$ (*Microbotryum*)). Helices were less likely to contain positively selected sites than expected in both study species (Figure 6.1). Also consistently in the three study species, coils contained more positively selected sites than expected. Strands were expected to contain fewer sites under selection than expected; this result was detected in the mammalian data set where $P_s \geq 0.9$. The *Microbotryum* genes contained fewer selected sites than expected by chance in strands where $P_s \geq 0.99$ and a slightly higher number than expected where $P_s \geq 0.9$. Where $P_s \geq 0.9$ in the mammalian data set, error bars did not overlap; this might suggest that more power is needed to resolve a similar result at the higher probability. Selected sites in β-turns were found to be under-

represented. The proportion of selected sites in β-turns in *Microbotryum* was similar to the expected value.

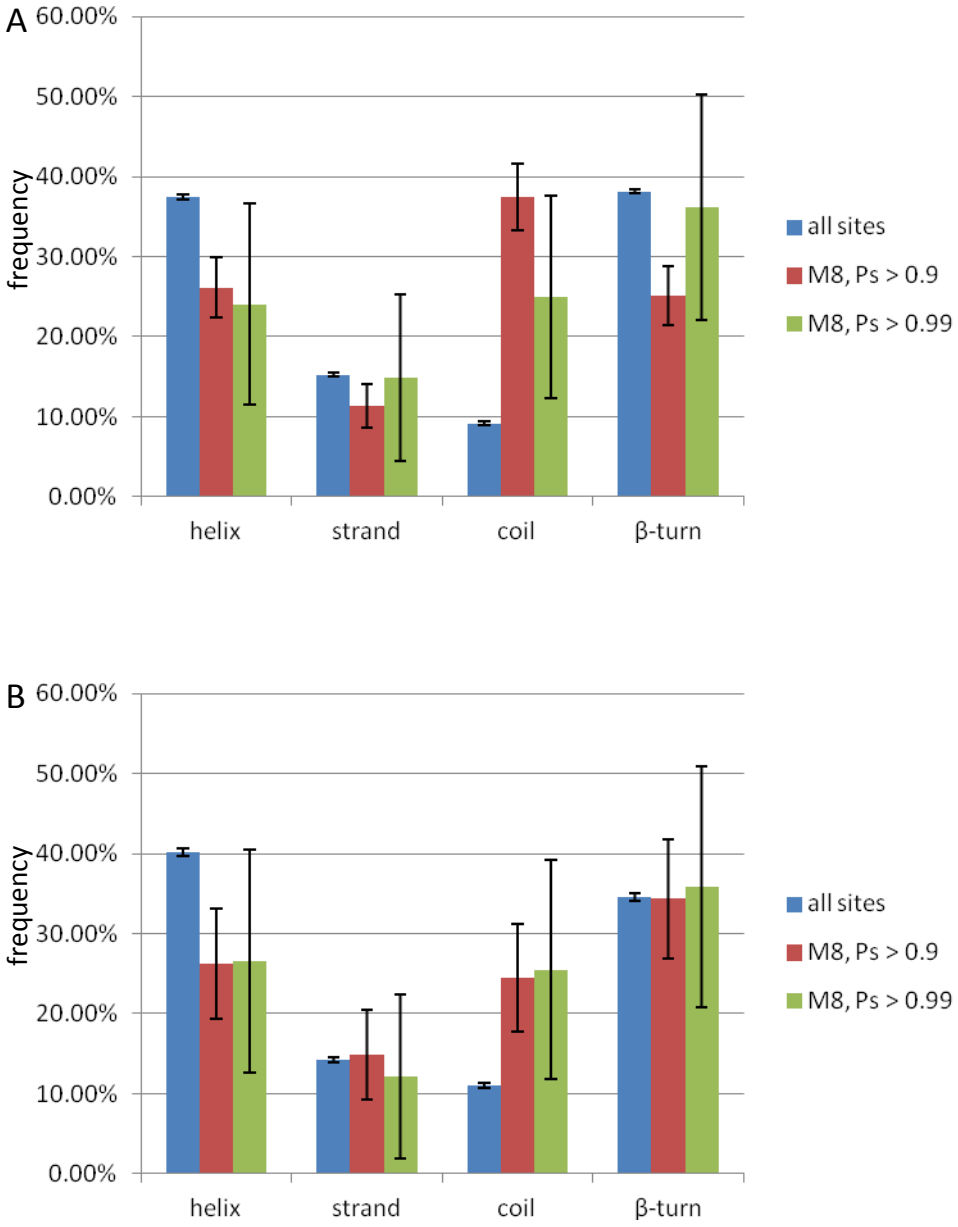

Figure 6.1. Proportions of all sites in selected genes (blue bars) in different secondary structures and selected sites in different secondary structures in A) mammals and B) *Microbotryum*.

Table 6.1. Positively selected sites in predicted secondary structures under model M8. Percentages represent the proportion of sites in all genes with a significant LTR. Total sites are only for genes with a positive LRT.

	Mammals					***Microbotryum***				
	Total Sites	**$P_s \geq 0.9$**	**$P_s \geq 0.99$**	**ENC**	**Mean ω**	**Total Sites**	**$P_s \geq 0.9$**	**$P_s \geq 0.99$**	**ENC**	**Mean ω**
Helix	40,675	138.98 (0.34%)	10.79 (0.027%)	55.33	0.263	14,948	41.25 (0.28%)	10.33 (0.069%)	46.89	0.247
Strand	16,514	60.40 (0.37%)	6.67 (0.040%)	55.20	0.300	5,314	23.36 (0.44%)	4.20 (0.079%)	46.92	0.257
Coil	9,963	198.94 (2.00%)	11.18 (0.11%)	55.56	0.319	4,133	38.53 (0.93%)	9.91 (0.240%)	49.77	0.294
β-Turn	41,416	133.13 (0.32%)	16.20 (0.039%)	55.55	0.327	12,858	53.99 (0.42%)	13.95 (0.108%)	48.65	0.318
Total	108,568	531.46 (0.49%)	44.84 (0.041%)	55.47	0.302	37,253	157.14 (0.42%)	38.88 (0.100%)	48.14	0.279

To investigate the effects of codon bias on the distribution of selected sites in secondary structures, the effective number of codons (Wright, 1990) was recorded in each of the four structures. Although codon bias is different (where smaller numbers represent more strongly biased usage) in mammals (average: 55.47) and *Microbotryum* (average: 48.14), the same trend can be seen in both analyses. Helix and strand structures, which experience a lower proportion of positively selected sites had stronger codon bias than the less ordered structures – the opposite to what we might expect if the result were caused by codon bias. This result suggests that secondary structure is correlated with codon bias, whereby the most ordered structures experience the strongest codon bias, regardless of positively selected sites. Mean ω was calculated in all structures to examine the relationship between ω and secondary structure. Mammals had the highest overall mean ω, at 0.302, and helix and sheet regions had lower mean ω values and coil and β-turn higher. *Microbotryum* genes had a loxer overall mean ω (0.279), and followed the same distribution of ω between structures. Within an organism, codon bias was stronger where ω was lower, however, this correlation did not persist between species. Consistently higher ω in the

less ordered structures support the hypothesis that these structures can physiochemically afford more mutations without disrupting the structure and therefore experience slightly relaxed purifying selection.

6.3.2 SPACING OF SELECTED SITES

If the spacing of selected sites was random, the distribution of the intervals between sites (of a spliced gene) should follow a geometric distribution. Examining the spacing of selected sites in *Drosophila* revealed that this was not the case; the likelihood of a site being selected decreased with distance from the nearest selected site. In order to test for this effect in mammals and *Microbotryum* the distances from each selected site to its nearest adjacent selected site was measured. In both data sets, the likelihood of a selected site decreased with distance from the nearest selected site (Figure 6.2).

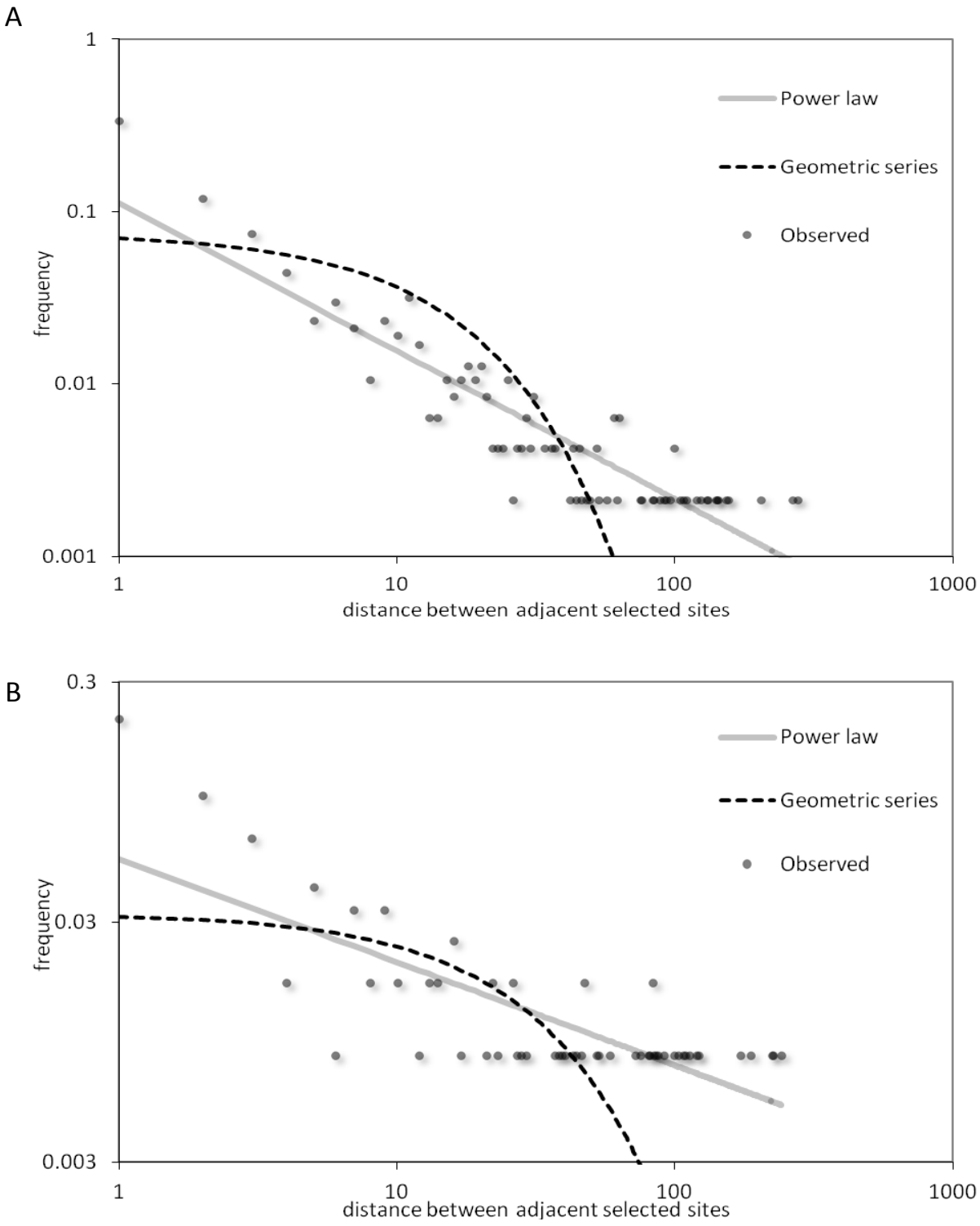

Figure 6.2. The distribution of sizes of the intervals between adjacent selected sites on a log-log scale, with a geometric curve expected with no clustering of sites, and a power law fitted to the curve in A) mammals and B) *Microbotryum*. Sites are considered selected where $P_s \geq 0.9$.

6.3.3 SELECTION AT THE GENE ENDS

In *Drosophila*, the proportion of positively selected sites was higher towards the start and stop codons of the genes, which we label here the 'gene ends'. Numbers of positively selected sites were calculated along the length of each gene divided into 20 equal segments, in mammals and *Microbotryum* (Figure 6.3). *Microbotryum* genes contained a sharp increase in the proportion of positively selected sites at the ends of the genes (Figure 6.3B). Proportions of selected sites in mammalian genes do not obviously follow this pattern, although there is a distinct increase in the proportions of selected sites approaching the C-terminus. The slight increase in the proportion of selected sites at the N-terminus is offset by the large peak almost halfway along the genes.

It is possible that an increase in the proportions of β-turn and coil sites or a decrease in the proportion of helix and strand sites at the gene ends could cause the increase in the numbers of positively selected sites. The proportions of the four secondary structures were measured along each gene divided into 20 equal segments (Figure 6.4). Neither mammals (Figure 6.4A) nor *Microbotryum* (Figure 6.4B) displayed the same pronounced increase in ordered structures and decrease in the less ordered structures. Both data sets revealed a marginal increase in the proportions of helices, coils and β-turns at the gene ends. Mammals, and to a lesser extent *Microbotryum* revealed a decrease in the proportion of strand structures towards the termini over the remainder of the gene.

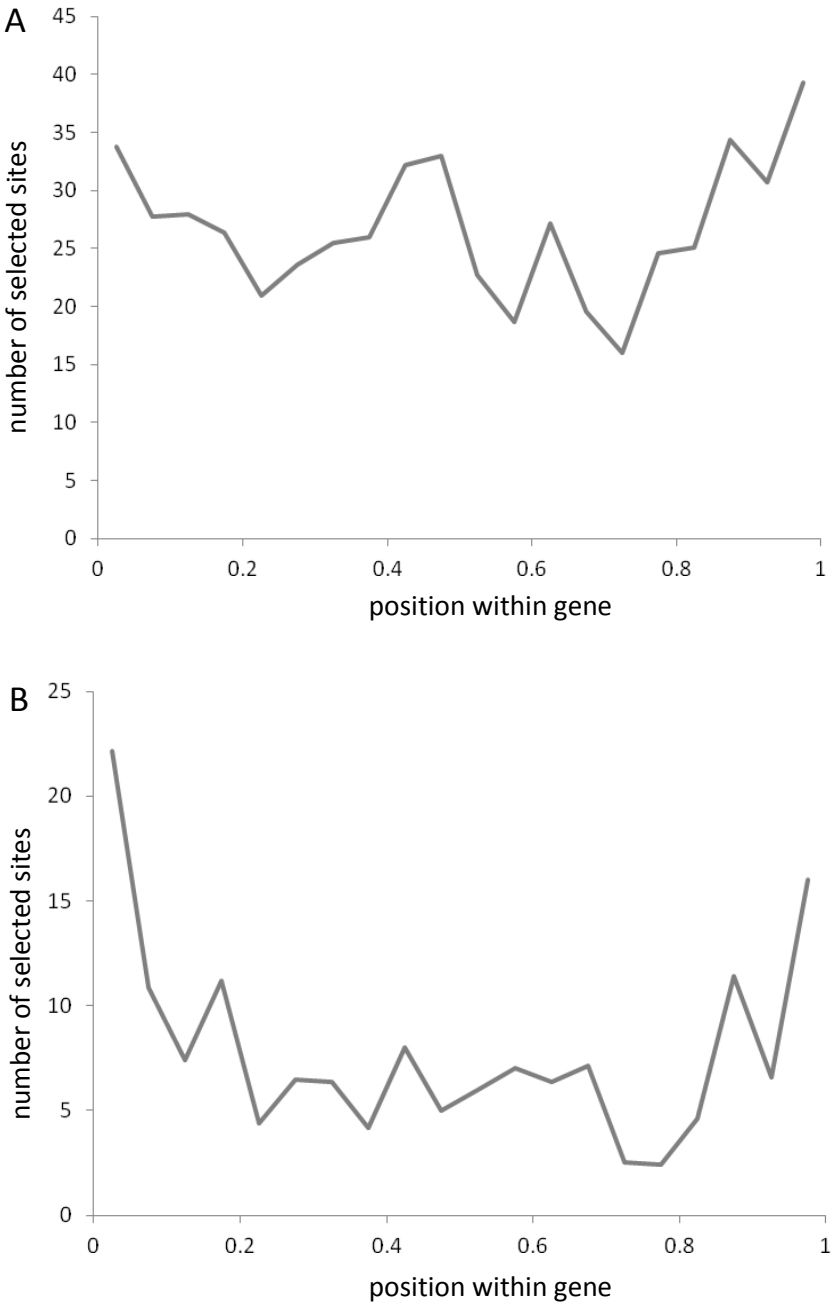

Figure 6.3. The distribution of selected sites along the length of the gene, when divided into 20 equal segments in A) mammals and B) *Microbotryum*. Zero marks the N-terminus and 1 the C-terminus. Sites are considered selected where $P_s \geq 0.9$.

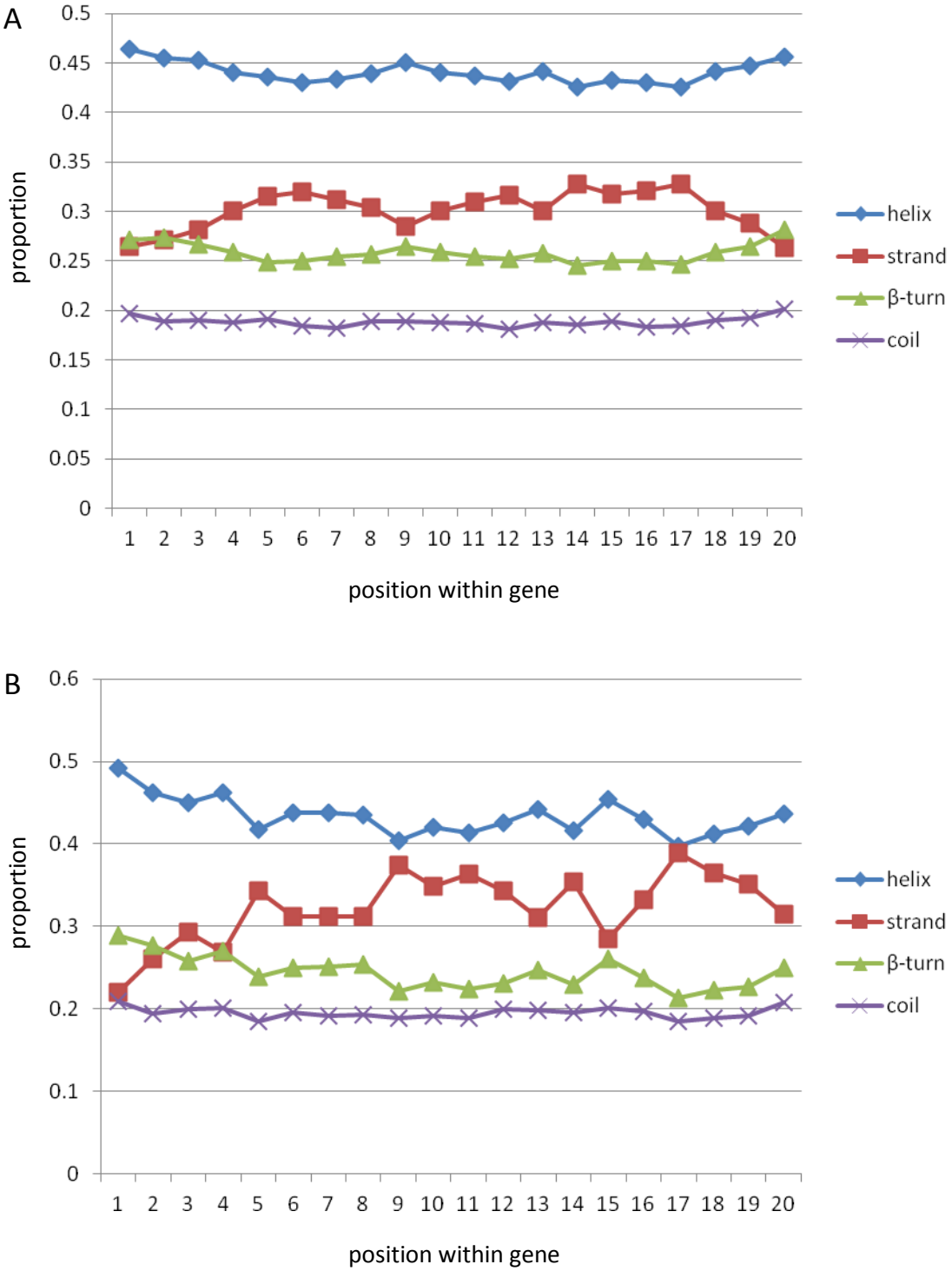

Figure 6.4. The distribution of secondary structures along the length of the gene, when divided into 20 equal segments in A) mammals and B) *Microbotryum*. Zero marks the N-terminus and 1 the C-terminus.

6.3.4 SELECTION AT INTRON BOUNDARIES

The distance of every selected site ($P_s \geq 0.9$) from the nearest intron in the selected *Microbotryum* genes was recorded. Genes containing no introns were not included in this analysis. Selected residues were found to be unevenly distributed with distance from the intron boundaries. Observed and expected numbers of selected sites were binned based on distance from the nearest intron in order to provide sufficient data for the chi squared test of significance. The bins of distances from the nearest intron were set to allow for an approximately equal number of selected sites in each category (19-20 residues). There was a significant deviation from the expected distribution of selected sites ($P < 0.05$). The numbers of selected sites were higher than expected in the first 3 bins (0-70 bp, 71-200 bp, 201-320 bp): (observed/expected) 19/14, 20/18, 19/13 respectively. When examining the first 320 residues from the nearest intron only, the distribution of selected sites does not differ significantly from the expectation. The remainder of the residues extend to a maximum of 1,865 bp from the nearest intron, and contain fewer selected sites than expected by chance: observed 19, expected 30.

The distances of the four secondary structures were recorded and compared within the same binned distances as the selected sites. The four distributions examined were not independent of one another, and are presented here to demonstrate general trends in the data. All four structures differed significantly from the expectation; where structures were distributed evenly (and independently) along a gene. Significantly fewer helixes than expected ($P < 0.00005$) were observed in the first four bins (where there were a greater number of selected sites) and more than expected in the final bin (where there were fewer selected sites). Sheet structures were distributed with significantly fewer sheets than expected in the first bin and more in the following 3 bins ($P < 0.00001$). Turns were distributed with a significantly larger number of residues than expected in the first 3 bins and fewer in the final bin ($P < 0.00001$). Finally, there were significantly fewer

coil residues than expected ($P < 0.001$) in the first and third bins, and more than expected in the second and final bins.

6.4 DISCUSSION

Proportions of positively selected sites in the three distant taxa of *Drosophila*, mammals and *Microbotryum* demonstrate similar trends with regard to distribution in different protein secondary structures; therefore it might be assumed that some of these phenomena are general features of protein-coding sequences across a broad spectrum of the tree of life. The investigation into *Drosophila* was performed to determine the relationship between positively selected sites and protein secondary structure on a genomic scale, and it was concluded that the more ordered the secondary structure, the fewer instances of positive selection that structure would experience (Ridout, et al., 2010). Selected sites in mammals and *Microbotryum* were also found to occur at a higher concentration than expected in disordered coil regions and in a lower concentration in the highly organized helixes. Strand regions, which are also highly ordered regions, contained fewer selected sites than expected in *Drosophila* and mammals. Selected sites in strands were neither under, nor over represented in *Microbotryum*, suggesting that either there are no effects of a strand structure on the paucity of selected sites, or, that the result was not strong enough to detect given the volume of data. Overall, we have demonstrated that the increase in the proportion of selected sites in unstructured regions over the more ordered structures persists across distantly related taxa; thus supporting the theory that the physiochemical properties of different secondary structures has a role in the rate of selection of a gene. Selected sites were not found to be significantly unevenly distributed in mammals using the experimentally determined data where sites were considered to be selected where $P_s > 0.5$, however, it is likely that the result was not detectable with this amount of data.

Selection in β-turn structures was more variable, however, the degree of accuracy of turn prediction is ~70% (Shepherd, et al., 1999); lower than the accuracy of helix, strand and coil prediction (~80%) (Moult, et al., 1997). The *Drosophila* crystal structure results, expected to be the most accurate with respect to annotating secondary structures, contained fewer selected sites in β-turn structures than expected. The predicted mammalian β-turn residues also contained fewer selected sites than expected. Selected sites in predicted β-turns in *Microbotryum* appear to be neither over nor under-represented; similar to the strand sites. This may be due to no effect of β-turns on proportions of selected sites, or it may be due to the lack of data and the uncertainty of β-turn prediction, however, there was no difference in the way β-turns were predicted between the taxa, so this is less likely. It was suggested in the previous chapter that the difficulty in crystallising β-turn residues that exist in the disordered outer regions of the protein leads to an under-representation of solvent-accessible, and thus selected, β-turn residues. It is not clear however, how this same result might have been generated in the mammalian predicted structures. Examining the structure of the sections of genes removed from the mammalian data by Kosiol et al., (2008) might also reveal whether β-turn residues at the ends of proteins were removed, which could cause a similar result; this effect would then be consistent with the lack of an increase in the proportion of selected sites at gene ends, as was expected from both the *Drosophila* and *Microbotryum* results. It has been shown that some turn structures play a more active part in folding than others (Marcelino and Gierasch, 2008), therefore we would expect some disparity in the conservation of selected sites in β-turns; the proportions of which within the data sets tested might result in the uncertainty that we see in our data.

Codon bias was examined here to determine whether an excess of rare codons or strong codon bias could have caused a false inference of positively selected sites. As with the *Drosophila* data, in both mammals and *Microbotryum* codon bias was found to be stronger in the more ordered

helix and strand structures than in the aperiodic coil and β-turn structures – the opposite to what we should see if codon bias or rare codons were causing the structure result. This trend therefore is consistent across large evolutionary distances, and as such, is likely to be due to the physiochemical nature of the structures themselves. One interesting continuation of this analysis might be to examine the GC content of the structures to determine whether this is the result of a structural nucleotide bias – for example, Chiusano, et., al (1999) found that nucleotide content varies with structure at the 3rd codon position in mammals – or whether biased codon usage in highly ordered structures offers some type of selective advantage.

The clustering of selected sites detected in *Drosophila* was found to be consistent across mammals and *Microbotryum*; selected sites tended to be spatially closer together than expected by chance. In the *Drosophila* analysis it was determined that this result was likely to be due to compensatory mutations, that is, selection on the protein function, as adjacent selected sites within 30 amino acids of one another were significantly more likely to occur on the same branch of the phylogeny than those selected sites that were further apart. The persistence of this result in both mammals and *Microbotryum* support this explanation, as the effect does not appear to be due to anything species-specific. Callahan, et al., (2011) performed a similar analysis in *Drosophila,* and determined that the clustering effect was only present at non-synonymous sites, also supporting the theory that this is due to selection on protein function, which we refer to as compensatory mutations (or epistasis as suggested in the study).

In *Drosophila,* the increase in the proportion of positively selected sites towards the gene ends was suggested to be the result of relaxed purifying selection at the ends of genes relative to the middle (Ridout, et al., 2010). It was also demonstrated that there was a higher proportion of coil

and β-turn residues at the ends of the genes relative to the middle, although this did not fully explain the result. A similar increase in the number of positively selected sites at the gene ends can be seen in *Microbotryum*, and we can see a slight increase in coil and β-turn residues and a decrease in strand residues. However, the opposite trend can be seen in helixes, which are increased at the gene ends relative to the centre. This supports the theory that the increase in the number of positively selected sites at gene ends is not simply due to an increase in disordered structures and a decrease in the more ordered structures. The mammalian genes follow a similar pattern of structure distribution, with a very slight increase in coils, β-turns and helixes at the gene ends and a decrease in strands. Selected sites are not so obviously increased at the gene ends; indeed, there is an obvious increase in selected sites at a peak around the middle of the gene. There are a number of possible explanations for this deviation; the first is that the conservative nature of the data set resulted in loss of the ends of the genes, as low quality bases were trimmed. A second explanation might be that the presence of introns in the genes is affecting the distribution of selected sites; however, as mammals contain a greater number of introns we would expect this to result in a decrease in the number of selected sites around the middle of the gene, due to strong purifying selection on splice enhancers (Parmley, et al., 2007). This effect, if it exists in *Drosophila* and *Microbotryum* would not be so pronounced, as both species tend to have fewer introns per gene; 1.7 in *Drosophila* and 1.4 in *Microbotryum* compared to around 5 in mammals. Another possible explanation is that Mammals simply do not experience such a strong increase in the proportion of selected sites at the gene ends. These results suggest that the increase in the number of positively selected sites at the gene ends in *Microbotryum* and – to a lesser extent – mammals is not caused solely by the variation in the proportion of different secondary structures.

As previously mentioned, mammalian genes contain a higher proportion of introns (an average of 5 per gene) than *Drosophila* (1.7 per gene) and *Microbotryum* (1.4 per gene), and it has been demonstrated that areas immediately surrounding introns experience a decrease in evolutionary rate in mammals (Parmley, et al., 2007). From the *Microbotryum* data set, we were able to determine that distance to the nearest intron had a significant effect on the distribution of selected sites along the gene length, as well as the distribution of secondary structures. Contrary to the results detected in mammals, a greater number of selected sites than expected by chance were found closer to the nearest intron, suggesting that internal exons (i.e. with an intron either side) contain fewer selected sites in the middle compared to the ends. This might indicate that selection on splice enhancers is ongoing in *Microbotryum*, or it might be that the binned distances are too wide to detect strong purifying selection around splice sites. It is possible that the excess of positively selected sites in disordered structures with regard to the expectation is due to the uneven distribution of structures and selected sites with distance from the nearest intron. However, if this were the case we might expect results to be far more pronounced in mammals where there is a greater density of introns than the other species, and this does not occur. It is also possible that the decrease in selected sites at a greater distance from an intron is due to the increase in the more ordered helix and strand structures. It is worth noting that there is the possibility that a small number of introns might instead be non-coding regions, where exons might have been forced together by prediction programs, however, this is unlikely, as only ~9% of genes predicted in *M. violaceum* var. *latifolia* were found to be chimeric (i.e. multiple real genes forced together) (Table 4.5). These results suggest that there is a relationship between the number of selected sites and distance from the nearest intron in *Microbotryum*. It also seems likely that secondary structures are not distributed evenly with regard to the distance from the nearest intron. It is not clear from the results whether this uneven distribution of selected sites is linked to the distribution of structures, as there was not sufficient data to perform a test for significance once the selected sites were divided into structure classes. It is possible that there

might be a higher proportion of disordered structures than expected around splice sites, as alternative splicing might disrupt protein structure. However, research into alternative splicing has revealed that a higher number of splicing events occur within highly conserved domains than expected, as around half of the proteins studied were in these regions. It was therefore suggested that splicing might provide more flexibility for structural change within these regions (Birzele, et al., 2008).

When examining the *Drosophila* genome, a number of factors were identified that might be linked to the proportion of selected sites varying with gene structure. The most notable is the relationship between hydrophobicity, solvent-accessibility and protein secondary structure. In the previous chapter, it was determined that the more solvent accessible surfaces and more hydrophilic residues experienced a higher rate of evolution, i.e. a greater proportion of selected sites in these areas than expected by chance. We also determined that coil and β-turn residues were often more hydrophilic and solvent-exposed/accessible than the more ordered structures. Although this is not tested in mammals or *Microbotryum*, we would expect this result to extend to these species, as hydrophobicity and solvent exposure is directly linked to the physiochemical properties of protein secondary structures, rather than the particular organism.

There are many factors that are known to affect the rate of evolution in a gene. Here we have demonstrated that the rate of evolution is affected by the distribution of protein secondary structures, the nearest selected amino acid and the position of the residue along the gene in insects, mammals and fungi. Suggestions have been made to take the analysis further, and to fully investigate the relationship between selected sites, secondary structures and distance from the nearest intron. By advancing the understanding in factors that affect the rate of evolution we can begin to fully characterise the evolution of genes and genomes.

6.5 REFERENCES

Anisimova, M. and Yang, Z. (2007) Multiple hypothesis testing to detect lineages under positive selection that affects only a few sites, *Molecular Biology and Evolution*, **24**, 1219-1228.

Birzele, F., Csaba, G. and Zimmer, R. (2008) Alternative splicing and protein structure evolution, *Nucleic Acids Research*, **36**, 550-558.

Blüthgen, N., *et al.* (2005) Biological profiling of gene groups utilizing Gene Ontology, *Genome Informatics*, **16**, 106-115.

Callahan, B., *et al.* (2011) Correlated evolution of nearby residues in Drosophilid proteins, *PLoS Genet*, **7**, e1001315.

Chiusano, M.L., *et al.* (1999) Correlations of nucleotide substitution rates and base composition of mammalian coding sequences with protein structure, *Gene*, **238**, 23-31.

Conesa, A., *et al.* (2005) Blast2GO: A universal tool for annotation, visualization and analysis in functional genomics research, *Bioinformatics*, **21**, 3674-3676.

Csuros, M., Rogozin, I.B. and Koonin, E.V. (2011) A detailed history of intron-rich Eukaryotic ancestors inferred from a global survey of 100 complete genomes, *PLoS Comput Biol*, **7**, e1002150.

Götz, S., *et al.* (2008) High-throughput functional annotation and data mining with the Blast2GO suite, *Nucleic Acids Research*, **36**, 3420-3435.

Huelsenbeck, J.P. and Ronquist, F. (2001) MRBAYES: Bayesian inference of phylogenetic trees, *Bioinformatics*, **17**, 754-755.

Irimia, M., *et al.* (2009) Complex selection on 5′ splice sites in intron-rich organisms, *Genome Research*, **19**, 2021-2027.

Kabsch, W. and Sander, C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features, *Biopolymers*, **22**, 2577-2637.

Kolaczkowski, B. and Thornton, J.W. (2009) Long-branch attraction bias and inconsistency in bayesian phylogenetics, *PLoS ONE*, **4**, e7891.

Kosiol, C., *et al.* (2008) Patterns of positive selection in six Mammalian genomes, *PLoS Genet*, **4**, e1000144.

Larkin, M.A., *et al.* (2007) Clustal W and Clustal X version 2.0, *Bioinformatics*, **23**, 2947-2948.

Larracuente, A.M., *et al.* (2008) Evolution of protein-coding genes in Drosophila, *Trends in genetics : TIG*, **24**, 114-123.

Lindsay, H., *et al.* (2008) Pitfalls of the most commonly used models of context dependent substitution, *Biology Direct*, **3**, 1-17.

Liu, J., *et al.* (2008) Natural selection of protein structural and functional properties: a single nucleotide polymorphism perspective, *Genome Biology*, **9**, R69.

Logsdon Jr, J.M. (1998) The recent origins of spliceosomal introns revisited, *Current Opinion in Genetics andvelopment*, **8**, 637-648.

Marcelino, A.M.C. and Gierasch, L.M. (2008) Roles of β-turns in protein folding: From peptide models to protein engineering, *Biopolymers*, **89**, 380-391.

Moult, J., *et al.* (1997) Critical assessment of methods of protein structure prediction (CASP): Round II, *Proteins*, **1**, 2-6.

Parmley, J.L., *et al.* (2007) Splicing and the evolution of proteins in mammals, *PLoS Biol*, **5**, e14.

Pascarella, S. and Argos, P. (1992) Analysis of insertions/deletions in protein structures, *Journal of Molecular Biology*, **224**, 461-471

Ridout, K.E., Dixon, C.J. and Filatov, D.A. (2010) Positive selection differs between protein secondary structure elements in Drosophila, *Genome Biology and Evolution*, **2**, 166-179.

Ronquist, F. and Huelsenbeck, J.P. (2003) MrBayes 3: Bayesian phylogenetic inference under mixed models, *Bioinformatics*, **19**, 1572-1574.

Shepherd, A.J., Gorse, D. and Thornton, J.M. (1999) Prediction of the location and type of β-turns in proteins using neural networks, *Protein Science*, **8**, 1045-1055.

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**, 4673-4680.

Wright, F. (1990) The 'effective number of codons' used in a gene, *Gene*, **87**, 23-29.

Yang, Z. (2007) PAML 4: Phylogenetic Analysis by Maximum Likelihood, *Molecular Biology and Evolution*, **24**, 1586-1591.

Yap, V.B., *et al.* (2010) Estimates of the effect of natural selection on protein-coding content, *Molecular Biology and Evolution*, **27**, 726-734.

7. DISCUSSION AND CONCLUDING REMARKS

KATE E. RIDOUT

7.1 DISCUSSION

At the start of this thesis I set out to investigate factors of gene composition that might affect the rate of evolution of a protein. To perform this large-scale analysis, large amounts of sequence data were required. Genes for multiple closely related species were already available for *Drosophila* and mammals; however, the basidiomycete fungus *Microbotryum* was in the process of being sequenced by the Filatov lab. Preparing this data for use would generate three outcomes. The first was creating a number of assembled genomes for further investigations into this poorly studied group. The second was the creation of a bioinformatics pipeline designed to automatically annotate the genomes of eukaryotes. The final outcome was to produce a set of annotated genes in a number of genomes of *Microbotryum* that could be used for multiple types of analysis, including the investigations into factors affecting genome evolution. Using these three data sets; *Drosophila*, mammals, and the bespoke *Microbotryum* data, gene compositional factors that might affect the rate of evolution of proteins could be compared between extremely diverse taxa spanning two kingdoms.

Sequencing and annotating the genomes of *Microbotryum* resulted in seven species gene sets that could be used for evolutionary analysis. Annotating these genomes lead to the creation of the self-training pipeline FRANC, designed to identify repetitive elements and predict genes and coding regions for novel eukaryotic sequence data. After annotation and filtering, the genomes of *Microbotryum* were found to contain a maximum of 10,000 genes, and the small genome size of ~25 mb contained minimal repetitive content, which was consistent with other fungal genomes of a similar size. The intron content of the genes was low (around 1.4 introns per gene), compared to a number of other basidiomycetes, (for example *Cryptococcus*: 5.3 introns per gene), but high compared to the basidiomycete smut *Ustilago* (0.46 introns). Interestingly, the genome of *M.*

violaceum var. *caroliniana* was far more divergent than expected based on the phylogeny of the plant host species. This suggested that either the genome has undergone far more rapid evolution than the other species, or alternatively, that the *Microbotryum* species do not match the phylogeny of the plant host species. With such large amounts of data now available for these species of *Microbotryum*, it will be possible to investigate the evolution of many attributes of this fungus, such as mating-type evolution and host-pathogen interactions. With a reference genome available, and a better understanding of which assembly techniques are appropriate for *Microbotryum*, it will also be possible to easily assemble further genomes as they are sequenced.

In the introduction to this thesis, I discussed a number of factors that are correlated with the rate of protein evolution, such as gene expression, gene length and nucleotide content. In the chapters that followed, a number of factors that might potentially affect the rate of protein evolution were examined. These included protein secondary structure, distance from the nearest selected site within a spliced coding region, protein hydropathy, solvent accessibility, position along the length of the gene and distance from the nearest intron. Many of these factors were found to be consistent across two kingdoms, suggesting that they are general trends, rather than species-specific phenomena.

Here, it has been demonstrated that the proportion of positively selected sites varies between protein secondary structures across three taxa in two kingdoms. More specifically, the greater degree of physiochemical order – where helix and strand regions are highly ordered and β-turn and coil regions disordered – the fewer the number of selected sites. This is the first time that this result has been demonstrated on large-scale data. Secondary structure and solvent-accessibility and charge have also been correlated, as previously demonstrated (Chou and Fasman, 1974;

Koehl and Levitt, 1999), with disordered structures generally containing more hydrophilic residues. It has been demonstrated that more hydrophobic proteins tend to be more stable (Banerjee, et al., 2006), which might suggest that stable proteins contain more ordered structures and should be more strongly conserved. It has also been demonstrated that more robust proteins can experience a faster rate of evolution, as a robust protein tolerates more mutations (Rorick and Wagner, 2010). Therefore, highly ordered stable proteins are less robust to mutation and thus evolve more slowly, which is supported by the results presented in this thesis. This would suggest that the effect of protein secondary structure on positive selection does not necessarily confer a selective advantage, and is instead the result of protein physiochemical properties. From the results discussed here, it cannot be said that adaptively evolving sites are selected because of the structure that they are in; rather, there is a higher likelihood of an adaptive mutation occurring because the structure might tolerate more mutations and so there is more variability from which to produce an adaptive mutation.

Investigations in Chapters 5 and 6 into selection at the gene ends has demonstrated a general trend across taxa and kingdoms that there is a higher proportion of variable sites at the ends of genes relative to the middle, and a higher proportion of sites predicted by PAML to be positively selected. It was also determined that there was a greater proportion of disordered structures towards the ends of genes in *Drosophila*, which also experienced relaxed purifying selection. This relaxation of selection may allow greater variability for adaptive evolution to work on, however, it also might result in a larger number of weakly selected mutations and false inferences of positive selection.

Relaxed purifying selection can lead to difficult problems to overcome when identifying sites under positive selection, due to the increase in weakly selected sites and false positives. When identifying positively selected sites using PAML this can be controlled for by the use of conservative models for detecting selection, and by only examining sites with a high probability of positive selection, though a number of false positives are inevitable. Regions of high variability in proteins have been referred to as highly evolvable (Ferrada and Wagner, 2008; Rorick and Wagner, 2010) due to the increased tolerance of mutations, which are then available to be selected, however, there will be an increase in weakly selected sites in these regions over areas with stronger purifying selection. It might be inferred that when an area that is otherwise highly conserved experiences positive selection, this site should be under very strong positive selection (perhaps more so than at the more evolvable sites) in order to overcome the restrictions of the region. Recently, Bazykin and Kondrashov (2011) have suggested a different method of detecting positive selection; through sites currently experiencing negative selection. This test assumes that the results of past positive selection would now be undergoing negative selection as a method of maintaining that change. This method is likely to predict different sites to PAML's codeml program, though both identify positively selected sites. One possible way to uncouple a relaxation of purifying selection with strongly positively selected sites might be to use this alternative method to test for positive selection in different regions such as different secondary structures, and at the gene ends, to see if the trends detected through PAML are maintained.

Sattath et. al., (2011) recently examined the rate of neutral substitutions around amino acid polymorphisms in *Drosophila simulans*. In this analysis they detected the same result demonstrated in Chapters 5 (Ridout, et al., 2010) and 6; that positively selected sites tend to cluster together, and are not evenly distributed along the length of a gene. It was suggested that this observation could be in part attributed to the variation in constraint and potentially variation

in adaptability of the gene. This is consistent with the extended complexity hypothesis presented by Aris-Brosou (2005), which states that proteins with complex interaction networks would be more conserved by evolution. This theory however, does not explain the tendency for selected sites to be clustered on the same branch of a phylogeny described in Chapter 5 and by Bazykin et. al., (2004), which would support the theory that the clustering of selected sites was due to compensatory mutations – a suggestion echoed by Callahan et al., (2011). It seems likely from these results that both mechanisms might play a role in the distribution of selected sites along the length of a gene.

Examining factors that affect the rate of protein evolution can help us to understand how and why proteins evolve. In this thesis I have used a large amount of novel genomic data combined with publicly available data stets to interrogate a number of protein physiochemical features that might affect this rate of evolution. In order to further understand this topic, instances where selective events occur despite heavy physiochemical disadvantages could be investigated. In addition, the proportions of mutations that overcome disadvantages of other factors affecting the rate of evolution should be examined. For example, if a highly conserved structure contained a high degree of selection for rare codons, it might be suggested that selection on codon bias exerted a stronger selective pressure than the degree of conservation of ordered structures. Given that a single site in a protein is simultaneously experiencing many factors that shape evolution – for example, structure, nucleotide content, solvent accessibility – to fully understand factors that shape the rate of evolution in proteins, the relative contributions of these factors in varying situations must be understood.

7.2 CONCLUDING REMARKS

The studies carried out in this thesis have contributed towards a better understanding of sequence assembly techniques, the creation of a genome annotation resource and the identification of genes in *Microbotryum* species. Research advancing the knowledge of factors affecting positive selection has also been carried out, demonstrating the effects of protein secondary structure on the likelihood of positive selection on a genomic scale for the first time. The clustering of selected sites, suggested here to be due to compensatory mutations, and the increase in the number of selected sites at the ends of genes – potentially due to relaxed purifying selection on protein termini – have been reported. These large trends have not only been supported by smaller studies, and since been repeated in *Drosophila*; it has also been demonstrated here that these trends are consistent across two separate kingdoms. Further studies might expand upon the reasons behind these findings, and evaluate the strength of these factors when considering other aspects of evolutionary rate heterogeneity. Every site in a protein has an associated structure, function, nucleotide, codon, hydropathy, charge, solvent accessibility, physical location (relative to introns and protein ends), and degree of interaction with other loci; and all of these factors have been shown here to affect the rate of evolution of a protein, along with others that have not been discussed. Although I have investigated factors affecting the rate of evolution and the potential reasons for these, we have not yet begun to evaluate their relative strengths, and we will not fully understand which (if any) of these factors are truly selected for and to what degree until we understand how these factors affect each other.

7.3 REFERENCES

Aris-Brosou, S. (2005) Determinants of adaptive evolution at the molecular level: The extended complexity hypothesis, *Molecular Biology and Evolution*, **22**, 200-209.

Banerjee, T., Gupta, S.K. and Ghosh, T.C. (2006) Compositional transitions between *Oryza sativa* and *Arabidopsis thaliana* genes are linked to the functional change of encoded proteins, *Plant Science*, **170**, 267-273.

Bazykin, G.A. and Kondrashov, A.S. (2011) Detecting Past Positive Selection through Ongoing Negative Selection, *Genome Biology and Evolution*, **3**, 1006-1013.

Bazykin, G.A., *et al.* (2004) Positive selection at sites of multiple amino acid replacements since rat-mouse divergence, *Nature*, **429**, 558-562.

Callahan, B., *et al.* (2011) Correlated evolution of nearby residues in Drosophilid proteins, *PLoS Genet*, **7**, e1001315.

Chou, P.Y. and Fasman, G.D. (1974) Conformational parameters for amino acids in helical, beta-sheet, and random coil regions calculated from proteins, *Biochemistry*, **13**, 211-222.

Ferrada, E. and Wagner, A. (2008) Protein robustness promotes evolutionary innovations on large evolutionary time-scales, *Proceedings of the Royal Society B: Biological Sciences*, **275**, 1595-1602.

Koehl, P. and Levitt, M. (1999) Structure-based conformational preferences of amino acids, *Proceedings of the National Academy of Sciences*, **96**, 12524-12529.

Ridout, K.E., Dixon, C.J. and Filatov, D.A. (2010) Positive selection differs between protein secondary structure elements in Drosophila, *Genome Biology and Evolution*, **2**, 166-179.

Rorick, M.M. and Wagner, G. (2010) Structural Robustness Confers Evolvability in Proteins, *AAAI Fall Symposium Series (FS-10-03)*, 110-119.

Sattath, S., *et al.* (2011) Pervasive adaptive protein evolution apparent in diversity patterns around amino acid substitutions in *Drosophila simulans*, *PLoS Genet*, **7**, e1001302.