Understanding lineage-specific biology through comparative genomics.

Yang Li

St Cross College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Trinity 2014

This thesis is dedicated to my mother and my father
for their patience, understanding and support.

Acknowledgements

First and foremost, I would like to thank Richard Copley for his supervision, encouragement, and honesty. Richard let me work independently on problems and helped me keep a global view of my projects without losing myself in meaningless details. Second, I am extremely grateful to Chris Ponting, who allowed me to work on exciting projects while ensuring that the quality of my work is of the highest standard. In terms of supervision, I am most thankful to Wilfried Haerty, for acting as a great supervisor and a friend. I would also like to thank other members of the Ponting lab who helped me through my doctoral studies with minimal amounts of doubt and despair. Thank you to all friends and climbers for keeping my life interesting outside of work. Thank you Will for the gym sessions. Last, but certainly not least, I wish to thank my parents for their continued care and support.

Contents

List of Figures

Understanding lineage-specific biology through comparative genomics.

Name: Yang Li
College: St Cross College
Degree: *Doctor of Philosophy*
Term: Trinity 2014

Abstract
A major challenge in biology is to identify how different species arose and acquired distinct phenotypic traits. High-throughput sequencing is transforming our understanding of biology by allowing us to study genomes and cellular processes at genome-wide levels. Only a decade subsequent to the publication of the first human genome draft, genome assemblies of hundreds of organisms have been produced. Yet, genome analysis remains challenging and advances have lagged far behind our sequencing abilities and other technological advances. The next generation of comparative genomicists must therefore understand, invent and apply a wide number of computational tools in order to study biology in the most efficient manner and in order to pose the most interesting questions. This thesis spans areas covering evolutionary genomics, gene regulation, and computational methods development. A major aim was to understand how genetic variation contributes to variation in phenotypic traits. This was approached using a large variety of evolutionary and comparative genomics tools. In particular, high-throughput sequencing datasets were analysed to study single-cell transcriptomics, gene duplications, gene architecture evolution, and alternative splicing. Additionally, in cases where off-the-shelf analysis tools were inexistent, novel pipelines and programs were designed and implemented to solve algorithmic problems such as scaffolding genome assemblies and short-read mapping onto small exons.

1 Chapter 1: Introduction

Physical appearance, behaviour, and susceptibility to diseases are traits offspring inherit from parents. This simple observation embodies the founding principles of genetics. In 1944, DNA was first suggested to be an essential molecule to heredity (Avery et al., 1944); today we know that DNA is fundamental to life and is shared among all living organisms. All living organisms share a common origin, and there is overwhelming evidence that accumulation of genomic changes throughout evolutionary time played a major role in today's biodiversity. Nevertheless, how genomic variation directly contributes to variation in phenotypic traits remains under active investigation.

My thesis explores protein coding gene variation in terms of sequence, intron-exon architecture and copy number. A major aim of my work has been to use high-throughput sequencing data to understand how evolutionary forces shape past and contemporary genomes. Another major aim of my work has been to develop analysis approaches for high-throughput sequencing data. Though very different in nature, these subjects are increasingly becoming essential aspects of research in genomics.

Owing to the varied types of analyses presented in this thesis, I will briefly describe some biological and computational concepts in this introduction, leaving more complex and technical discussions that is relevant to each study to their appropriate chapters. In particular, I provide an overview of important mechanisms that underlie phenotypic evolution, followed by a description of relevant computational approaches.

1.1 Evolutionary genomics

Genome evolution is governed by four major evolutionary forces: mutation, recombination, drift and selection. These forces constantly act on the genome over time. Although my work focuses on genomic differences that underlie phenotypic variation between (possibly distantly-related) species, an understanding of evolutionary changes at the population level is required. Indeed, within-species variation lies at the root of species divergence, and only variants that survive the early polymorphic stage and reach fixation will contribute to variation among species (Lynch, 2007).

Spontaneous mutations in the germline of an individual can be carried over to the genomes of offspring and introduce new alleles (i.e. variants) to the population. These alleles can increase, decrease, or remain stable in frequency over time. In rare cases, new alleles can become fixed in the population, i.e. reach a frequency of 100%. Over time, alleles that increase reproductive fitness are more likely to increase in frequency and to be fixed in the population (positive selection), while deleterious alleles tend to decrease in frequency and to be purged from the population (purifying selection). However, most new alleles tend not to affect reproductive fitness or to have small effects: they are effectively neutral and are thus expected to disappear from the population eventually. Nevertheless, due to random sampling of alleles in a finite population, most alleles (including slightly deleterious and neutral ones) can become fixed, by chance alone. This phenomenon is known as genetic drift.

The strength of genetic drift heavily depends on (effective) population size. This is because the variance associated with random sampling of alleles depends on the total number of individuals with offspring in a population. If the pop-

ulation size is small enough, genetic drift can dominate weak selective forces, resulting in a situation in which the genetic landscape of a population evolves nearly entirely randomly (Ohta, 1973; Kimura, 1983). This is an important concept that I have borne in mind when interpreting results from my analysis of East African cichlid species (Chapter 3) that have small population sizes.

Though most spontaneous mutations within an individual are never fixed in a species, it is clear that the genomes of even closely-related species differ greatly. These differences are rare exceptions that have accumulated over millions of years. Most differences were once neutral or near neutral alleles (effectively neutral; Ohta (1973)) that have no phenotypic effects and became fixed by chance. However, some fixed mutations do generate phenotypic variation. In the remainder of this section, I describe genetic mutations that impact phenotype. Furthermore, I describe several genomic changes that show signatures of positive selection because they are generally assumed to impact phenotype. This, of course, does not exclude the possibility that effectively neutral or deleterious mutations contribute to phenotypic diversity (but these have proved to be much harder to find or show).

1.1.1 Gene duplications and family expansions

Gene duplication is a major evolutionary mechanism that can generate molecular diversity (Ohno, 1970), and can occur in several distinct ways. Gene duplicates can result from chromosomal regions being duplicated in tandem (Eichler and Sankoff, 2003; Fan et al., 2008) which are caused by unequal cross-overs or replication slippage (De Grassi and Ciccarelli, 2009). They can be produced from complete duplications of the genome (Holland et al., 1994; Meyer and Schartl, 1999; Dehal and Boore, 2005), and through retrotransposition by which copies

of the retrotranscribed gene are inserted within the genome at other locations (Xing et al., 2006; Ewing et al., 2013).

Following duplication, genes may be more prone to random genetic drift and may then be subject to adaptive pressures that reflect species-specific biology. A priori, newly duplicated genes have redundant functions (except those without regulatory elements), and their trajectories (even those that reach fixation) generally follow the route of non-functionalization (Han et al., 2009). However, some acquire non-redundant functional roles and are retained in the genome. Gene duplicates that have been retained over a long evolutionary period are expected to have acquired diverse functional roles. Redundant copies may acquire mutations that alter protein structure and function. Gene duplicates may therefore acquire a completely novel function (Neofunctionalization; Ohno (1970)). Duplicated genes may also maintain ancestral function and acquire novel expression patterns. For instance, novel genes, including duplicates, are often found to be first expressed in testes of mammals, and are predicted to then slowly acquire expression in other tissues (Kaessmann, 2010; Assis and Bachtrog, 2014). Furthermore, ancestral functions can be divided between the duplicated and original copies of a gene, a process known as subfunctionalization (Ohno, 1970). A simple model for subfunctionalization is the duplication-degeneration-complementation model which proposes that degenerative mutations in one copy are buffered by the presence of the other copy. The accumulation of buffered mutations in both copies therefore leads to a scenario in which the ancestral gene function is divided between the two copies (Stoltzfus, 1999).

Among genes with retained copies that appear to have been repeatedly duplicated in animals, are those that encode visual pigments in insects and vertebrates

(Yuan et al., 2010; Rennison et al., 2012), venoms in snakes, platypus and mollusks (Chang and Duda, 2012), olfactory and gustatory receptors in insects and mammals (Robertson and Wanner, 2006; McBride, 2007). The Hox gene family of transcription factors that regulate animal development (Barten et al., 2001; Nikolaidis et al., 2005; Laun et al., 2006; Pearson et al., 2005a) has also undergone repeated duplication events. The Hox family is well studied and one of the most celebrated example of gene family expansion with clear phenotypic consequences. Hox genes are homeobox transcription factors that have key roles in patterning the antero-posterior axis of animals (Pearson et al., 2005b). Their duplication and subsequent evolution have substantially contributed to the large diversity of morphological phenotypes in arthropods and chordates (Carroll, 1995). Analysis of Hox-like clusters revealed that multiple paralogous clusters are present in animals: the Hox cluster (Pearson et al., 2005a), the ParaHox cluster (Brooke et al., 1998) and NK clusters (Kim and Nirenberg, 1989). These clusters likely originated from successive duplications that trace back to an ancestral proto-Hox cluster predating the cnidarian-bilaterian transition (Garcia-Fernandez, 2005). Duplications and subsequent evolution of Hox and Hox-like clusters represent one example of gene family expansion that played an instrumental role in animal phenotypic variation.

1.1.2 Non-synonymous substitutions

Non-synonymous substitutions are known to impact phenotypic traits, possibly by altering protein function. They arise when point mutations alter amino acids encoded in exonic regions. Analyses of coding sequence differences between *Drosophila* species and between primates estimated that over 40% and 10–20% of non-synonymous substitutions were adaptive during *Drosophila* and primate evolution, respectively (Clark et al., 2003; Eyre-Walker, 2006; Boyko

et al., 2008; Messer and Petrov, 2013). These percentages vary greatly probably owing to differences in effective population size and population history. Nevertheless, because of the prevalence of adaptive non-synonymous substitutions, they are widely believed to affect phenotypic variation (Evans et al., 2004; Dorus et al., 2004; Bustamante et al., 2005).

Most non-synonymous substitutions have unknown consequences on protein function. However, genome-wide scans revealed several interesting proteins with positively selected sites and known function (Clark et al., 2003; Nielsen et al., 2005; Kosiol et al., 2008). For examples, genes involved in immunity and sensory perceptions tend to harbour positively selected sites (Kosiol et al., 2008). These findings are somewhat expected because immune defence responds to changing viral and bacterial populations, and interestingly, some of these sites are located within protein domains predicted to bind viruses (Kosiol et al., 2008).

Another interesting example of a positively selected gene is Foxp2. Human Foxp2 differs by the substitution of two amino acids compared to its chimpanzee ortholog. Comparative analyses suggest that these two substitutions occurred in the human lineage and subsequently reached fixation (Zhang et al., 2002; Enard et al., 2002). Interestingly, one mutation in Foxp2 is associated with speech and language deficits in human. Furthermore, the two human-specific amino acids alter Foxp2 function by altering its transcriptional targets (Konopka et al., 2009). Therefore, these two substitutions may have contributed to the emergence of human language.

1.1.3 Gene expression and alternative splicing evolution

Gene regulatory evolution

Evidence increasingly suggests that non-coding mutations rather than mutations in coding regions underlie differences between closely related species (Wittkopp et al., 2004; Prud'homme et al., 2006, 2007). These non-coding mutations are believed to impact phenotypes by their effects on gene expression. In mammals, gene regulation evolves under tissue-specific selective pressures (Blekhman et al., 2008; Brawand et al., 2011), which is consistent with the idea that natural selection can act on diverse organ phenotypes through changes in tissue-specific gene regulation. According to models of gene expression level evolution, over 10% of all genes evolved under directional positive selection (i.e. preferential fixation of alleles increasing or decreasing expression levels to an optimum) in humans and other mammals (Gilad et al., 2006; Blekhman et al., 2008). Using population genetics data, positively selected mutations were found to affect gene regulation more often than they affect protein sequences (Kudaravalli et al., 2009; Fraser, 2013), which suggests an important role of gene regulatory changes on phenotypic evolution.

In an elegant study from the coelacanth genome project, Amemiya et al. (2013) identified a regulatory region that is proximal to the Hox-D cluster and conserved in tetrapods and coelacanth, but not in ray-finned fish. Hox-D is involved in the formation of digits in tetrapods (Montavon et al., 2011), which motivated Amemiya and colleagues to study the regulatory roles of the region they identified. Interestingly, the sequence of this region was shown to drive reporter expression in a limb-specific pattern in a mouse transgenic assay. The developmental regulation of limb part was therefore likely derived from a regu-

latory element present in the common ancestor of coelacanth and tetrapods.

There exist several other examples of gene regulatory evolution that contributed to phenotypic changes. Prud'homme and colleagues (2007) studied colour pigments divergence at the tip of male wings in several *Drosophila* species. Remarkably, differences in cis-regulatory sequence were found to cause expression level differences across Drosophila species, which in turn resulted in variation in wing colouration. Other examples include gene expression evolution that are associated in differences in stickleback pelvic morphology (Shapiro et al., 2004), and immune response (Tournamille et al., 1995) and dietary habits in human (Perry et al., 2007).

Alternative splicing evolution

Despite clear evidence that non-coding (cis-regulatory) changes play important roles in phenotypic diversity, some argue that variation in gene expression is unlikely to account for a large proportion of phenotypic variation across vertebrates because gene expression is largely conserved (Barbosa-Morais et al., 2012). Moreover, protein expression levels in primates evolve under even stronger evolutionary constraint than mRNA levels (Khan et al., 2013), further suggesting that other evolutionary mechanisms, such as changes in alternative splicing patterns (Barbosa-Morais et al., 2012; Merkin et al., 2012), likely played a greater role in phenotypic diversity than previously thought.

Alternative splicing (AS) is recognised to be a major player in generating protein diversity (Graveley, 2001; Brett et al., 2002; Hiller et al., 2004). AS allows novel exonic regions or splicing patterns to be tested within existing transcripts at low frequencies (Modrek and Lee, 2003). The inclusion of these novel

transcriptional events may then increase through genetic drift or through positive selection in the case they confer a fitness advantage. Moreover, because alternative splicing can also be regulated in a tissue-specific and organ-specific manner, these changes allow existing proteins to be expressed in a new spatial and temporal context. Additionally, the alternate usage of exons can contribute to gene regulation by inducing nonsense-mediated decay (Cuccurese et al., 2005), and by affecting the stability or translation of protein-coding mRNAs (Kelemen et al., 2013).

Unlike gene expression levels, alternative spliced events in mammals are generally species-specific (Pan et al., 2005; Takeda et al., 2008). Notably, profiles of alternative splicing events tend to be more strongly related among different tissues from the same species than they are among the same tissue from different species, even when comparing chimp and human organs that diverged only 6 million years ago (Barbosa-Morais et al., 2012). These results can be interpreted as evidence for a high turnover of alternative splicing events, which provides a large number of novel transcripts for natural selection to operate on.

Many lineage-specific splicing events may be solely driven by noisy splicing and do not underlie functional differences (Pickrell et al., 2010b). However, some AS events are expected to generate multiple functional isoforms from single genes (Xing and Lee, 2006). In primates, a recent study (Reyes et al., 2013) identified 3,800 AS exons from 1,643 genes that show conservation of tissue-dependent usage patterns (which corresponds to nearly 10% of all multi-exonic genes). Therefore, it is clear that alternative splicing possesses functional roles and that evolutionary changes affecting exonic splicing patterns can generate phenotypic innovations.

1.2 Genome projects and Next-generation sequencing

Modern sequencing technologies are revolutionising the field of genetics. Like PCR (polymerase chain reaction), which is now considered to be indispensable to molecular biology research, high-throughput sequencing is changing the way biologists study genetics. Owing to the constant decline of sequencing costs, decoding the genome of most living organisms is now possible at a modest price. However, one major limitation of high-throughput sequencing (compared to older sequencing methods with much lower throughput) is the short length of the reads produced (typically $\leq$ 100bp). Although read lengths are steadily increasing (e.g. $\sim$250bp with MiSeq, multiple kbs with PacBio, and up to 10kb with Moleculo; Kuleshov et al. (2014)), a large number of sequencing runs still produce read sizes of 39 to 100bp. High-throughput sequencing therefore produces new challenges for sequence analysis particularly in alignment and assembly of short reads.

Genome projects have occupied the time of a large number of genomicists in the early 2000s. Since the completion of the human genome project, the genomes of hundreds of animals have been assembled and analysed yielding a wealth of biological insights. Nevertheless, perhaps owing to the rapid maturation of sequencing technology, most researchers have moved from genome analysis to the study of molecular machinery, population history, and human disease. Yet much remains to be learned about the genomic basis of specific morphological traits and environmental adaptations.

Indeed, the genome of newly sequenced species are now being used to study specific evolutionary phenomena. To name but a few, just last year genome projects yielded insights into avian influenza virus hosts from duck (Huang et al., 2013), turtle-specific body plan development from the softshell and green tur-

tles (Wang et al., 2013), echolocation from bats (Parker et al., 2013), parasitism from tapeworms (Tsai et al., 2013), bilaterian evolution from spiraleans (Simakov et al., 2013), and vertebrate evolution from lamprey (Smith et al., 2013) and coelacanth (Amemiya et al., 2013).

Despite the large number of successful genome projects, completing a genome project is far from routine and often requires a large amount of resources. From start to finish, a genome project generally takes several years (the reference genome of the zebrafish took over 10 years to be finalised), and depends on the collaboration of several dozens of researchers. Most genome projects follow different strategies to obtain a genome assembly, to annotate the genome, and to infer the evolutionary history of diverse elements using comparative genomics. In this section, I share some of my personal experience on these topics from several genome projects.

1.2.1 Genome assembly

The genome assembly of a species is a representation of the DNA sequences that collectively make up its genetic material. The assembly is the most important part of a genome project as most downstream analyses benefit from a high-quality assembly (I describe three metrics for quality later on). To sequence a genome, DNA must be extracted from the tissue of one or more individuals, and the tissue is selected according to ethical or practical concerns. For example in humans, DNA can be extracted from white blood cells, sperm, or saliva (Lander et al., 2001).

To improve assembly quality, several measures can be taken. For example, polymorphism within sequenced DNA has been a long-standing issue for genome

assemblers (Aparicio et al., 2002; Huang et al., 2012). This is owing to most assemblers' internal representation of the genome as a graph. Because this graph must first recapitulate each possible haplotype (see Figure 1) and is only then trimmed down to a linear sequence, the more polymorphic a genome is, the more complex this graph and pruning steps will be. For this reason, the levels of polymorphism within the sequenced individual(s) should be as low as possible. In many cases, this is impossible. However, if DNA is pooled from more than one individual, selecting closely-related individuals avoids increased numbers of polymorphic sites. Furthermore, in some organisms, individuals can be inbred, which result in genomes with low levels of polymorphism. Uneven genome coverage can also cause assembly problems and errors (Hunt et al., 2013; Chen et al., 2013). For species with XY or ZW sex-determination systems, sequencing the homogametic sex (i.e. the sex with the same two copies of the sex chromosomes) is advantageous because the coverage of the sex chromosomes will be equal to that of the autosomes.

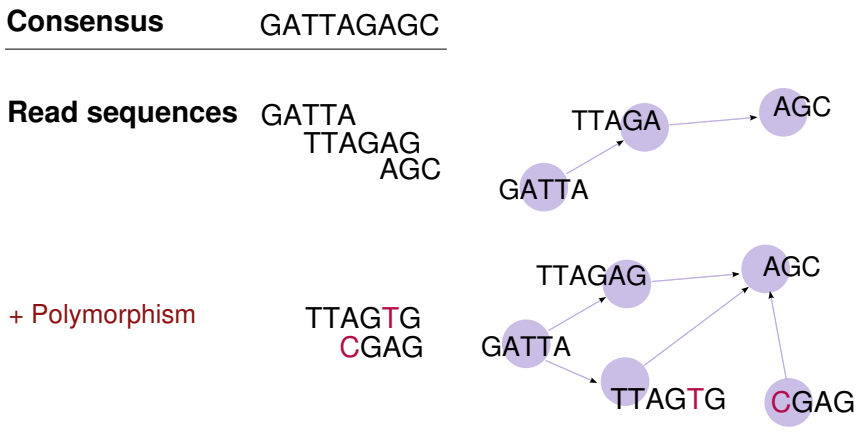

Figure 1: Graph complexity increases as polymorphism increases.

These considerations help the assembly process of both long reads generated by Sanger/Capillary sequencing and shorter reads generated from next-generation (Illumina) sequencing. However, short read lengths (i.e. 100-250bp)

present additional challenges for assembling sequences into a complete genome. Indeed, short-read assemblers must handle a much larger amount of sequence data compared to traditional assemblers. Special data structures, such as the de Bruijn graph (Zerbino and Birney, 2008; Li et al., 2010), must therefore be employed. Assembling repetitive regions within a genome is also difficult owing to the inability of short reads to span repeats longer than few hundred nucleotides. This problem can be circumvented by using paired end or mate pair reads with different insert sizes long enough to span repetitive regions (Gnerre et al., 2011). Nevertheless, mate paired reads with long insert sizes (>50kb) are difficult to generate (Personal communications de Magalhães, di Palma). Some genome projects use a hybrid approach in which high-throughput sequencing is complemented by lower-throughput sequencing (Roche 454, or Sanger sequencing), or use extra information (e.g. orthologous protein sequences see Chapter 2) to improve the assembly.

Popular assembly software include ALLPATH-LG (Gnerre et al., 2011), ABySS (Birol et al., 2009), Velvet (Zerbino and Birney, 2008), and SOAP-denovo (Li et al., 2010). Although all of these programs can assemble short sequence reads into a genome assembly, published assemblies largely vary in quality. For instance, ALLPATH-LG (Broad Institute) and SOAP-denovo (BGI-Shenzhen) have consistently produced high-quality genomes in the last several years. Although there exists no single metric that can single-handedly measure the quality of an assembly, I describe the three most relevant metrics below: the assembly coverage, the contig/scaffold N50, and the gene coverage.

Sequencing coverage

An informative statistic is the coverage at which a genome has been se-

quenced. The coverage of a genome assembly is the expected number of times each base in a genome has been sequenced, or in other terms, it is the number of bases sequenced divided by the size of the genome. Note, however, that biases in the sequencing and assembly process can result in uncovered genomic regions (Sims et al., 2014). In particular, a bias in a standard Illumina DNA amplification protocol sometimes results in an under-representation or absence of sequences with extreme GC-content (Dohm et al., 2008). Under-sequencing GC-rich/poor regions can cause large genomic regions to be missed or mis-assembled as some genomes (notably birds and mammals) are known to possess large segments with elevated or decreased GC-contents (Bernardi, 2000; Costantini et al., 2009). Though, there now exist modifications to the amplification protocols that minimizes GC bias (Aird et al., 2011), and should be used if uneven coverage is a limiting step.

While a higher sequencing coverage is generally better for the quality of a genome assembly, differences in sequencing technology also affect assembly quality. For example, sequences produced by Sanger sequencing have fewer fragments, lower coverage, but cover larger genomic regions. This allows a better reconstruction of the genome owing to reads spanning across regions difficult to sequence, e.g. repeats. Assembling low coverage Sanger reads can therefore produce better assemblies than assembling short reads at a much higher coverage. Nevertheless, a higher coverage is generally preferable when comparing genomes produced using the same technologies as it reduces the number/lengths of unsampled regions. A higher coverage also allows a better estimate of heterozygosity within an individual, a feature that can be used to infer the population history of the sequenced species (Li and Durbin, 2011).

N50 statistics

A genome assembly rarely consists of contiguous sequences that capture full chromosomal sequences. Instead, chromosomes are divided into small chunks of contiguous sequences commonly known as contigs. More precisely, contigs are contiguous strings of DNA sequence representing genomic regions which have been completely assembled, mostly without gaps. Contigs can be linked together using various information, e.g. paired-end reads or linkage maps, to form scaffolds. Thus, scaffolds are separated by regions of unknown sequences and lengths (though sometimes the length can be estimated from paired-end read insert sizes).

The contig and scaffold N50 are widely used statistics to describe the lengths of contigs and scaffolds relative to the size of an assembly. Given a set of contigs/scaffolds of various lengths, the N50 is defined as the length for which the set of all contigs/scaffolds of that length or longer contains half of the total of the lengths of all contigs/scaffolds. The N50 statistic is often used as a simple, yet informative, statistic in genome projects to obtain an idea of the contiguity of an assembly.

In general, repeat-rich genomes will have low N50 values owing to difficulties in assembling repetitive regions. This is problematic for applications that require assemblies with high N50s. For instance, large contiguous sequences allow us to infer tandem duplication events and sometimes large-scale genomic rearrangements. They may also help us to study the evolution of gene clusters – a difficult task when genes from the same clusters are located on different contigs/scaffolds. For many other applications, high N50s are optional. For instance, because genic regions are often depleted in repeats in the forms of transposable elements (Nel-

laker et al., 2012), gene-rich regions of the genome may be well-assembled into single contigs despite low N50s. This allows individual genes to be studied in terms of their evolutionary history. Of course, if the N50 is very small, e.g. the same length as (or shorter than) the average length of a gene, then a large number of genes will likely be split into different contigs.

For reference, a contig N50 of 15–25kb is now standard for vertebrate genomes and considered to be of good quality (Broad Institute blog, `http://www.broadinstitute.org/software/allpaths-lg/blog/`).

Gene coverage

Gene coverage, i.e. the number of complete genes covered by the assembly, is for many the most important quality measure for an assembly. This is because most of the important information within a genome is captured by genes, either at the expression or protein level. However, gene coverage is hard to measure, and more often than not, the total number of genes is used as a measure instead. As a rule of thumb, the total number of complete genes predicted should be around twenty thousand (for vertebrate genomes) to be considered of high-quality. However, this requires gene annotation efforts that require months and sometimes years to undertake – a delay that is unacceptable at the stage of quality checking.

There are several different ways gene coverage are assessed in the initial stages of a genome project. One way to assess gene coverage is to map proteins from a related species to the new assembly using TBLASTN. However, not all proteins are represented in every genome, therefore the gene coverage may be largely underestimated. Para and colleagues (2007) circumvented this problem by compiling a list of 458 "Core Eukaryotic Genes" (CEG) from eukaryotic

gene clusters that are conserved in *Arabidopsis thaliana*, *Caenorhabditis elegans*, *Drosophila melanogaster*, *Homo sapiens*, *Saccharomyces cerevisiae*, and *Schizosaccharomyces pombe*. This list therefore consists of highly conserved genes which are predicted to be essential in all species and therefore present in their genomes. Using this list, gene completeness can be assessed by mapping the CEGs of a well-annotated, closely-related species onto the genome. Assuming these highly conserved genes are representative (in gene architecture, lengths and sequence composition) of other genes in the genome, the estimated number of complete and partial genes is expected to approximate the true gene coverage of the assembly. Indeed, Parra et al. (2009) later showed by simulating genome assemblies that the average discrepancy between the estimated completeness and true completeness was less than 10%.

1.2.2 Genome annotation

Once the initial genome assembly has been assessed for N50 and gene coverage, genes and other elements of interest can be annotated. Over the last decade, several methods have been successfully used to annotate genes within newly assembled genomes. These can be grouped within three classes of methods: empirical methods, prediction by homology, and *ab initio* gene prediction.

Empirical methods

A class of successful annotation methods has been empirical methods. Briefly, polyadenylated RNA are sequenced (also see section on short read mapping) and processed into expressed sequenced tags (ESTs) or assembled cDNA fragments (from RNA-sequencing). The ESTs or cDNA fragments are then mapped onto the genome which allows transcribed clusters, i.e.

complete or partial transcript sequences, to be constructed (Haas et al., 2008). Subsequently, coding sequences can be predicted by identifying long open reading frames and particular amino acid composition (Guigo et al., 1992). Owing to the high quality annotation produced by this method, empirical methods have become very popular. In fact, this strategy has been used in several recent genome projects including our annotation of five East African cichlid genomes (Brawand et al., under review; also see Chapter 4), and annotation in other genomes (Wang et al., 2013; Amemiya et al., 2013; Kim et al., 2011).

To briefly illustrate empirical methods with a concrete example, I used the following procedures as part of the annotation effort for the cichlids genome project. First, RNA-seq libraries were generated from ∼11 different tissues in each of the five East-African cichlids whose genome we assembled. Illumina RNA-sequencing of these libraries yielded 20-60 million reads for each tissue which were assembled into transcripts using Trinity (Grabherr et al., 2011). Mapping the transcripts to the genomes with GMAP/PASA, and identifying long open read frames resulted in 18-21K protein-coding gene predictions for each cichlid genome.

Not only is this approach very effective, but it is the only method that allows the different isoforms of a gene to be predicted (with low or no false positives). However, a limitation to empirical approaches is that only a fraction of all genes (and isoforms) is expressed in any single tissue or developmental stage. The RNA of many tissues must therefore be sequenced to obtain a (near) comprehensive annotation. High costs incurred by the production of multiple libraries, or a limited number of RNA sample (especially from rare specimens) can therefore be problematic.

Orthology/homology-based methods

Protein coding sequences from related species can also be used to predict gene structure. This prediction strategy is known as prediction by homology or orthology. Because protein coding sequences generally evolve at very slow rates, they can be used to predict homologous protein coding genes from another species. Of course, a complete set of protein coding sequences from a closely-related species is ideal (divergence <50My (million years)). However, a pipeline such as GPIPE (Heger and Ponting, 2007) and the Ensembl annotation pipeline (`http://www.ensembl.org/info/genome/genebuild/genome_annotation.html`) are able to use distantly related proteins to predict the exon-intron structure of many genes. For example, I predicted 20,622 protein-coding genes in the painted turtle genome using chicken (>250My divergence), lizard (>280My divergence) and mammalian protein sequences as template models (Shaffer et al., 2013). Evidently, fast evolving genes and lineage-specific genes will likely be missed. However, novel genes that arose by duplication can often be identified by this class of methods as they tend to resemble ancestral copies, e.g. beta-keratins (Shaffer et al. (2013); Li et al. (2013); see Chapter 3).

Ab initio methods

Ab initio methods predict gene exon-intron structures by exploiting sequence composition and motifs that are characteristic of splice sites, translation initiation and termination sites (Burge and Karlin, 1997; Salzberg et al., 1998). *Ab initio* gene predictors often assign a score to putative coding exons based on their lengths, the presence of splice site motifs, sequence composition and other characteristics such as the presence of an upstream poly-pyrimidine tract. Sub-

sequently, they may join consecutive high-scoring exons with consistent reading frames into full genes (minus the untranslated regions). Despite their usefulness in the past, *de novo* gene predictions tools have fallen out of fashion owing to increasing availability of homologous protein coding sequences and/or cheap transcriptome sequencing that led to the success of other gene prediction methods. Nevertheless, they are often used to complement other prediction methods (Bat genome; Zhang et al. (2013); Softshell turtle genome; Wang et al. (2013)).

Combining multiple methods

Each different gene prediction approach has its own merits and faults. Therefore combining the predictions of several methods has been a common strategy among genome annotation projects. EVidenceModeller (EVM; Haas et al. (2008)) allows gene prediction models from different sources to be combined into a single annotation. EVM also allows weights or priors to be assigned to each different model source. These priors represent the confidence in which we believe a prediction from a particular source to be correct. For example, gene models from *de novo* predictions are generally scored lowly in terms of confidence, while alignments from transcripts assembled from native RNA sequences are scored highly. Combining all sources of predictions allows us to generate annotation with varying confidence levels: a high confidence set of protein coding genes with evidence from all prediction methods, medium confidence sets consisting of genes with evidence from some methods, and a low confidence set consisting of genes with relaxed coding evidence. Each of these annotation sets may be useful, depending on the analysis of interest.

1.2.3 Orthology/paralogy prediction

A set of predicted genes in a new species is generally of little use on its own. Gene orthology and paralogy relationships must be inferred to fully exploit newly annotated genes. Orthologs are genes derived through speciation from a single genes, while paralogs are genes that result from duplication events either before (out-paralogs) or after (in-paralogs) speciation. Because orthologs and paralogs are derived from a common ancestor, they are often assumed to have the same or similar functions. Thus, genes can be assigned functions based on ortholog/paralog functions that have previously been established in other organisms.

The relationships between genes can be represented by a phylogenetic tree, in which genes from the same family can be traced back to a common ancestor (see Figure 2 for an example; also see `http://www.ensembl.org/info/genome/compara/homology_method.html`). Once this gene tree is constructed, evolutionary changes such as mutations, gene duplication and loss can be assigned to species lineages by reconciling the gene tree with the species tree.

Gene clustering

A first step to orthology/paralogy prediction is to cluster similar genes into families. Most approaches differ by how similarity is defined. For closely-related genes (i.e. those that can be traced back to a recent gene ancestor), clustering genes with high pairwise percent nucleotide identity is an effective strategy. For distantly related genes, pairwise amino acid identity is often used instead. For instance, OPTIC (Heger and Ponting, 2008) performs pairwise BLASTP alignments between protein sequences from each species pair and then uses a tree-based orthology assignment method to infer orthology relationships between

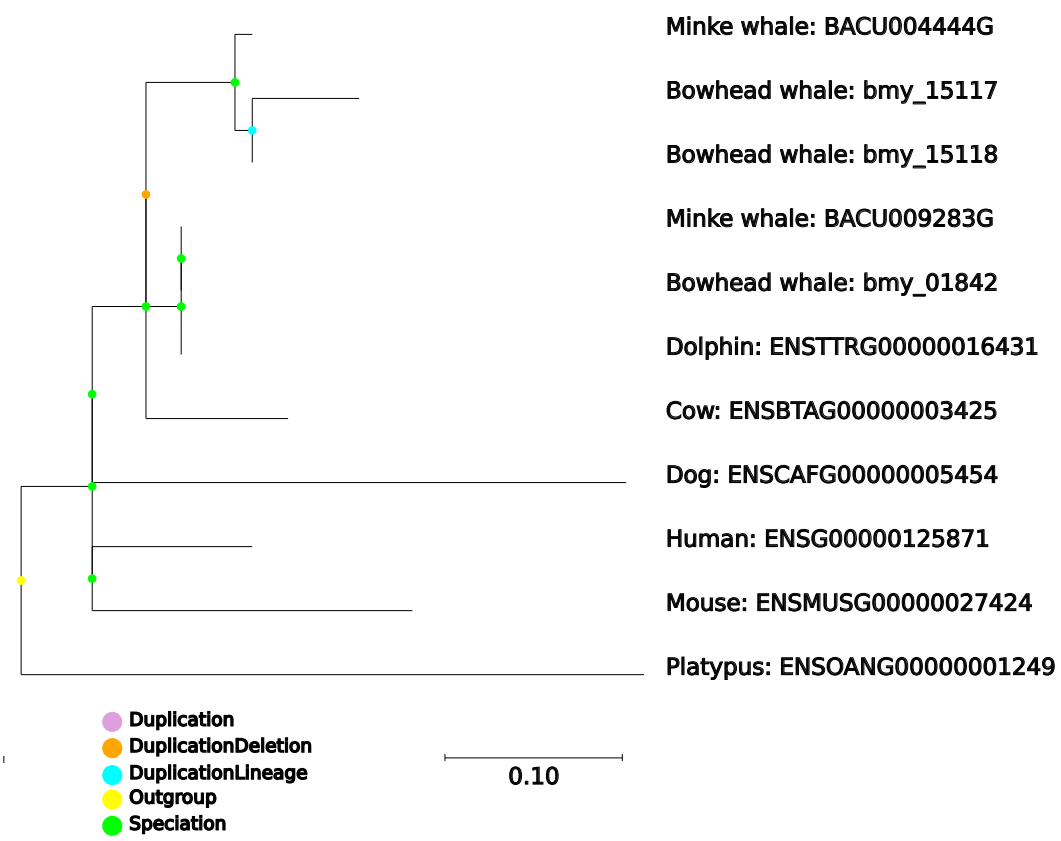

Figure 2: Example of a gene tree (MGME1).
Example taken from a gene predicted to have underwent duplication/loss in marine mammals (prediction by OPTIC; see Chapter 3). Branch length is measured as dS (the number of synonymous substitutions per site).

gene pairs. Using these pairwise assignments, gene clusters are constructed by exploring the graph with pairwise orthology relationships as edges. Gene clusters therefore consist of genes that are homologous to one another.

Gene tree reconstruction and reconciliation with species tree

Several algorithms can then be used to reconstruct the evolutionary history (gene trees) of the gene clusters. In general, these algorithms take multiple alignments of gene sequences in which each column is derived from a single ancestral base/amino acid, and predict the most likely phylogenetic tree (according to several parameters) describing the evolutionary history of this gene. For small gene families with a small number of duplication/loss events, simple

tree reconstruction algorithms that use parsimony or gene distant are able to infer correct evolutionary histories (Felsenstein, 1989). However, maximum likelihood or Bayesian approaches, e.g. PhyML (Guindon and Gascuel, 2003), RAXML (Stamatakis, 2014) or BEAST (Drummond and Rambaut, 2007), are widely thought to produce better reconstructions for larger gene clusters as they allow all model parameters to be estimated (Posada, 2008).

Some methods can also incorporate multiple trees together. For instance, TreeBeST (`http://treesoft.sourceforge.net/treebest.shtml`) combines five different trees (see Tree building section at `http://www.ensembl.org/info/genome/compara/homology_method.html` for details), which allows it to take advantage of strengths from multiple methods: trees based on DNA are often are more accurate for closely related parts of trees, whilst trees based on protein sequences are better at inferring the relationships of distantly-related genes.

Despite advanced methods for gene tree reconstruction, the gene trees of several clusters are notably difficult to reconstruct due to a lack of informative sites in the individual genes owing to short sequences or slow substitution rates (Rasmussen and Kellis, 2007). Furthermore, when gene clusters harbours both fast-evolving and slow-evolving lineages, genes that are fast-evolving tend to be mistakenly clustered together due to sequence convergence, a bias known as long branch attraction (Bergsten, 2005). A genome-wide analysis of gene duplications and losses is therefore likely to yield many false positives (Rasmussen and Kellis, 2007). For these reasons and because gene tree reconstruction heavily affects gene duplications/losses analyses (see below), manual inspection of the sequence alignments and tree constructed is performed in our analyses (see Chapter 3)

and others.

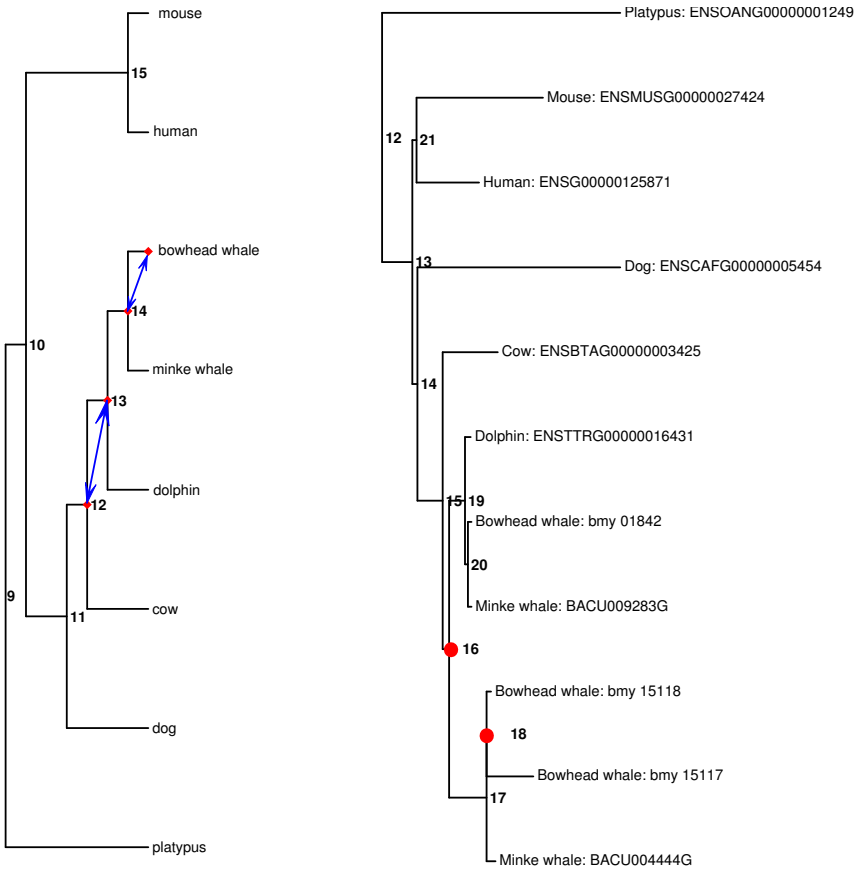

Figure 3: Example of gene-species tree reconciliation (MGME1). Blue arrows represent lineages where duplication occurred and red dots (right) represent duplication events on the gene tree. In this case, a duplication of MGME1 is predicted to have occurred in the ancestors of marine mammals, followed by a copy loss in a lineage leading to dolphins, and an additional duplication event in bowhead whales. Numbers on the tree nodes represent internal representations of the trees.

For applications such as a gene duplication analysis, the topology of the gene tree alone is often sufficient. Because gene trees and species trees are sometimes incongruent due to gene duplications and losses, the gene tree must be reconciled with the species tree in order to assign duplications/losses to tree branches (see Doyon et al. (2009) for a review). Missing genes from assemblies can affect tree reconciliation. However, if the species tree is relatively large (e.g. >5 species) and

the gene coverage high (>90%), the general reconciliation will be correct and gene duplication events tend to also be assigned correctly. To give an example, because the dolphin genome is relatively poorly annotated (15.8K predicted genes), the predicted loss of a MGME1 copy in dolphins (Figure 3) may simply reflect an incomplete annotation. Nevertheless, despite the possibility that the MGME1 copy loss is an artefact of the dolphin annotation, the general tree reconciliation is likely to be correct.

1.2.4 Short read mapping

The last topic I introduce here is short-read mapping. Although short-read mapping is not essential for genome projects, it is increasingly used to study molecular pathways in non-model organisms. For instance, Wang et al. (2013) used RNA-seq to identify genes that showed turtle-specific increases in expression during embryogenesis. They found 233 genes that showed an increased in expression compared to chicken orthologs (at the same embryonic stage), and a significant number of these genes are involved in ossification and extracellular matrix regulation according to a Gene Ontology analysis. Further studies of these genes may therefore shed light on the extensive ossification or the folding of the body wall involved in turtle shell formation (Wang et al., 2013).

Over 60 mappers currently exist for short read mapping (Fonseca et al., 2012). The aim of these aligners is to find the true location of each read within a reference while taking into account sequencing errors and genetic variation. All this must be done in a reasonable amount of time. Next-generation aligners must align millions (sometimes billions) of short reads to a reference sequence, a task that is intractable using conventional algorithms such as Smith-Waterman (Smith and Waterman, 1981) or BLAST (Altschul et al., 1990). The development

of a large number of different mappers has been motivated by both technological advances and novel biological applications (e.g. RNA sequencing). For example, paired-end reads, produced by the sequencing of both ends of a DNA fragment instead of one, led to the development of algorithms which used pairing information to achieve higher mapping accuracy (Langmead et al., 2009; Li and Durbin, 2009). The increase in read length from 39bp to expected reads of up to 1Mbp has also motivated the development of algorithm variants such as BWA-MEM and BWA-SW (Li and Durbin, 2010; Li, 2013). Some aligners, such as stampy, were motivated by an interest in characterising and aligning reads containing structural variation that previous aligners failed to map (Lunter and Goodson, 2011).

RNA sequencing

Measuring the gene expression in cells has yielded an incredible amount of biological insights, for example in terms of gene expression changes over cellular differentiation, and in terms of how gene expression differs between a diseased and a healthy cell. Previous high-throughput analyses of gene expression were generally conducted with microarrays. However, because RNA-sequencing offers several advantages compared to microarrays (see Wang et al. (2009) for review), RNA-seq has steadily increased in popularity. For instance, to quantify gene expression levels with microarrays, probes must be designed according to existing gene annotations, which limits our ability to detect novel transcripts. In contrast, RNA-seq outputs sequence fragments which allows us to study the transcriptomes of non-model organisms that lack annotations. For this reason, a large number of aligners were developed for the specific task of mapping reads from RNA-seq data to the genome.

Compared to reads originating from DNA fragments, mapping RNA reads to the genome has the additional difficulty that they can overlap one or multiple splice-junctions. RNA short read aligners must therefore be able to map reads to multiple genomic regions separated by intronic sequences. When a complete or near-complete transcript annotation exists (i.e. when nearly all transcripts have been sampled within a species), RNA reads can be mapped directly to transcript sequences (as DNA reads are mapped to genomic sequences), and converted to genomic coordinates. However, when no annotation or only a partial one exists, RNA aligners must be able to map these spliced reads onto the genome. Some RNA aligners, such as TOPHAT (Trapnell et al., 2009), map reads in two steps. First, RNA reads are mapped to the genome in the same way DNA reads are mapped; some of these reads, e.g. reads that are within a single exon, will map contiguously to the genome. The remaining, unmapped, reads are then divided into small fragments, which are mapped to the genome independently. The assumption is that several fragments will map to the correct genomic location and the spliced alignments can be recovered by extending the alignments (Trapnell et al., 2009).

Recently, a large-scale evaluation of RNA-seq aligners was performed (Engstrom et al., 2013). In total, the authors compared 11 programs and pipelines and identified major differences in performance in terms of alignment yield, mapping accuracy, exon junction discovery among others. Because each aligner uses different heuristics or assumptions to speed up the mapping process, a good understanding of the strength of different mappers is useful if we wish to take advantage of multiple aligners. As an example, I used two different aligners in one project (Chapter 5) to exploit the speed (bwa) and sensitivity towards indels (stampy) of existing short-read aligners.

1.3 Overview of the Thesis

My work is divided into four chapters (Introduction and Discussion excluded), each based on analyses that has been published or is under preparation. In **Chapter 2**, I start with the theoretical portion of the thesis. I present a workflow and pipeline that uses protein sequences to link genome assembly contigs together (published in *Bioinformatics*, Li and Copley (2013)). I also describe my contribution to the assessment of a novel protocol that aims to sequence DNA and RNA from a single cell (in preparation, Macaulay et al.). In **Chapter 3**, I present a large-scale study of beta-keratin gene family expansion in the turtle lineage (published in *Genome Biology and Evolution*, Li et al. (2013)), which I discovered whilst annotating the Painted Turtle genome (published in *Genome Biology*, Shaffer et al. (2013)). I also present a gene duplication analysis I have conducted for the bowhead whale genome project (in preparation, Keane et al.). In **Chapter 4**, I highlight my contributions to the analysis of five East African cichlids genomes (2nd round of reviews, Brawand et al. (co-first author)). These include an overview of our annotation efforts, the characterisation of incomplete lineage sorting within their genomes, an analysis of positive selection in candidate colour and morphology genes, and lastly a scan for novel exons and alternatively splicing events. **Chapter 5** describes my work on small exons that are 51nt or shorter (in preparation, Li et al.). I show how comparative genomics can be used to study micro-exons at the DNA, RNA and protein structure levels. In **Chapter 6**, I discuss strengths, weaknesses, and future directions of my work.

2 Chapter 2: Exploiting high-throughput genomics

Evolutionary genetics has shifted from a theory-dominated field into a data-driven field. The amount of data produced at sequencing centres far outpaces our ability to analyse them. High-throughput sequencing is changing the way we study biology. A deep understanding of its strengths and limitations is therefore crucial if we wish to utilise novel approaches to address questions that were previously intractable.

Data in genomics often require substantial processing and careful analysis before biological information can be extracted. In this chapter, I describe strategies I have used or developed to study several different large datasets. Although the focus of this chapter is not on biological findings, I later discuss several avenues in which these strategies can be employed to study biological questions.

This chapter is divided into two sections. First, I present a method that uses protein sequences to improve the contiguity of *de novo* assemblies, which has been published in *Bioinformatics* (Li and Copley, 2013). Then, I describe my analysis of DNA and RNA sequencing data from single cells, part of which will be included in (Macaulay et al., unpublished).

All results presented in this chapter are my own. Richard Copley contributed to the guided assembly project, in which he designed the study and significantly edited the manuscript. Iain Macaulay generated all single cell data and contributed the GT-seq protocol.

2.1 Scaffolding low quality genomes using orthologous protein sequences

2.1.1 Background

This project was motivated by the large number of low-quality assemblies that have been generated in the high-throughput sequencing era. Many (unpublished) reference assemblies have contig N50s of less than 1kb, leaving a majority of genes split across multiple contigs. However, even basic comparative studies require an accurate map between orthologous genes, which is difficult to obtain if genes are fragmented. Richard Copley and I wished to use protein sequences to improve the usability of these genomes for large-scale analyses.

Several groups have worked on incorporating available related assembled genomes into the assembly process (Pop and Salzberg, 2008). For instance, whole genome alignments can be used to build super scaffolds by aligning the new assembly to the high quality genome of closely related species (e.g. Locke et al. (2011)). Many species, however, are sequenced because they belong to an unrepresented taxon for which no complete genome is available. Thus, other approaches have been adopted to increase assembly contiguity. In Mortazavi et al. (2010), for example, researchers exploited eukaryote exon-intron gene structure and used RNA-seq data to scaffold contigs from a *Caenorhabditis* nematode genome, increasing the initial N50 from 5.0kb to 9.4kb. The disadvantage of this strategy is the requirement of existing RNA-Seq data in addition to genomic sequence. Furthermore, genes that are not expressed in the RNA-Seq library will not be available for scaffolding.

As protein sequences are generally well conserved across distant taxa (compared to non-coding or DNA sequences), a possible way to deal with fragmented

initial assemblies is to use orthologous proteins to guide the scaffolding of contigs which encode fragments of the same proteins. Surget-Groba and Montoya-Burgos used orthologous proteins to guide transcriptome assembly by mapping contigs onto a related protein set and improved the N50 of their zebrafish transcriptome assembly by up to 42% (Surget-Groba and Montoya-Burgos, 2010). However, using orthologous protein sequences to scaffold a genome is a fundamentally more difficult problem due to the presence of introns, large intergenic regions and the much larger volume of data involved in genome assemblies. Salzberg and co-workers introduced a 'gene-boosting' algorithm that used amino acid sequences from predicted proteins to improve the scaffolding of a 6Mb bacterial genome, although again, as bacterial proteins lack introns, this is a somewhat easier problem to approach (Salzberg et al., 2008).

In this section, I present an algorithm which uses proteome sets from different species to guide scaffolding of genome assemblies (including highly fragmented ones, e.g. N50 $<$ 3kb). The term scaffolding refers to the process of ordering and orienting contigs without estimating distances between contigs. I call this pipeline SWiPS (Scaffolding With Protein Sequences). Existing algorithms and strategies, including the Ensembl genebuild process for low-coverage assemblies, and GPIPE (Heger and Ponting, 2007), are not suited for highly fragmented assemblies where a gene may be spread over multiple contigs. To deal with fragmented genes within the assembly, SWiPS uses orthologous proteins (possibly from distantly related species) as guides to link the exons of a protein that may be situated on different contigs. At the time of writing, only one other published method, ESPRIT Dessimoz et al. (2011), does this. ESPRIT was used to bridge 666 *Callorhinchus milii* contigs from a Sanger-based assembly together and reported precision of about

80%. ESPRIT, however, assumes a uniform evolutionary rate over the entire length of split proteins, something that is clearly questionable in the general case.

SWiPS is a novel approach for scaffolding contigs by a greedy optimisation strategy that improves the contiguity of novel assemblies, reduces their fragmentation, as well as predicts the content of their exomes. I tested SWiPS on several genomic assemblies comprising of simulated *Ciona intestinalis* next generation sequencing data, real sequencing data from *Drosophila melanogaster*, as well as the Sanger-sequenced assembly of *Callorhinchus milii* and a pre-assembled *Homo sapiens* genome.

2.1.2 Methods

Outline of method

The overall method is outlined in Figure 4 and described in detail later on. Briefly, SWiPS first identifies contigs that include protein coding exons. Next, the contigs are ordered and oriented according to their mappings to proteins. At this stage, multiple contigs are allowed to map to the same protein. Similarity scores are then computed between proteins and contigs according to a distance matrix. Subsequently, SWiPS uses these scores to scaffold contigs together greedily by iteratively picking the best scaffold model (protein sequence) for any subset of contigs. Lastly, SWiPS links scaffolds into super-scaffolds by examining contigs encoding multiple proteins which, in turn, best model multiple contigs.

Determining coding contigs and exonic regions

In order to determine which contigs might be protein-coding, proteins from user specified guide proteomes are mapped onto the set of all contigs using

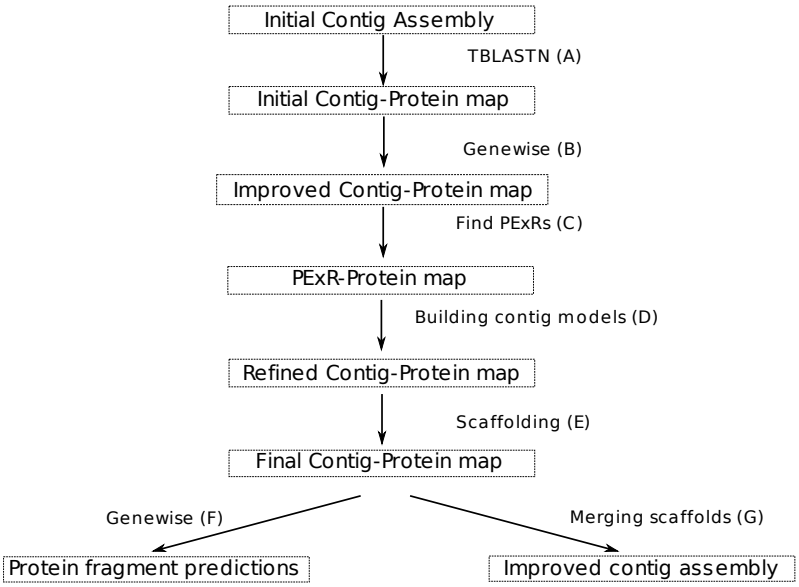

Figure 4: SWiPS workflow.
A) Proteins in the chosen guide proteome are mapped to contigs in the initial assembly using TBLASN. Low scoring overlapping mappings are removed. B) Proteins with high scoring mappings are realigned with genewise to obtain nucleotide level resolution. C) Putative Exonic regions (PExRs) are predicted by combining regions in the contig that encode multiple homologous protein segments. D) PExRs within each contig are used to establish a refined contig-protein orthology model, only allowing contigs to map to different proteins if the associated PExRs show no overlap. Then, contig gene models are built using this mapping. E) A score optimisation heuristic is used to scaffold contigs together according to the total score of protein models. F) genewise is used once again on the protein scaffolds to predict protein fragments. G) A graph theoretic approach is used to find the local order of the contigs and their strand.

tblastn (Gertz et al., 2006). An initial protein-contig mapping is obtained by assigning a contig to a protein if it (1) has a mapping e-value below 10^{-5} and (2) that e-value is under 5 orders of magnitude less significant than the most significant contig mapping to the same protein region. Proteins are allowed to map to multiple contigs. Two contigs are considered to map to the same protein region if the regions they map to overlap by 15 amino acids. These parameters are chosen to be relaxed compared to similar usage in previous studies e.g. (Hahn et al., 2007). However, more conservative parameters can be chosen at the cost of sensitivity to gain running speed.

SWiPS refines the tblastn-based protein to contig mapping using genewise, to more precisely define exon boundaries (contigs of sufficient length will contain multiple exons from the same gene). SWiPS aligns the template protein to the contig using the "623" model of genewise (Birney et al., 2004), successively searching the regions of contig flanking each alignment until no more high scoring regions are found (although this is in principle possible with the looping models of genewise, in practice these yield results that are hard to interpret). The results from all iterations are then merged and all mapping coordinates are corrected to yield one or multiple protein-contig pairwise alignments.

After re-aligning with genewise, a mapping from contig intervals to protein intervals is established. Then, for every contig, the longest contiguous regions with at least one protein mapping to every base in each region are found and defined to be Putative Exonic Regions (PExRs), i.e. PExRs are intervals on contigs for which every base is covered by at least one protein mapping. This definition of PExR is useful even if we expect a single orthologous protein region to be homologous and map to the entirety of a PExR. I found that the mapping

of orthologous proteins suffered from edge effects, i.e. orthologous regions with little similarity can only be aligned if they are flanked by orthologous sequences showing high similarity. Because of the lack of potential flanking region with high similarity at the edges of orthologous regions, there is a drop in sensitivity in the detection at the edges of these regions. Furthermore, splice variants can have exons with alternative 3' and 5' splice sites. By combining all protein mappings of the same contig region, SWiPS was able to predict regions coding for exons with higher specificity.

Scaffolding contigs by optimization of overall protein to contig mappings

At this stage, SWiPS has a refined mapping of proteins to contigs. Using these mappings, SWiPS computes a similarity score for each protein-contig mapping by summing the bit scores across the aligned regions defined by genewise according to a distance matrix (BLOSUM62 in our case). SWiPS now must order the contigs into scaffolds, by defining orthology relationships between the template protein and contigs mapping to different regions of that protein. In cases where there is a clear one-to-one orthologous relationship between the guide protein and its cognate in the newly sequenced genome, this task is conceptually straightforward. SWiPS, however, must be able to disentangle orthology relationships within gene families using fragmentary sequences. This complicates the procedure considerably.

The protein-contig similarity scores only give us a partial picture of the orthology relationships between proteins and contigs. The distribution of similarity scores is strongly correlated with their lengths, and the strength of stabilizing selection can be highly variable across loci. Therefore, orthology relationships

between a particular protein and contig cannot be called without considering all other mappings between the contig and other proteins. In order to take into consideration all mappings at once, the problem of scaffolding contigs is posed as an optimisation problem for which the similarity score between contigs and guide proteins is maximised without creating inconsistencies. That is, SWiPS assign contig gene models, i.e. regions of a contig that are predicted to be the exons from a single gene, to protein regions so that (1) each pair of contig gene model and protein region is the pairing most likely to be orthologous (according to similarity scores), (2) there are no two contig gene models associated to the same protein regions, and (3) as few proteins are used as possible as scaffold models. The reasons for (1) and (2) should be fairly obvious, as contigs can be scaffolded together (and extend protein gene models) only if they contain PExRs orthologous to different regions of the same protein. Importantly, it can be seen that the implementation of condition (1) in SWiPS is analogous to the best reciprocal hits approach to identifying orthologous proteins. Condition (3) is necessary for resolving the problem of contigs mapping to legitimate orthologs from different proteomes. By using as few protein scaffolds as possible, the spread of contig mappings over multiple protein orthologs or splice variants is reduced. In SWiPS, the implementation of these ideas is summarised by the definition of two concepts, the scaffolding power and the maximal scaffoldable set, which are described next.

Scaffolding power and Maximal scaffoldable set

For each potential guide protein, I retrieve all refined contig mappings and define regions of the guide protein into which the contigs cluster (Figure 5). To allow for mapping edge effects, contigs cluster together if they overlap by 15 nucleotides or more. In all contig clusters, contigs which have a low similarity

score are removed if they are lower than α times the highest similarity score (default $\alpha = 0.75$) (e.g. contig 1 in Figure 3). The mappings of multiple contigs to the guide protein can then be viewed as a graph (Figure 5). The score of all possible paths through this graph is calculated by exhaustively combining all contigs from different regions and summing the individual contig-protein match scores to get an overall score for each path. All paths having scores with at least β (default $\beta = 0.5$) times the highest scoring path are then compared, and contigs present in all higher scoring paths are kept and form a maximal scaffoldable set, with a score referred to as the scaffolding power. The default values of $\alpha = 0.75$ and $\beta = 0.5$ were chosen to be conservative, and make SWiPS less likely to produce chimeric sequences from different paralogs within the target genome. These two parameters, and how they affect the performance of SWiPS in the presence of paralogs, are discussed next. However, I tested several values for α and β on the *Ciona intestinalis* runs and found that, for both, increased values resulted in increased numbers of scaffolded contigs (and N50), but also in increased rates of scaffolding error.

Contigs encoding in-paralogs, i.e. genes with many-to-one or many-to-many relationships with members of the guide proteins, cannot be reliably scaffolded because they result from one or more duplication events subsequent to the divergence from genes in the guide proteome. Contigs resulting from duplication events prior to the divergence from genes in the guide proteome (out-paralogs), however, can be scaffolded. In SWiPS, α is used to uncouple contig-protein pairings that have lower similarity scores than the one with the highest score (e.g. contig 1 and P1 in figure 5). For example, an $\alpha = 1$ forces every contig to be coupled unambiguously with the protein with which it shares the highest similarity score, whereas an $\alpha = 0$ implies that all contig-protein pairings are

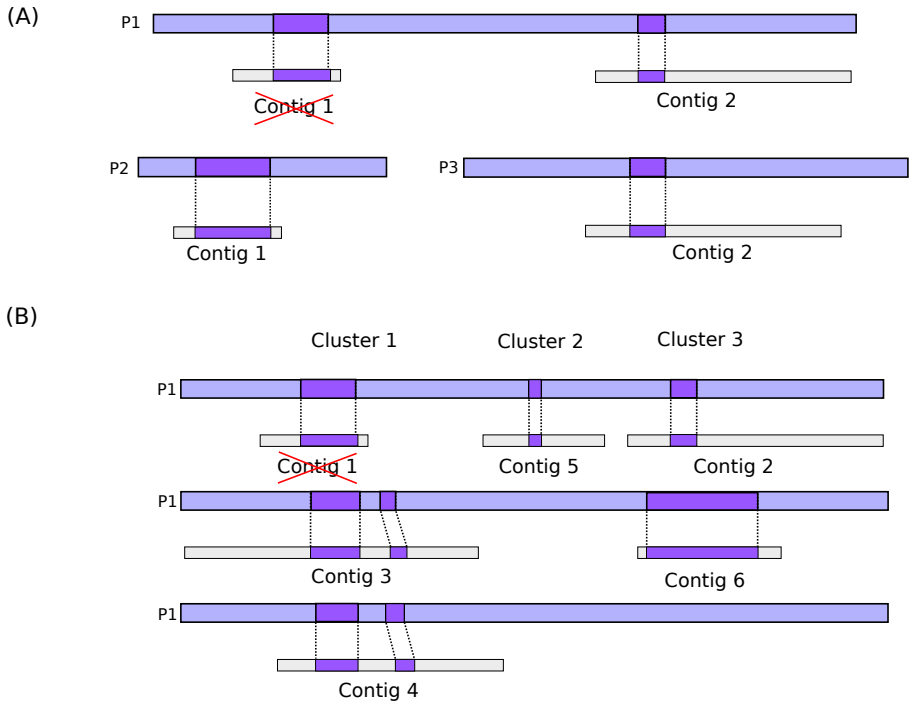

Figure 5: Determining protein-contig mappings.
(A) Although contig 1 and 2 are mapped to protein P1, they have better mappings on protein P2 and P3, respectively. The similarity score between contig 1 and P1 is lower than $\alpha = 0.75$ times the similarity score between contig 1 and P2, and is removed from P1. The similarity score between contig 2 and P1 is within α of the similarity score between contig 2 and P3, and both mappings are kept. (B) Contigs with similarity to P1 are clustered according to the region of similarity. Here, contigs 3 and 4 are clustered together in contig cluster 1, contig 5 is the only contig in contig cluster 2, and contig 2 and 6 are clustered in contig cluster 3.

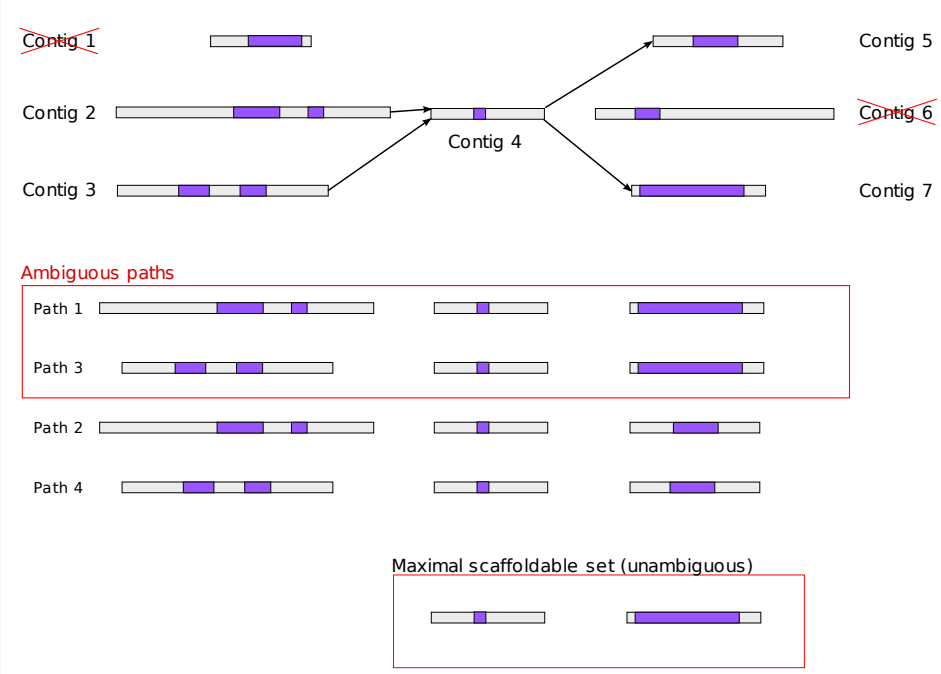

Figure 6: Finding the maximal scaffoldable set.
The scores for all possible combinations of contigs in each region are computed (including combinations with no contigs from clusters), and the lowest scoring paths are removed. High scoring paths (i.e. paths with at least $\beta < 1$ times the maximal scoring path) are then compared; contigs present in the highest scoring (ambiguous) paths are kept and determine the maximal scaffoldable set.

kept for subsequent steps. Consequently, a high α can be used when a complete set of closely-related proteins is used as guide proteome, and contigs encoding out-paralogs will be coupled unambiguously with their orthologous proteins. However, if the set of closely-related proteins is incomplete, a high α may result in erroneous contig-protein couplings that are considered unambiguous, thereby leading to scaffolding errors. My choice of 0.75 as the default value for α is thus conservative and allows the use of incomplete protein sets.

While α takes incomplete protein sets into consideration, our maximal scaffoldable set approach (which relies on β) takes into account contigs encoding in-paralogs that cannot be scaffolded. If two contigs encode in-paralogous sequences (contig 3 and 4 in Figure 5B and 6), they will form distinct paths that have similar scaffolding powers (e.g. path 1 and 3 in 6). Therefore, they will be excluded from the maximal scaffoldable set (Figure 6). Because the number of in-paralogs are potentially high, I chose β conservatively.

Once the scaffolding power of each protein is computed, the protein scaffold with the highest scaffolding power is iteratively picked, and the maximal scaffoldable sets (and scaffolding power) of all other proteins are updated so that a contig gene model (the contig region) is only used once. Picking the protein scaffold with highest scaffolding power ensures that as many contigs as possible are linked. After all guide proteins above a minimum scaffolding power have been chosen, the scaffolded contigs go through another scaffolding phase in which contigs with multiple gene models are used to merge the scaffolds together. The relative positions of contigs is clear when a single protein maps to them because of the linear structure of a gene. However, the loci of some proteins overlap and inferring the relative positions of these different contigs can be problematic. To

deal with simple cases, directed graphs are constructed according to the relative mapping position of the contigs on the proteins. The order in which the contigs are traversed in these graphs determines the contig ordering of the scaffolds. After this scaffolding step, SWiPS takes all scaffolded contigs and uses genewise to build the final gene models.

Genome assembly data and guide protein sets

The *Ciona intestinalis* genome sequence, version 2.0 (Dehal et al., 2002) was retrieved from Ensembl (release 63). 80x coverage Illumina 50bp paired-end reads were simulated using simLibrary and simNGS (http://www.ebi.ac.uk/goldman-srv/simNGS/) with default parameters. *Ciona intestinalis* was chosen owing to our underlying interest in deuterostomes and for the pragmatic reasons that it is small and has not undergone a whole genome duplication. SOAPdenovo was used with default parameters to assemble the simulated *Ciona intestinalis* reads (Li et al., 2010). For linguistic convenience, I refer to all SOAPdenovo generated sequences as contigs, even though these sequences include contigs that have been scaffolded together using the insert size distributions of paired end sequence data.

Next-generation sequencing paired-end data from *Drosophila melanogaster* DNA sequence were obtained from the short read archives (accession SRX021790; Mackay et al. (2012)), and were used to build an assembly (N50=1441bp) with SOAPdenovo. The human (NA18507) genome assembly (N50=61,980bp), built using SOAPdenovo, was obtained from `http://yh.genomics.org.cn/download.jsp`.

For the comparison with ESPRIT (Dessimoz et al., 2011), I downloaded the genomic assembly of *Callorhinchus milii* from `http:`

//esharkgenome.imcb.a-star.edu.sg. This 1.4x assembly consists of 647,131 contigs (754MB) with an N50 of 1464bp.

All protein sets used were retrieved from Ensembl, except for *Strongylocentrotus purpuratus*, and *Saccoglossus kowalevskii* (Genbank); and *Branchiostoma floridae*, *Drosophila grimshawi*, *Heliconius melpomene*, *Culex quinquefasciatus*, *Anopheles gambiae*, *Anopheles merus*, and *Anopheles quadriannulatus* (Uniprot).

Estimating errors

All simulated contigs were mapped to the *Ciona intestinalis*, *Drosophila melanogaster*, or *Homo sapiens* reference genome (all three from Ensembl) using BLASTN and assigned a reference scaffold, strand, and position. SWiPS predicted scaffolds were considered correct if their mapped contigs are ordered correctly (within 100kb of each other as eukaryotic introns are rarely over 100kb; for human, I used 200kb) across the reference scaffold on the same strand. Note that contigs are allowed to be intercalated between exon-containing scaffolds without affecting the assessed accuracy. I used two measures of accuracy: 1) the percentage of contig joins that are correct - local link correctness, and 2) the percentage of scaffolds (with at least one contig link) for which all local contig links are correct.

2.1.3 Results

Mapping ability with proteins from the same species

The extent to which contigs can be scaffolded by protein sequences will depend on the amount of protein coding content in the genome, and the ability to assign contigs to particular protein coding sequences. The latter will

depend on the extent of similarity between the guide proteins and the genome under consideration. In order to understand the best possible outcome for the scaffolding process, I first tested SWiPS's ability to assemble a genome using a protein set from the same species as the genome.

I created a *Ciona intestinalis* genome assembly of the kind that might be produced using a simple next generation sequencing strategy, with low N50, and applied SWiPS, using *Ciona intestinalis* proteins as guides for the scaffolding process.

My algorithm improved the N50 of the simulated assembly from 3,851bp to 6,245bp (62.2% increase) and reduced the number of scaffolds from 108,424 to 94,680 with 4134 out of 4510 (91.66%) scaffolding to be entirely correct (it should be noted that the parameters used in this analysis were not optimised to take into account that the guide proteins were of the same species).

Use of orthologous protein sequences & dependence on read depth

As a more realistic test, I scaffolded simulated *Ciona intestinalis* assemblies using orthologous protein sequences from *Danio rerio*, *Strongylocentrotus purpuratus*, *Gallus gallus*, *Homo sapiens* and *Saccoglossus kowalevskii*. To test the dependency on sequencing depth I produced assemblies based on the *Ciona intestinalis* genome at 10x, 20x, 40x, and 80x coverage.

Irrespective of coverage depth, SWiPS was able to produce a $> 20\%$ improvement in the N50 of the genome assembly, with a corresponding increase in the numbers of scaffolded contigs (Table 1). These numbers show that SWiPS appears to be useful for improving assembly quality for even low coverage (by

next generation standards) datasets. Figure 7 shows an example of a scaffolded region, with two proteins, one from Sea Urchin and one from *Saccoglossus* providing evidence to link five contigs from the SOAPdenovo assembly.

To test SWiPS on real next-generation data, I produced a *Drosophila melanogaster* assembly from Illumina Genome Analyzer II sequence data. Using protein sequences from *Drosophila grimshawi*, *Heliconius melpomene*, *Culex quinquefasciatus*, *Caenorhabditis elegans*, *Anopheles gambiae*, *Anopheles merus*, and *Anopheles quadriannulatus*, SWiPS produced an improvement of 11.1% on the assembly N50 and reduced the number of contigs from 130,312 to 123,309 (Table 2.1.3). I also tested SWiPS on a human assembly using protein sequences from *Canis familiaris* and *Mus musculus*. In this case, SWiPS produced an improvement of 18.2% on the assembly N50 and reduced the number of contigs from 314,877 to 306,882. I also found that, when allowing only 100kb between scaffolded contigs, the assessed scaffold correctness was 70.6% and the local scaffold correctness was 90.3%. However, allowing 200kb (500kb) resulted in an assessed scaffold correctness of 85.3% (92.1%) and a local scaffold correctness of 96.0% (98.3%).

I also tested SWiPS on low coverage Sanger sequenced data, in the form of the 1.4x coverage genome of the elephant shark *Callorhinchus milli* (Venkatesh et al., 2007). I saw an improvement in N50 from 1464bp (647,131 scaffolds) to 1506bp (630,493 scaffolds), a 2.9% increase. This difference in N50 improvements compared to the one for *Ciona intestinalis* is mainly due to the difference in contig length distribution.

Comparison with other methods

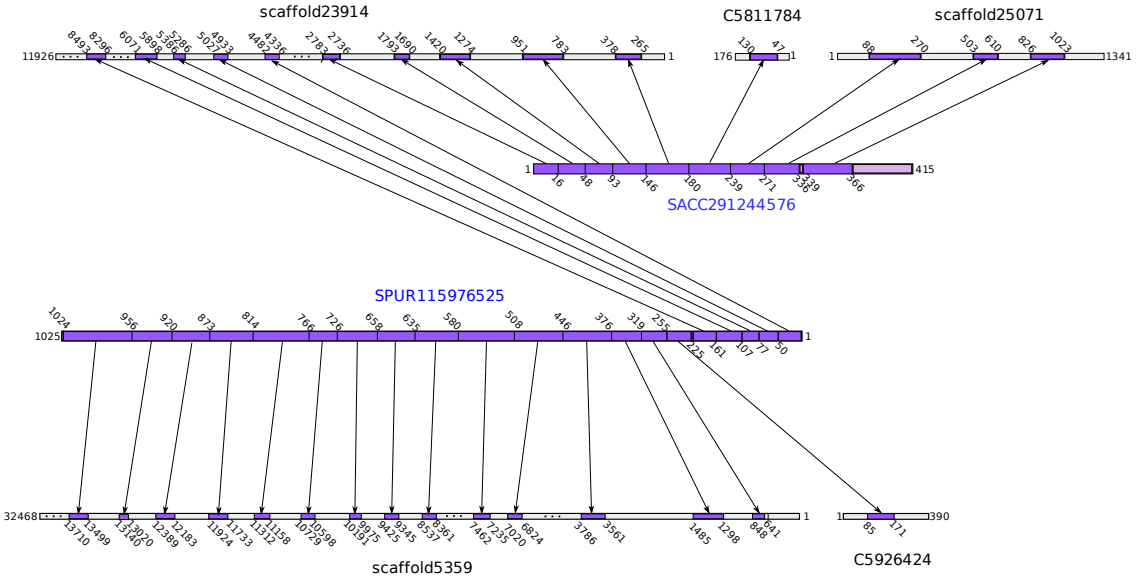

Figure 7: Example of 5 contigs scaffolded using SWiPS.
Protein SACC29144576 links contigs "scaffold23914", "C5811784", and "scaffold25071" together, while SPUR115976525 links contigs "scaffold23914", "C5926424", and "scaffold5359" together. The two protein models allow SWiPS to scaffold all contigs together. Each arrow represents a region of homology (in purple); the intron-exon structure (gray and purple) of the orthologous proteins is clearly represented within each scaffold.

Assembly (coverage)	N50 improvements (bp)	CEGMA improvements (complete/partial genes)	Scaffold correctness	Local link correctness
Ciona 10x	1767 to 2101 (18.9%)	95 to 136 (41.2%) 140 to 190 (35.7%)	88.3%	97.05%
Ciona 20x	3714 to 4544 (22.3%)	134 to 166 (23.9%) 175 to 204 (16.6%)	87.53%	97.42%
Ciona 40x	3799 to 4790 (24.4%)	138 to 166 (20.3%) 169 to 210 (24.3%)	85.65%	96.78%
Ciona 80x	3851 to 4719 (22.5%)	139 to 162 (16.4%) 171 to 210 (22.8%)	89.43%	97.60%
Drosophila melanogaster	1441 to 1601 (11.1%)	178 to 201 (12.9%) 241 to 245 (1.7%)	98.9%	99.7%
Human (NA18507)	61,980 to 73,255 (18.2%)	95 to 97 (2.1%) 199 to 206 (3.5%)	85.3%[*]	96.0%[*]
Elephant Shark	1464 to 1512 (3.3%)	7 to 22 (214.3%) 49 to 107 (118.4%)	N/A	N/A

The only other program currently available to scaffold genomic contigs using protein sequences is ESPRIT (Dessimoz et al., 2011). To directly compare SWiPS with ESPRIT, I applied the CEGMA pipeline to the *Callorhinchus milii* reference genome. CEGMA uses a set of 248 core eukaryotic genes that are generally present in low copy number to assess the completeness and quality of a genome assembly, and can serve as a complementary metric to the N50 length (Parra et al. (2007); also see Genome Assembly section in Chapter 1).

In my hands, the CEGMA pipeline detected 7 complete and 47 partial genes when assessing the initial reference genome, while Dessimoz and colleagues reported 14 complete and 35 partial genes. This difference is likely due to the use of a more recent version of CEGMA. The overall similarity of the results, however, suggests a comparison of the performance of SWiPS and ESPRIT is valid.

To link *Callorhinchus milii* contigs together with SWiPS, I used the same set of proteins used by ESPRIT, which consists of human, mouse, anole lizard,

chicken, African clawed frog, zebrafish, medaka, *Ciona intestinalis* and *Branchiostoma floridae* proteins. After linking 666 contig pairs with ESPRIT, Dessimoz et al. reported an increase from 14 to 16 complete genes and from 35 to 38 partial genes, i.e. a 14.3% and a 8.6% improvement, respectively. In comparison, SWiPS was able to merge 27,659 contigs into 9,121 scaffolds, and increased the CEGMA gene space from 7 complete genes to 22 and 47 partial genes to 107, i.e. a 214.3% and a 118.4% improvement, respectively. Even using Dessimoz and colleagues' higher baseline of 14 complete genes, the new assembly produced by SWiPS improved significantly the complete gene space when compared to ESPRIT (8 compared to 2, i.e. a 4-fold improvement).

2.2 Discussion

I have shown that SWiPS is able to use protein sequences to order and orient contigs into scaffolds, thus improving assembly contiguity and gene prediction, showing significant improvements over the only other available tool. Although the overall improvements in N50 found by SWiPS may appear modest, they should be understood with reference to the fact that the majority of most animal genomes is composed of non-coding sequence, and thus not amenable to scaffolding via protein coding sequences.

In order to confidently scaffold exon-containing regions on different contigs, we need to be sure that they come from the same gene. In the context of this problem, the most obvious way of doing this is by assuming that the guide protein comes from a gene that shares a 1:1 or many:1 orthologous relationship with the gene encoded by the exons. That each exonic region on a contig is orthologous to the guide protein needs to be tested. Rather than constructing a full phylogeny for each exon region, I maximize the overall similarity score

of templates to exons for all proteins that match equivalent sets of regions. As the ability to discriminate orthologs from paralogs is fundamental to my method, SWiPS is expected to perform best for proteins that have no similar sequences in the genome, and less well where proteins come from gene families with multiple closely related members. Furthermore, the more closely related the template set of proteins is to the target genome, the less likely my method is to be confounded by gene duplication events that have occurred after the template and target lineages have diverged.

Although inference of orthology is most obvious via overall comparisons of sequence similarity, it is conceivable that other sources of information may be useful. For instance, conservation of exon/intron boundaries may help to discriminate between orthologs and paralogs in cases where orthologs share intron locations to the exclusion of paralogs. In practice, I was able to find few examples where this provided a significant improvement on my current results (data not shown; also see gene architecture evolution in East African cichlids in Chapter 4).

For many applications, the primary interest of a genome sequence lies in its encoded proteins. Coupled with the difficulty of *de novo* genome assembly and gene prediction, this suggests that transcriptome sequencing may be a more useful strategy for generating an initial survey of genome content. Transcriptome sequencing, however, has major disadvantages compared to genome sequencing. Firstly, not all genes are expressed in all developmental stages and cell types, and for many taxa, it may be difficult to sample sufficient transcriptional libraries to obtain representation of all genes. Without this, it is difficult to make reliable inferences of the absence of particular genes. Secondly, genome sequence contains a wealth of information not encoded in the transcriptome, in the form of

regulatory elements, synteny information, gene structures and so on. Preliminary genome sequence is thus likely to be of greater long term usefulness than incomplete transcriptome data. Although genome assemblies produced from single libraries of Illumina data (i.e. with only one insert size) are likely to be very fragmentary (whatever the sequencing depth), protein coding content is likely to be depleted in the repetitive sequence that causes assembly problems, at least at the exon level. By allowing the scaffolding of exons on different contigs my method allows the maximum usefulness in terms of protein coding content to be extracted from these preliminary assemblies.

2.3 Sequencing DNA and RNA from a single cell

High-throughput sequencing is one of the most influential technological advances in the field of biology. With it, genomes and transcriptomes of hundreds of species are being sequenced in the hope of gaining insights into diverse adaptive traits or evolutionary history. Not only does the amount of sequencing increase rapidly in breadth, but also in depth. For example, the Genotype-Tissue Expression project (GTEx) hosts the transcriptomes of up to thirty different tissues or organs from 150 individuals. Now, researchers continue to reduce the scale at which we study biology by analysing single cells.

The cellular landscape across different tissues can vary substantially, and a significant proportion of the variation is conserved in mammals (Brawand et al., 2011). These conserved differences suggest the existence of important tissue-specific gene regulatory networks. This is rather unsurprising considering that different organs are well characterised to possess different functional roles. However, recent analyses of transcriptomes in single cells of seemingly homogeneous populations of cells and tumours revealed a surprisingly large amount of heterogeneity (Shalek et al., 2013; Dalerba et al., 2011). This heterogeneity was previously overlooked, owing to technical difficulties in sequencing low amount of DNA or RNA: nearly all studies to date rely on the quantification of molecular phenotypes of thousands to millions of cells in bulk. However, advances in microfluidics and DNA amplification technologies allow both DNA and RNA to be quantified from single cells (Macaulay and Voet, 2014).

A new method to sequence the DNA and RNA from a single cell, GT-seq, is currently being developed by Macaulay et al. (see Materials and Methods). My goal was to help evaluate the data produced by GT-seq and explore potential

uses of such methods. Here I present analyses I conducted to assess GT-seq compared to other methods, and to evaluate a potential future use. Although the work I present here is largely technical in nature, I will describe applications of single cell sequencing in a biological context in the Discussion chapter.

2.3.1 Results and Discussion

Single cell transcriptomics

One of our objectives was to quantify the transcript diversity that GT-seq is able to capture. This is important because previous single-cell transcriptomics studies often suffered from a poor sensitivity to lowly expressed transcripts (Macaulay and Voet, 2014). Macaulay and colleagues (unpublished) sequenced breast cancer cell line (HCC38) transcriptomes with Illumina (11 cDNA libraries prepared according to GT-seq; 96 libraries prepared using Fluidigm C1), and PacBio (4 normalized cDNA libraries prepared according to GT-seq) technology (Materials and Methods). In this case, the normalisation of the cDNA library is appropriate as we aimed to characterise the different transcripts captured by GT-seq.

We used two approaches to measure transcriptome complexity. A colleague measured the number of genes that are detectable by counting genes with a minimum threshold of sequencing reads mapped (data not shown). To complement this approach, I measured the amount of transcriptional complexity by the number of genes with splice junctions detected using each method.

I first started by assessing the transcriptional complexity using PacBio sequencing. Although PacBio sequencing yields low throughput relative to other

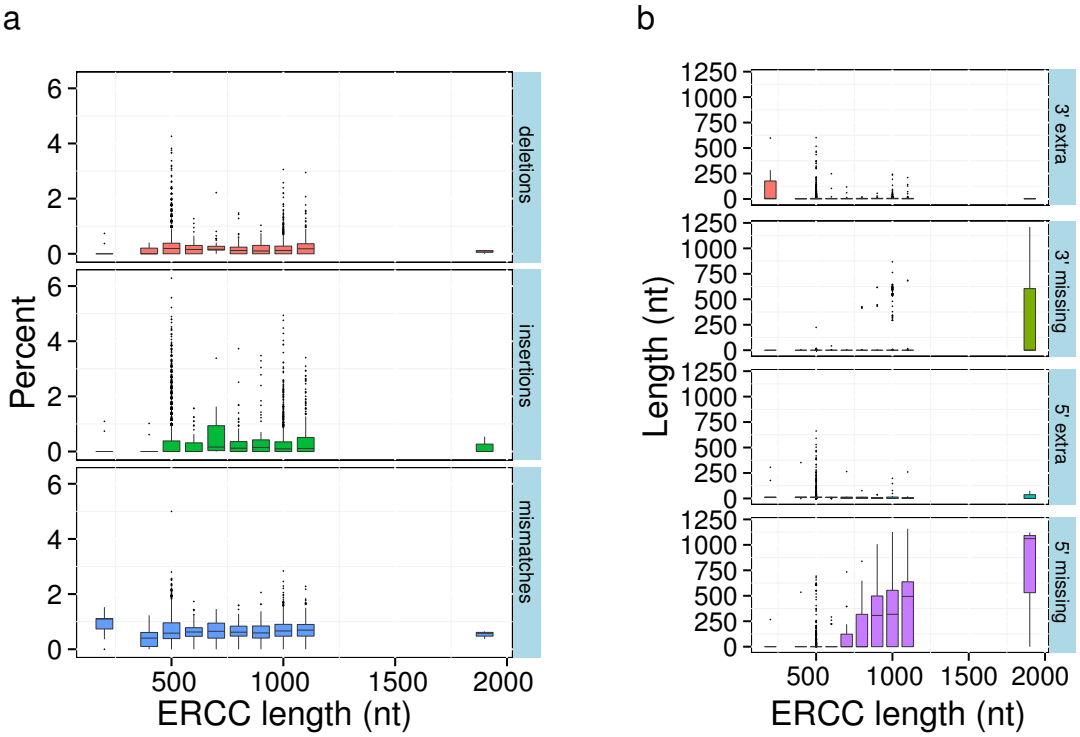

Figure 8: PacBio sequencing errors by length.
(a) Average number of insertions, deletions or point mutations by 100nt from PacBio sequencing. (b) Sequence appended or missing from the 5' or 3' end of the ERCC spike-ins.

popular sequencing methods, it is reported to produce long reads of over 1kb that allow splice junctions to be predicted easily. Owing to high reported error rates of PacBio, I wished to quantify the error rates in our PacBio dataset by first mapping the ERCC spiked-in reads onto the spike-in reference sequences and then by identifying differences between the reads and reference sequences. As mentioned previously, spike-in sequences are known and designed to be different from sequences in the human genome (Jiang et al., 2011), therefore differences between read and reference sequences are highly likely to correspond to sequencing errors. I used all 9,536 PacBio reads that mapped to ERCC transcripts to quantify PacBio error rates and sequencing performance. I found the number of mismatches to be generally under 2% (Figure 8a). Furthermore, the numbers of insertions and deletions are also low – well under 2 insertions

or deletions per 100bp (Figure 8a). I also observed that although the lengths of the reads are significantly longer than standard Illumina reads (39–100bp), many reads were truncated at their 5' end when they originated from ERCC transcripts of length 700bp or longer (Figure 8b). This, however, likely reflects a technical artefact from size selection (personal communication, Macaulay) and not a limitation of PacBio sequencing (Sharon et al. (2013) report PacBio read lengths that average 1.5kb).

Next, I used GMAP to map trimmed PacBio reads onto the human reference genome. The adapter sequences from PacBio reads were trimmed using a seed matching strategy (Materials and Methods). As such, adapter sequences were identified in 88.5–89.7% of the reads (4 sequencing runs); the remaining 11–12% were removed from further analysis. Because the human genome is repeat-rich and comprises of many genes with paralogs, PacBio reads were sometimes found to map to multiple ambiguous location. Therefore, in instances of reads mapping to multiple locations, I only kept the mapping with highest percent identity. In total, 152,355 PacBio reads were mapped onto the human genome assembly (hg19).

To assess the gene coverage of PacBio sequencing, I computed the number of genes with at least one splice junction supported by a read. Of 18,756 multi-exonic protein-coding genes annotated in Ensembl, PacBio libraries supported 1,546–1,718 genes, and 3,232 genes were supported when the four libraries were pooled together. This number is far lower than the maximum number expressed according to bulk sequencing of HCC38. Indeed, counting the number of genes supported by at least one splice junction in the bulk dataset yielded 13,235 multi-exonic genes (Figure 9). This means that additional PacBio sequencing

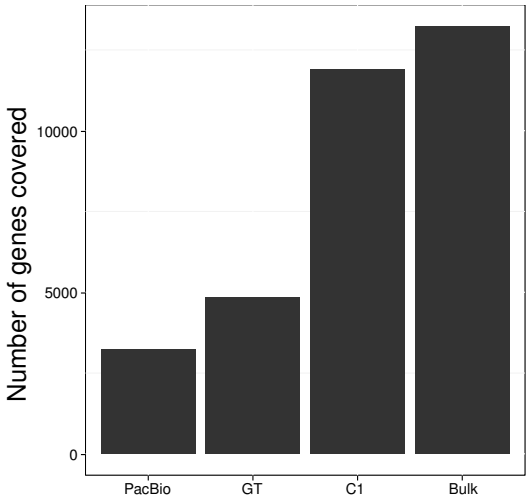

Figure 9: Gene coverage by method.
Number of genes with at least one splice junction supported. These numbers are calculated from pooled samples from each method. Neither the sequencing depth nor the number of libraries are normalised here. Sequencing by PacBio, GT-seq, C1 Fluidigm, and bulk sequencing of breast cancer cell line (HCC38).

must be performed to achieve a comprehensive transcriptional landscape in single cells. This low gene coverage may also reflect a current limitation of GT-seq. Indeed, Sharon et al. (2013) sequenced a pool of 20 different normal human tissues and organs with PacBio technology (yielding 476,000 reads) and identified nearly 14K multi-exonic genes.

Using this same metric, I next assessed the gene coverage of GT-seq and C1 protocols. To do this, all reads from GT-seq and C1 protocols were mapped onto the human genome (hg19) complemented by annotated exon junction sequences as before (Piskol et al., 2013). The number of unique splice junctions predicted ranged from 3,545 to 10,390 (median 5,854) for the 11 GT-seq samples, while it ranged from 172 (several libraries yielded very few reads) to 32,371 (median 24,561) for C1 samples. This difference cannot be explained by sequencing depth since GT-seq samples yielded more sequencing reads (median 7.1M reads) than C1 samples (median 2.6M reads). This therefore suggests that libraries prepared

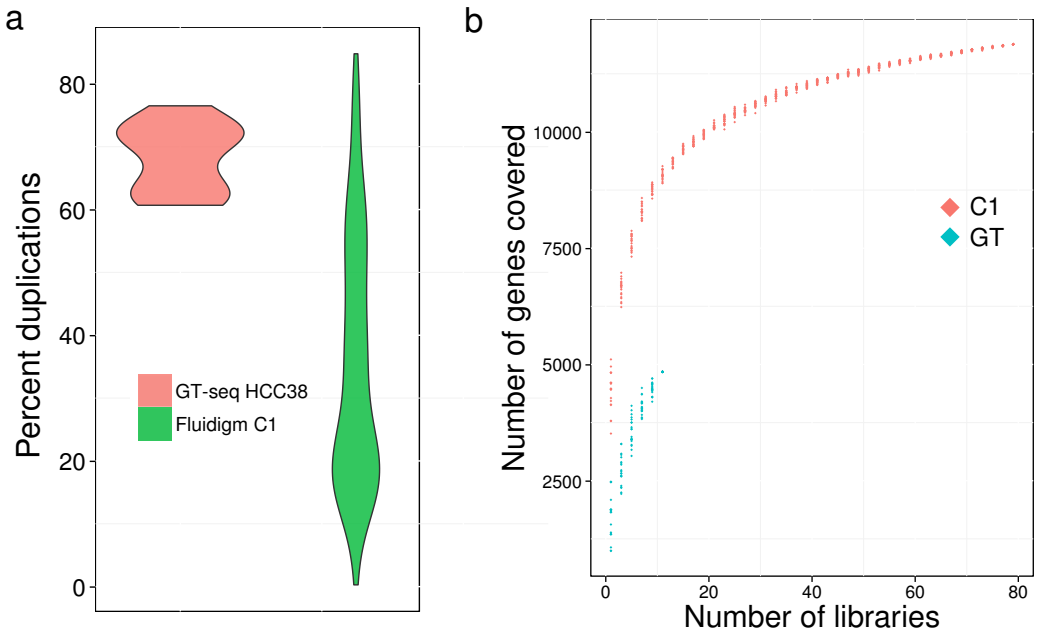

Figure 10: C1 versus GT-seq duplications and coverage.
(a) Duplicated reads in single cell sequencing. Percentage of reads marked as duplicates by MarkDuplicates of PicardTools from HCC38 (11 samples; red) sequenced with GT-seq and HCC38 single cells libraries (94 samples; green) sequenced with C1 Fluidigm. (b) Number of genes with at least one junction covered by reads from C1 versus GT-seq protocols.

using GT-seq contain lower transcriptome complexity than libraries prepared using the C1 protocol. Upon investigation, I found that a large percentage of reads (median 70.21%) in GT-seq sequenced reads were classified as duplicates by Picard tools (`http://picard.sourceforge.net/`), while a much smaller percentage (median 30.4%) were marked as duplicate in C1 sequenced reads (Figure 10a). This likely reflects a larger percentage of redundant reads (i.e. reads from the same transcripts or PCR artefacts) in GT-seq compared to C1 sequencing and explains, at least partially, the reduced complexity of the transcriptome captured by GT-seq.

To obtain an estimate of the transcriptome complexity captured by pooling C1 or GT-seq samples, I computed saturation curves by pooling randomly-

chosen samples together (Figure 10b). Due to the large differences in detected splice junctions among C1 samples, I removed 17 libraries for which fewer than 7,000 unique splice junctions were detected. I found that, on average, by pooling 11 randomly picked C1 samples, the number of genes covered increases to 9K (from 3-5K in one sample), and slowly continues to increase to 11,885 when all 79 C1 samples are pooled together. In contrast, pooling all 11 GT-seq samples resulted in 4,842 genes with splice junction support. Again, these results suggest that a higher number of single cells are needed to recapitulate the transcriptome complexity of HCC38 cells when using the GT-seq protocol in its current form.

These results also suggest that C1 technology can be used to study transcriptomes of single cells because 11,885 genes (compared to 13,235 from bulk sequencing) had a splice junction covered when all samples were pooled together, albeit 79 single cells (totalling 288M reads) had to be sampled to reach this number.

Shared DNA and RNA variants

Sequencing both DNA and RNA from a single cell provides the interesting opportunity to validate DNA variant calls (SNPs) from RNA sequence. Assuming a mutation occurring in a RNA library to be 1 in 1000 bases, the joint variant calling using both DNA and RNA data may reduce false positives by a thousand-fold. An exciting application of this lower false positive rate lies in the identification of somatic variation. Other experimental designs may uncover somatic mutations with low false positives by sequencing DNA alone or by sequencing both RNA and DNA from bulk cells. However, sequencing both DNA and RNA from a single cell permits us to study mutations that occur during a time-frame as short as one cellular division, or mutations in rare cells

that cannot be cultivated *in vitro*.

To verify that joint calling is feasible, I first wished to ensure that variant calls from single cell RNA-seq data are reliable. This question is not trivial because the starting RNA quantity in a single cell is low, which allows mutations to be duplicated in the PCR amplification steps and subsequently sequenced with high coverage. For RNA variant calling, I used all 96 HCC38 single-cells sequenced with C1 Fluidigm technology. This is because GT-seq is under active development and currently produces transcriptome data of poorer quality compared to C1 protocols. Nevertheless, variants identified from RNA-seq data produced by C1 is likely reproducible by GT-seq or similar methods in the future.

To identify variants from the C1 RNA-seq data, I used the pipeline introduced in (Piskol et al., 2013). This pipeline uses the GATK variant calling approach and further verifies that variants do not occur in repeated regions, homopolymers or RNA-edited sites. I therefore obtained a set of high-confidence variants for each library. I then compared these variants to a comprehensive list of all HCC38 variants (personal communication, Dabas). True RNA variants are thus expected to be replicated in this list. Indeed, I found that nearly all variants ($> 95\%$) identified in two or more cells could be replicated in the list of HCC38 variants called from bulk DNA sequencing (Figure 11). Variants called in two or more samples are therefore expected to be real variants in HCC38 cells. However, only 35.7% of the variants could be replicated if they were identified in one cell alone. Because most variants are only called within one sample (84,610 or 83.1%), around half (46.2%) of all RNA variants called are known DNA variants. Nearly all of the remaining 53.8% are expected to reflect technical artefacts. Nonetheless, a few of these variants may represent true somatic mutations.

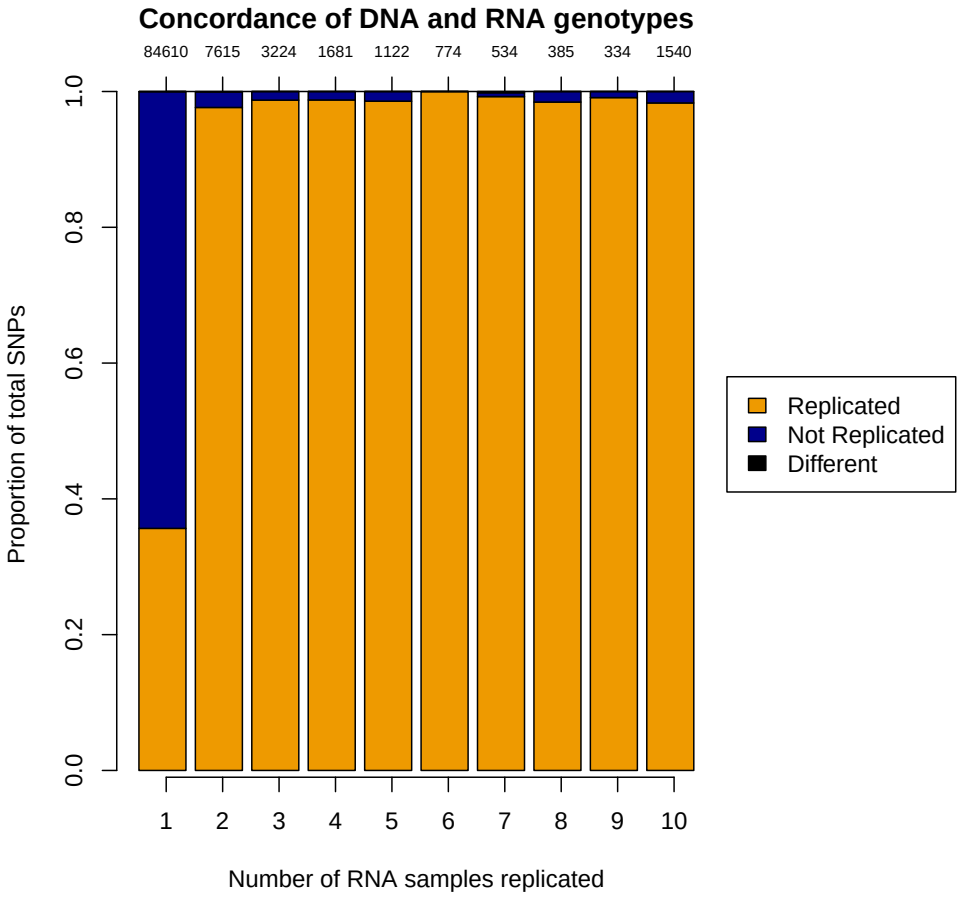

Figure 11: Concordance of RNA variants call with HCC38 variants.

Using protocols such as GT-seq, somatic variations may be detected if RNA variants from a single cell can be replicated in its DNA. Assuming the probability of a DNA sequencing error is 1 in 1000, a rough estimate yields a false positive rate less than 5.38×10^{-4} (0.538×0.001) for RNA variants replicated in DNA sequences. This estimate, of course, assumes independence of samples and the absence of mapping artefact.

Overall, these results suggest that somatic variation can be called in single cells by combining RNA-seq data and DNA sequence data using GT-seq. Cur-

rently, single cell RNA-seq and whole genome DNA-seq data produced by GT-seq are too low to reproduce variant calls (data not shown). However, it is likely that using exome sequencing will allow variants to be called in transcribed DNA regions, which in turn will allow variants to be replicated in the RNA and DNA of a single cell.

2.4 Material and Methods

Combined transcriptome and genome analysis of a single cell (GT-seq)

Single cells were isolated directly into lysis buffer, from which mRNA was first captured using a modified oligo(dT) primer, containing a universal priming site and conjugated to a magnetic bead. The supernatant, containing genomic DNA and non-polyadenylated transcripts, was then removed. The beads were subsequently washed four times, and the washes pooled with the original supernatant.

First strand cDNA synthesis and subsequent PCR-based amplification were then performed directly on captured RNA following the SMARTer protocol (Ramskold et al., 2012). Genomic DNA was isolated from the supernatant and wash pools by ethanol precipitation. Subsequently the precipitated genomic DNA was amplified using the MALBAC protocol (Zong et al., 2012).

Sequencing libraries were prepared from amplified gDNA and cDNA using the Nextera XT kit and following the manufacturer's instructions (Illumina) Multiplexed libraries were sequenced in fast mode on an Illumina HiSeq 2500.

For full-length cDNA sequencing on the Pacific Biosciences RSII, SMRTbell

libraries were prepared from the cDNA as per the manufacturer's protocol for a 2 kb library. One SMRT cell was used to sequence a library from each single cell.

C1 sequencing

cDNA was prepared from single cells using the Fluidigm C1 platform. Briefly, a suspension of HCC38 cells was loaded onto a microfluidic chip, single cells were captured in individual capture sites, lysed and cDNA prepared on-chip using the SMARTer amplification protocol (Clontech/Takara). Nextera library preparation was performed on the cDNA as previously described, the individual libraries were then pooled and sequenced (100bp, paired end) on two lanes of HiSeq 2500.

Trimming PacBio reads

To find the 5' (AAGCAGTGGTATCAACGCAGAGTACATGGG) and 3' (GTACTCTGCGTTGATACCACTGCTT) SMRT primers from the PacBio reads, I searched for 15-mers shared between the primers and the reads using a sliding windows approach. The read was then trimmed to the last 5' and first 3' primers found (from a 5' to 3' scan). Reads were discarded in cases where either 5' or 3' primers were not found, or the 5' primer was found to be 3' of the 3' primer.

Counting duplicated reads

To find duplicated reads in each library (.bam file), the MarkDuplicates program from Picard tools was used: *java -Xmx4G -XX:MaxPermSize=512m -jar MarkDuplicates.jar AS=true VALIDATION_STRINGENCY=SILENT.*

3 Chapter 3: Gene family evolution and expansion

Owing to the importance of gene duplications in phenotypic innovation (Kaessmann, 2010), I became interested in gene families that underwent significant changes in clades exhibiting particular phenotypic traits. Over the last few years, I studied gene duplications in two newly assembled genomes: that of the painted turtle and that of the bowhead whale. The painted turtle genome was the only Chelonian (turtle) genome sequenced (at that time). Therefore the study of painted turtle genes promised to shed light on over 250 million years of gene family evolution during Chelonian evolution, subsequent to their split with the bird lineage.

In this chapter, I present an in-depth analysis of the evolution of a specific gene family, the beta-keratins, in the Chelonian lineage. The study of beta-keratin gene family evolution in Sauropsids (reptiles and birds) revealed a large expansion in the number of turtle-specific beta-keratins. I draw links between this expansion and the birth of the unique carapace and plastron of turtles. Analyses of the turtle gene expansions were published in *Genome Biology and Evolution* as (Li et al., 2013) and were included in the genome sequence publication in *Genome Biology* (Shaffer et al., 2013).

I also present a gene expansion analysis in the bowhead whale lineage, subsequent to their split with minke whales. Owing to the short divergence time between bowhead and minke whales, no large-scale duplication was identified. Nevertheless, I identified several gene duplications of interest. These results are included in Keane et al. (in preparation).

In the beta-keratin study, Wilfried Haerty contributed several analyses including the manual curation of beta-keratin sequences, the estimation of gene conversion bias, and the PhyML, secondary structure and positive selection analyses. I annotated the painted turtle genome, predicted orthology relationships, discovered the expansion of the beta-keratin gene family, identified their segmental duplication and estimated phylogenetic dates. Wilfried and I wrote the manuscript with notable contributions from Chris Ponting.

In the bowhead whale gene duplication analyses of bowhead whales, David Brawand contributed code and methods to reconcile species and gene trees.

3.1 Evolution of beta-keratin genes in turtles

3.1.1 Background

Previously, our annotation and gene duplication analysis of the painted turtle genome (Shaffer et al., 2013) revealed that 957 genes show expansion in the western painted turtle lineage. 15 of the 27 gene families with four or more genes were annotated as being involved in immune response an additional large expansion. We also predicted a large scale expansion of the beta-keratin gene family in the genome of the painted turtle. This motivated us to fully characterise the expansion of bete-keratin genes because they were previously associated with the hard keratinous material which covers turtle shells.

Many gene family expansions contribute to the emergence of novel, lineage-specific, morphological features. However, few are more striking than those of alpha-keratin genes that have led to the independent appearance of, for example, hair and nails in mammals, wool in sheep, and baleen in whales

(Vandebergh and Bossuyt, 2012). Beta-keratins, on the other hand, are specific to the sauropsids (reptiles and birds) and add much more rigidity to the scales of reptiles than alpha-keratins. They contribute to the formation of the hard keratinous claws and scales of reptiles, as well as to the formation of the beaks and feathers in birds (Alibardi et al., 2009).

Previous genome-wide comparative analyses in chicken and zebra finch identified several clusters of beta-keratin genes, the largest two of which occur on chicken microchromosomes 25 and 27 (Greenwold and Sawyer, 2010). When compared with expressed sequenced tags (ESTs) derived from beak, claws and feathers, the cluster on microchromosome 27 appears to be composed of beta-keratin genes whose proteins exclusively form the feathers. In contrast, the clusters on microchromosome 25 harbor beta-keratins involved in the formation of claws, feathers and scales, which have been postulated to have originated from a single progenitor cluster of beta-keratin genes on the same microchromosome in ancestral birds (Greenwold and Sawyer, 2010).

In turtles, beta-keratins are assembled into filaments in the outer corneous layers of the scutes (Alibardi, 2002; Alibardi et al., 2009) that cover the ventral plastron (12 to 13 scutes) and the dorsal carapace (37 to 38 scutes), the two parts of the shell that are unique to the Chelonians. These scutes form hard structures that are thought to protect turtles from predators (Solomon et al., 1986). The evolution of the turtle shell has long fascinated biologists. However, the paucity of ancestral turtle fossils with intermediate shell forms has made its study difficult. Nevertheless, the formation of the shell has been extensively studied in the context of paleontology (Li et al., 2008; Joyce et al., 2009), comparative anatomy (Nagashima et al., 2009), and the development of the turtle

bone plates (Nagashima et al., 2007; Lyson and Joyce, 2012). In comparison, there have been very few investigations tackling the molecular evolution of the shell. Additionally, these studies have been limited to exploring genes involved in the musculoskeletal development of the shell (Kuraku et al., 2005). The major aim of this study is to gain a better understanding of the evolution of the turtle shell at a molecular level, by studying the evolution of beta-keratins that are components of the turtle scutes. In contrast to the alpha-keratins, for which studies revealed deep conservation across vertebrates (Eckhart et al., 2008), we still lack an understanding of the evolutionary relationships among the beta-keratins found in Reptiles.

While a large number of beta-keratins have been annotated in bird genomes (International Chicken Genome Sequencing Consortium, 2004; Warren et al., 2010), only 17 beta-keratin genes have thus far been identified as being expressed in the skin from the shell, limbs, neck, and tail of the turtle *Pseudemys nelsoni* (Dalla Valle et al., 2009b). A phylogenetic analysis indicated that 16 out of these 17 beta-keratin genes form a *P. nelsoni*-specific clade. The remaining beta-keratin gene, which is expressed specifically in the digits and claws, clustered with chicken sequences derived from scales and keratinocytes. Still, it remains unknown whether this set of beta-keratin genes is comprehensive, whether lineage-specific beta-keratin duplications can be identified among turtle species, and how these genes relate to avian or other reptilian beta-keratin genes.

With the newly sequenced genomes of three turtles (*Chrysemys picta*, *Chelonia mydas*, *Pelodiscus sinensis*; (Shaffer et al., 2013; Wang et al., 2013), that represent approximately 160 million years of chelonian evolution, we have identified a total of 211 turtle beta-keratin genes. Further analysis of these genes

reveals lineage-specific duplications of the beta-keratin gene family in turtles in the syntenic location to the expansions of this family in birds. The timing of the emergence of the turtle-specific beta-keratin clade (173–273 My ago) is predicted to have coincided with the emergence of the turtle shell, 230 to 270 million years ago (Li et al., 2008). Consequently, expansions of beta-keratin genes may have contributed to the evolution of two major morphological innovations, the turtle shell and avian feathers.

3.1.2 Results

Identification and characteristics of beta-keratins in turtles and other reptiles

We identified, through homology searches and manual curation (see Materials and Methods), 92, 40 and 79 beta-keratin genes in the genomes of three turtle species: *Chrysemys picta* (the painted turtle), *Chelonia mydas* (the green sea turtle), and *Pelodiscus sinensis* (the Chinese softshell turtle), respectively. 11, 31 and 43 additional loci containing beta-keratin-like sequences were either truncated or disrupted by a premature stop codon in *C. picta*, *C. mydas* and *P. sinensis*, and thus were discarded from further analyses. To allow comparison, 106 and 133 beta-keratin genes were identified in the genomes of chicken and zebra finch, respectively, using OPTIC orthology predictions (Heger and Ponting, 2007, 2008). However, because of likely differences in genome assembly qualities, it is difficult to draw any conclusion based on differences in the numbers of beta-keratin genes found.

Synteny of non-specific bird and turtle beta-keratins

To identify bird genomic regions in synteny with the beta-keratin clusters of

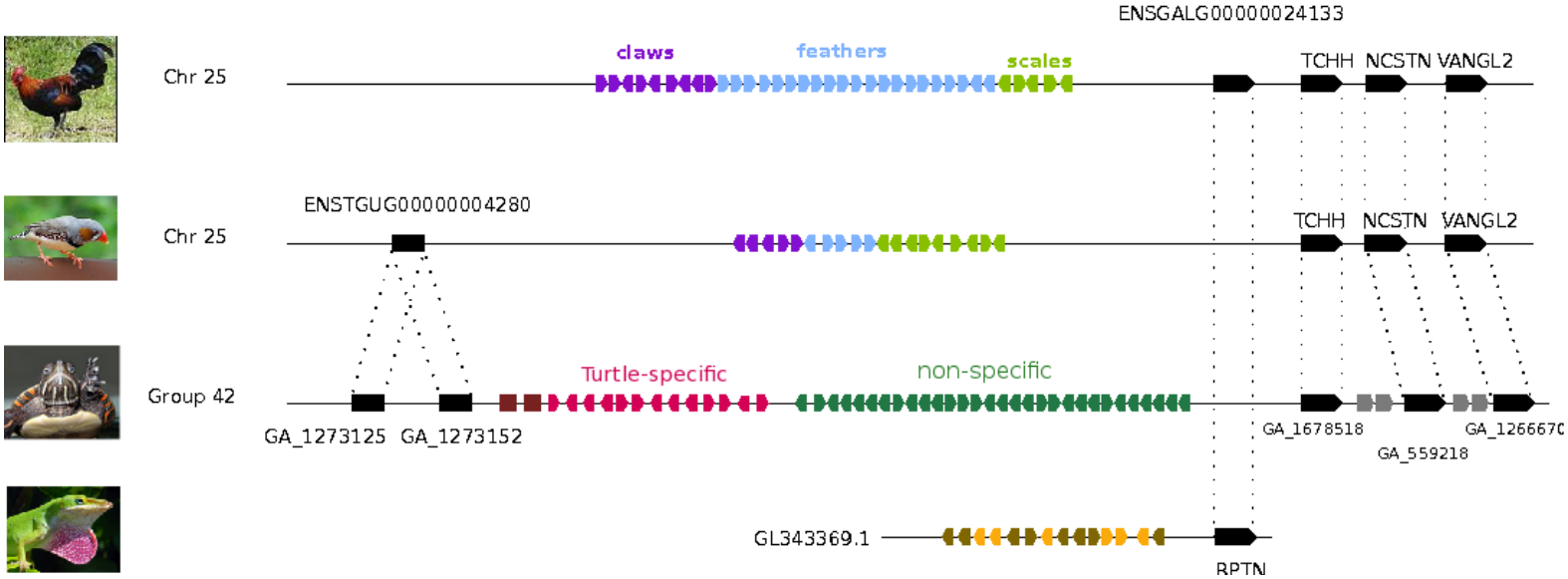

Figure 12: Synteny conservation for the beta-keratin cluster on the bird microchromosome 25 across reptiles. Data from Greenwold and Sawyer (2010) and Dalla Valle et al. (2009) were used to identify beta-keratins involved in the formation of the claws (violet), feathers (blue), and scales (green), and those that are turtle-specific (red) and non-specific, i.e. present in both turtles and birds (dark green). Orthologous genes that are not beta-keratins are depicted in black. Protein coding genes without orthologous relationship are in gray. *A. carolinensis* genes displayed in brown are missing due to poor alignments.

turtles, we performed analyses using 1:1 non-keratin orthologous genes flanking beta-keratin genes. Of 92 beta-keratin genes in the *C. picta* genome, 45 were located on a single scaffold (Group42) that is syntenic to chicken microchromosome 25 (Figure 12). Although additional *C. picta* beta-keratin genes were located on smaller scaffolds, these have undetermined syntenic relationships with the chicken genome. This is because they consist of beta-keratin genes only, or they lack one-to-one orthologs in either chicken or zebra finch which unambiguously map to a single region. Each of the annotated beta-keratin sequences encode a region of 20-30 amino acids, which represents the highly conserved core of beta-keratins from other reptiles (Figure 13). Likewise, we determined that a beta-keratin gene cluster on scaffold GL343369.1 of the green anole lizard (*Anolis carolinensis*, (Alföldi et al., 2011) lies in conserved synteny with chicken microchromosome 25 (Figure 12). Furthermore, we were able to retrieve EST data for all beta-keratin genes on scaffold GL343369.1 but two, and found that they are expressed in the skin of the green anole lizard.

To investigate synteny and orthology relationships among turtle beta-keratin genes, we annotated all beta-keratin genes located in scaffolds containing at least two beta-keratin genes and used phylogenetic reconstruction to determine their orthology relationships (Figure 14). Although scaffolds harboring *C. mydas* and *P. sinensis* beta-keratin genes are shorter due to their more fragmentary genome assemblies, beta-keratin genes with close orthology relationships to Group42 *C. picta* beta-keratin genes tend to be situated on the same scaffolds, suggesting that they are also located in regions with conserved synteny to microchromosome 25 of chicken (Figure 14). Other *C. mydas* and *P. sinensis* beta-keratin genes have undetermined syntenic relationships, suggesting either that they are in a region of the same microchromosome that is hard to assemble (often associ-

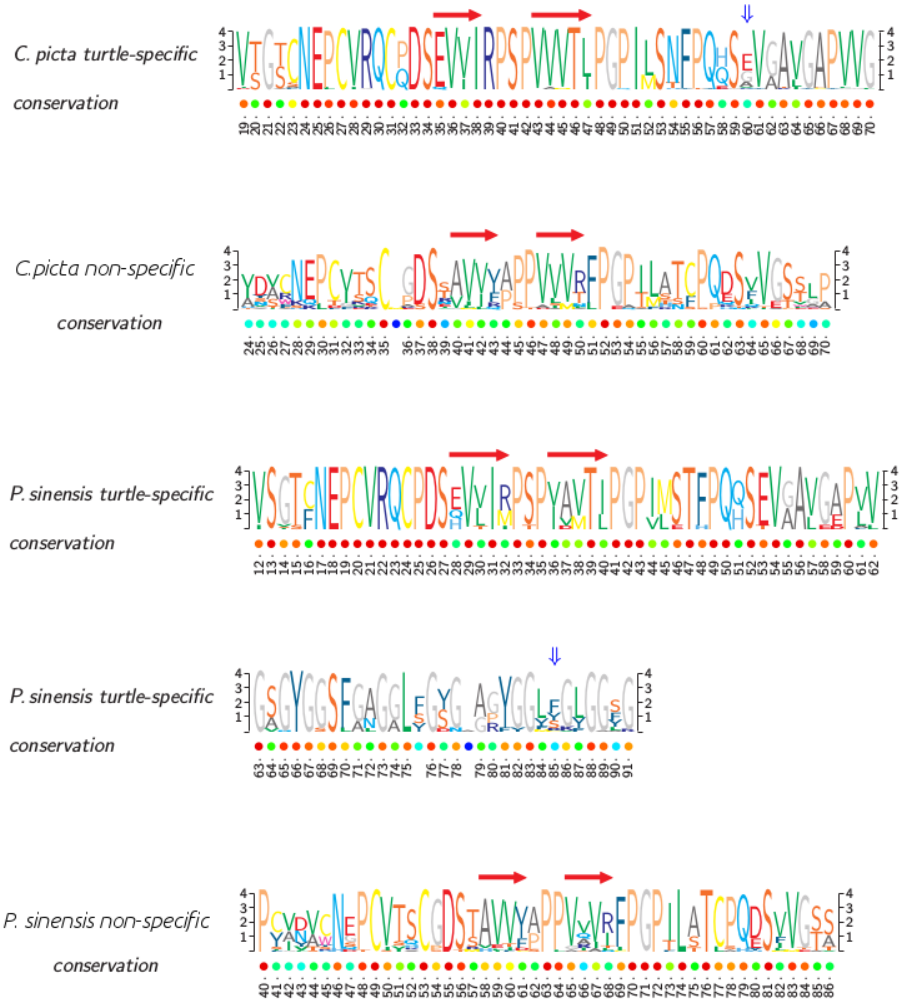

Figure 13: Sequence logo

Sequence logos for the "turtle-specific" and "non-specific" beta-keratins in *C. picta* and *P. sinensis*, and amino-acid sequences downstream of the beta-strands (tail) within "turtle-specific" beta-keratins in *C. picta*. The identification of the shell beta-keratins is based on clustering with cDNA sequences from the precursor cells of the shell in *P. nelsoni* (Dalla Valle et al., 2009b). The red arrows represent the beta-strands identified using PSIPRED (Jones, 1999)). The blue arrows represent the sites with evidence for positive selection using PAML.

ated with repeat-rich regions) or on different chromosomes.

Turtle-specific and avian-specific beta-keratin gene family expansions

Phylogenetic analyses of reptilian beta-keratin genes (Figures 15 and 16) revealed that turtle beta-keratin genes are represented within at least two major clades, one of which is turtle-specific, whilst the other is shared with birds (Figures 15 and 16). Consistent with previous analyses (Dalla Valle et al., 2009b), 16 of the 17 beta-keratins that were obtained in a study of the *P. nelsoni* shell skin and soft epidermis lie within the turtle-specific clade, while the one found to be expressed in the claws was located within the clade shared with birds (Figure 15). It is notable that the turtle-specific beta-keratins are less divergent from one another than are non-specific beta-keratins that clustered with avian beta-keratins (average amino acid identity 81.1% ±1.0% versus 56.2% ±1.7%). This would be consistent with turtle-specific beta-keratin genes tending to have arisen, through duplication, more recently than other such genes. The bird beta-keratin genes in this shared clade were previously associated with the formation of scales based on comparison with expressed sequence tags (Greenwold and Sawyer, 2010). The largest clade in this phylogeny is avian-specific and contains both chicken and zebra finch beta-keratin genes which are known to be expressed in feathers (Greenwold and Sawyer, 2010). As expected, turtle beta-keratins lack a tail sequence specific to previously described feather beta-keratins (Sawyer et al. (2005); Figure 13).

Several studies in Squamates and Sauropsids revealed that beta-keratins tend to differ in their amino acid composition according to the morphology and hardness of the tissue (Alibardi et al., 2007). Using the clustering of the turtle beta-keratins with previously identified beta-keratins from *P. nelsoni* (Dalla Valle

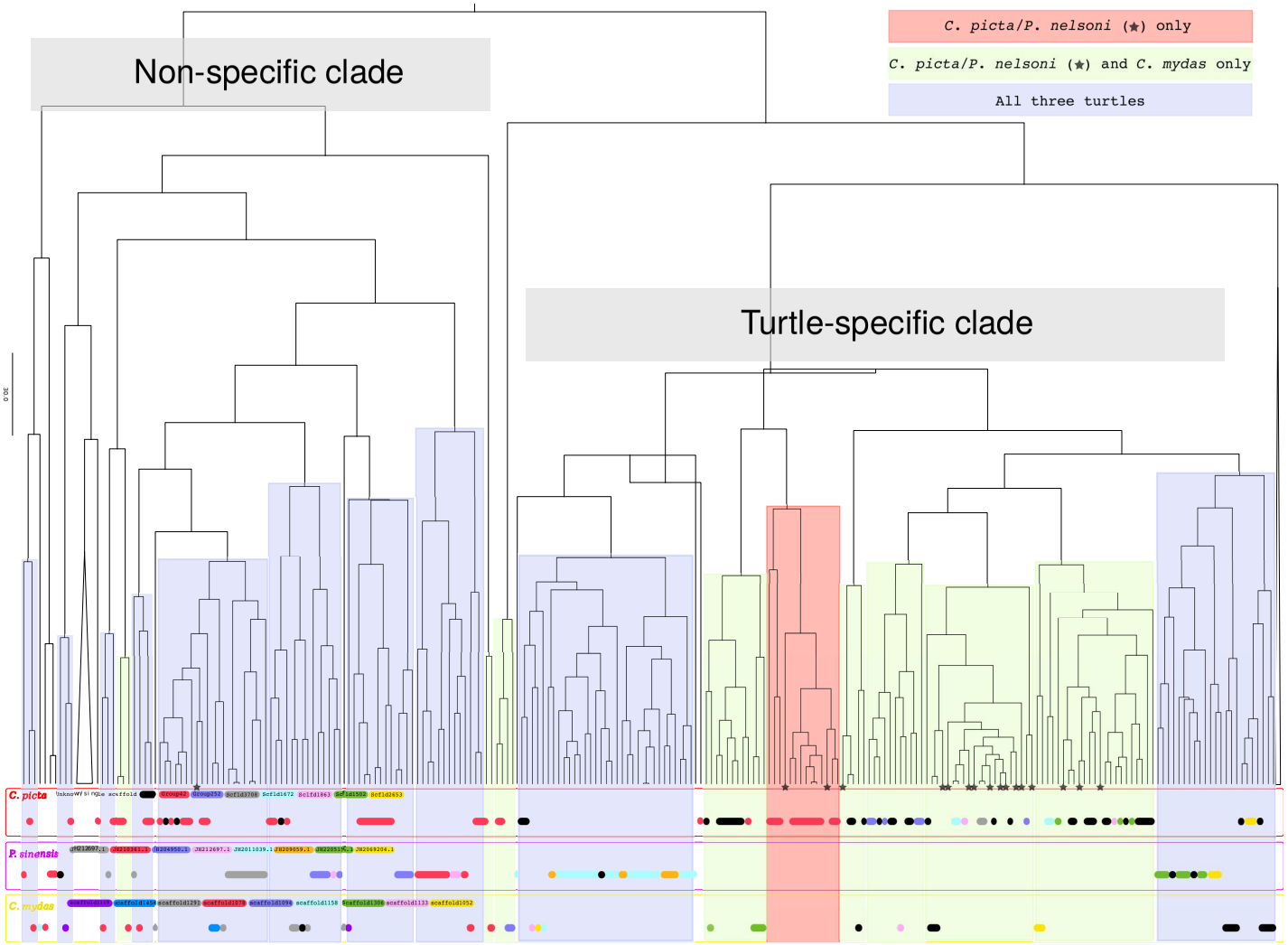

Figure 14: Synteny analysis of beta-keratin genes in the three turtles.

In this analysis, orthology relationships between beta-keratin genes from *C. picta*, *P. sinensis*, and *C. mydas* were determined by identifying clades with beta-keratin genes from all turtles (in blue) and in *C. picta*, and *C. mydas* (in green). Beta-keratin genes lying in the same clade likely share syntenic positions and their scaffold location are provided at the bottom of the figure: colours represent scaffolds. *C. picta* Group42 beta-keratin genes represent almost all beta-keratin genes located in the non-specific clade and include some of the turtle-specific beta-keratin genes, suggesting that Group42 is a hotspot for beta-keratin gene evolution in turtles. Although more fragmentary, the synteny analysis of *P. sinensis* and *C. mydas* beta-keratin genes also supports this model as beta-keratin genes situated in clades consisting of Group42 orthologs tend to be located on similar scaffolds. Colour keys are defined as follows, *C. picta* black: single or unknown scaffold, red: Group42, blue: Group252, grey: Scfld3708, little cyan: Scfld1672, pink: Scfld1863, green: Scfld1502, and yellow: Scfld2653. *P. sinensis* grey: JH21297.1, red: JH210361.1, blue: JH204950.1, pink: JH212697.1, light cyan: JH2011039.1, orange: JH209059.1, green: JH2205150.1, and yellow: JH2069204.1. *C. mydas* purle: scaffold169, light blue: scaffold1456, grey: scaffold1291, red: scaffold1078, blue: scaffold1094, light cyan: scaffold1158, green: scaffold1306, pink: scaffold1133, and yellow: scaffold1052.

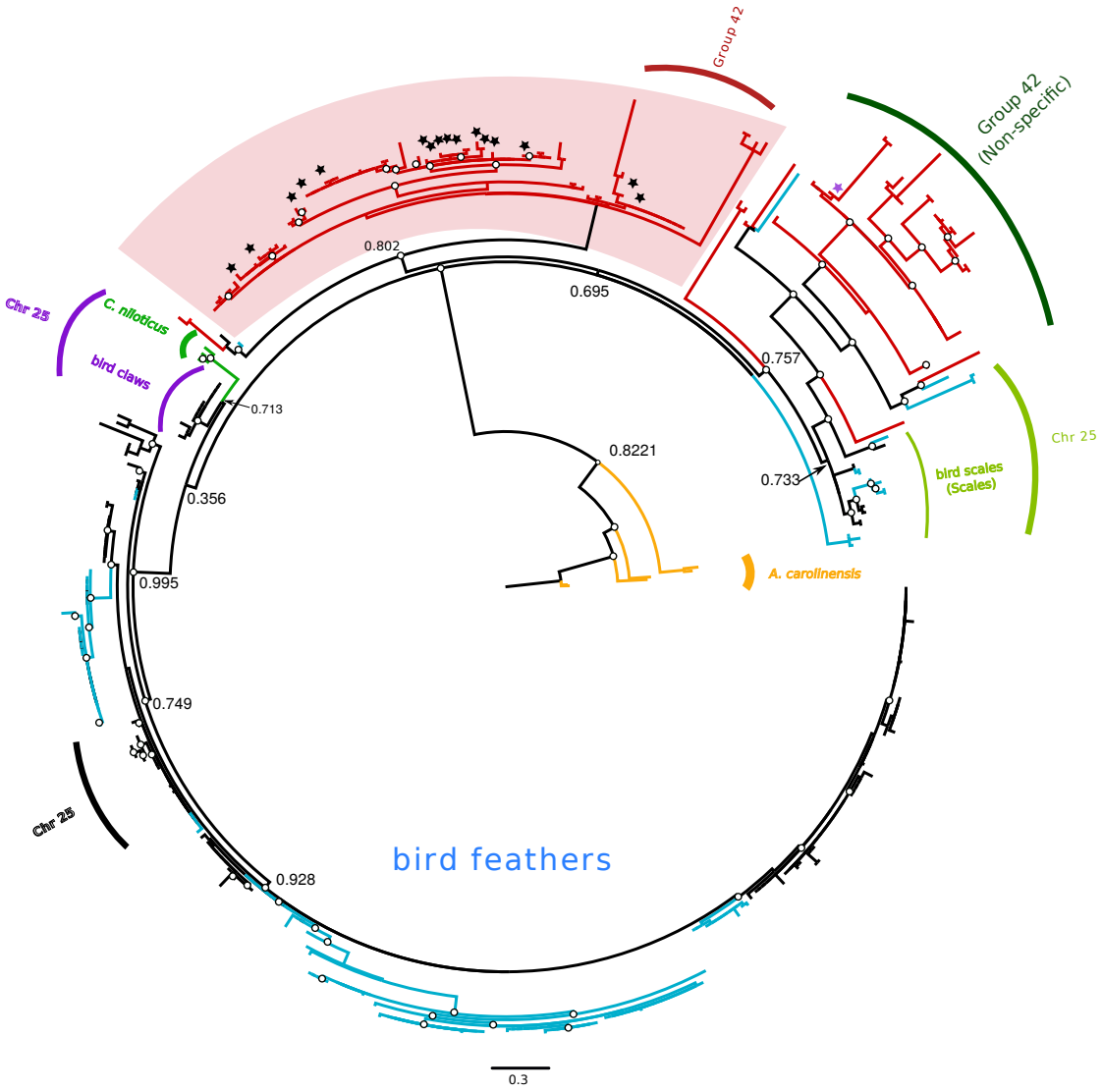

Figure 15: Identification of turtle beta-keratins potentially associated with shell formation.
Phylogenetic tree of the beta-keratins in *C. picta* (red), *Anolis carolinensis* (yellow), *Crocodillus niloticus* (green, 3 proteins), and *Gallus gallus* (black), and *Taeniopygia guttata* (blue). The 16 translated cDNA sequences expressed in the skin from shell, soft skin, claws and digit-tip of *Pseudemys nelsoni* are represented by black stars; the cDNA sequence with tissue-specificity to claws and digit tip is labelled with a purple star. The red shaded area highlights the putative "shell" beta-keratin clade in *C.picta*. Numbers next to internal nodes represent bootstrap confidence values.

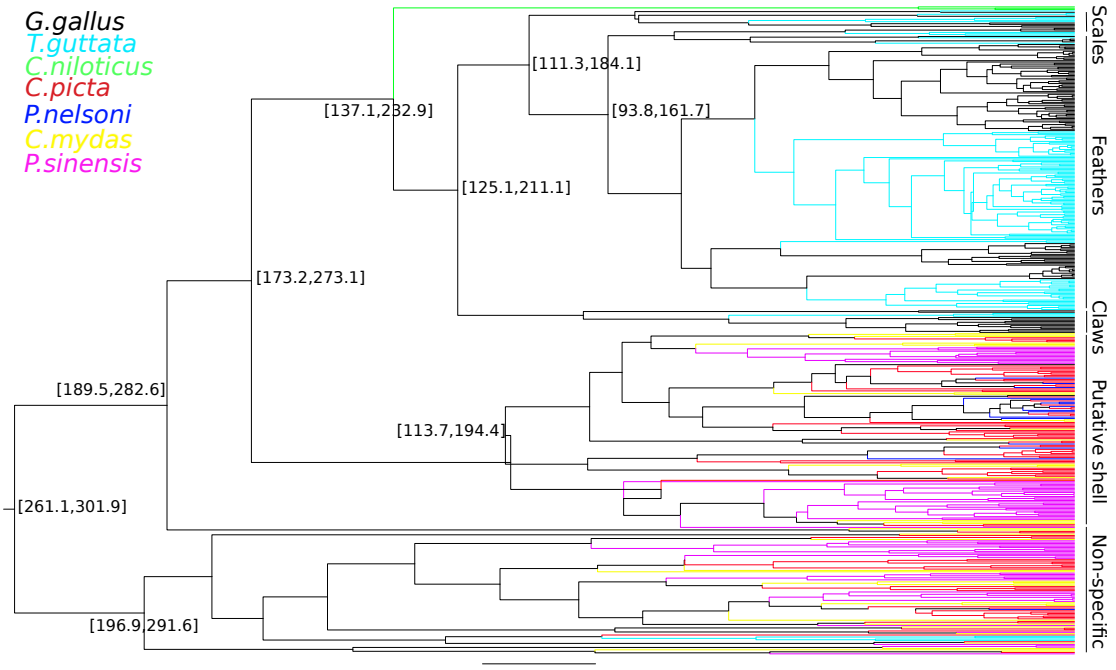

Figure 16: Phylogenetic trees of beta-keratins from all species and dating using BEAST.
The datings (in million years), which correspond to 95% confidence intervals, are in square brackets. The turtle-specific clade with 16 out of 17 beta-keratins from *P. nelsoni* is labelled as a putative shell clade. The bird-specific clade has been labelled in the same way as in Figure 15. The clade annotated as non-specific corresponds to the clade with turtle, zebra finch beta-keratin genes.

et al., 2009b), some of which were found to be highly expressed in the beta-layers of the scutes, we classified the beta-keratins into candidate (turtle-specific) shell and non-specific keratins, and compared the amino acid composition of the two groups. We found that turtle-specific beta-keratins tend to have a glycine- and tyrosine-rich tail, which is also observed in all beta-keratin genes of the Nile crocodile scales sequenced thus far (Dalla Valle et al., 2009a) or non-feather beta keratins in birds (Alibardi et al., 2009). In contrast, non-specific beta-keratins appear to have a cysteine and proline rich tail relative to turtle-specific beta-keratins (Figure 17).

Evidence for positive selection in turtle-specific beta-keratin genes

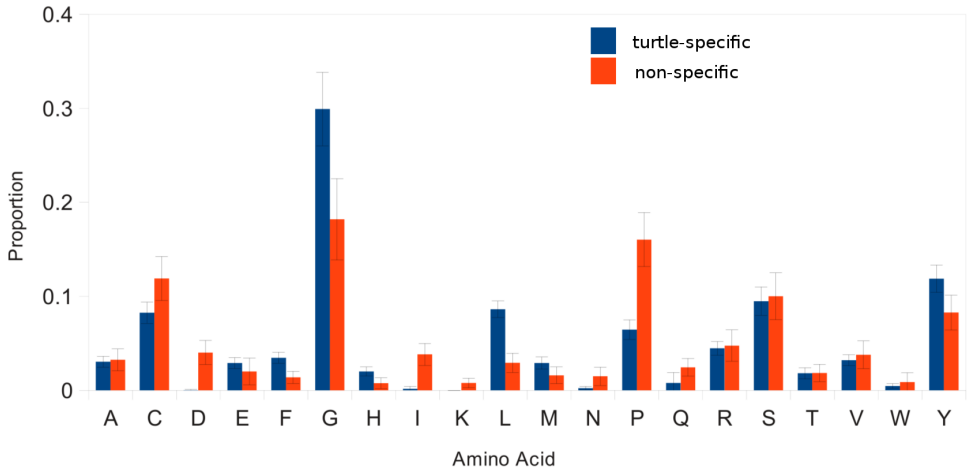

Figure 17: Amino acid composition of beta-keratins
Average composition of beta-keratins in *C. picta* of regions upstream and downstream of the beta-strands. The error bars represent standard error.

We investigated codons of duplicated beta-keratin paralogs for *C. picta*, *C. mydas* and *P. sinensis*, independently, for evidence of positive selection acting on amino acid-changing nucleotide substitutions. We found no evidence for positive selection acting on non-specific beta-keratins in *C. picta* (M1 vs M2, P> 0.05, M7 vs M8, $P>$ 0.05). For both *C. picta* and *P. sinensis* candidate shell beta-keratins, the likelihood ratio tests (M7 vs M8) were significant (P=1.8 $\times$ 10^{-3} and P=7.2 $\times$ 10^{-4} respectively). However, the sites that were identified as being under positive selection differ (Table 1, Figure 13). No evidence for positive selection was found for *C. mydas* beta-keratin genes.

We searched the scaffold containing the largest number of beta-keratin genes (44), Group42, in *C. picta* and more specifically the beta-keratin cluster for an increased GC3 content that could be suggestive of biased gene conversion (Duret and Galtier, 2009). Like microchromosome 25, its syntenic element, Group42 in *C. picta* displays an elevated G+C content relative to the rest of the genome. However, we found a weak but significant increased GC3 content

Table 1: Test for positive selection in beta-keratin genes from *C. picta* and *P. sinensis*.

[a]M0: one ratio model, [b]M1: neutral, [c]M2: positive selection, [d]M3: discrete, [e]M7: beta, [f]M8: beta and ω

Species	Tests	2ΔL	df	P-value
C. picta	M0[a] vs M3[d]	10.08	3	0.0179
	M1[b] vs M2[c]	4.74	2	0.0935
	M7[e] vs M8[f]	12.68	2	0.0018
P. sinensis	M0 vs M3	138.27	3	$< 2.2 \times 10^{-16}$
	M1 vs M2	10.14	2	0.0063
	M7 vs M8	14.48	2	0.0007
C. mydas	M0 vs M3	22.428	3	5.31×10^{-5}
	M1 vs M2	0.06	2	0.97
	M7 vs M8	1.1	2	0.577

within candidate shell relative to non-specific beta-keratin genes (Mann-Whitney test, P=0.043). In addition, we ran GENECONV (Sawyer, 1999) on the beta-keratin coding sequences located on Group42. However, no further evidence for gene conversion was reported.

Turtle-specific beta-keratins in the softshell turtle

Softshell turtles, such as *P. sinensis*, are characterized by the absence of scutes on the carapace and plastron leading to a soft and leathery shell. We therefore compared the beta-keratins of the Chinese softshell turtle *P. sinensis* to those of the hard shell turtles, *C. picta* and *C. mydas*. Using PSIPRED (Jones, 1999), we identified two beta-strands within *P. sinensis*' beta-keratins like those in *C. picta* (Figure 13). We also interrogated a preliminary assembly of the spiny softshell turtle *Apalone spinifera* and found 15 well-conserved beta-keratin genes using BLAST searches (see Materials and Methods). Contrary to previous speculations (Toni et al., 2007), we observed that beta-keratin genes in both softshell turtles possessed the conserved core-box.

Divergence time of the turtle-specific beta-keratin expansion

To estimate the divergence time of the turtle-specific beta-keratins from the other beta-keratin genes, we used BEAST (Drummond and Rambaut, 2007), a Bayesian evolutionary analysis program (see Materials and Methods). The sole molecular dating prior used was the height of the tree, 278.4 million years, which corresponds to the split between turtles and birds (Pereira and Baker, 2006b). Results from BEAST indicated that the turtle-specific beta-keratins diverged from the other beta-keratins between 173 and 273 million years ago (Mya) (95% confidence interval). A further estimate from BEAST indicated that additional beta-keratin duplications have taken place, between 114 to 194 Mya (95% confidence interval), after the divergence of the putative shell beta-keratins from other beta-keratins (Figure 18).

We noted that the topology of the beta-keratin gene tree reconstructed by phyML (Figure 15) was different than that of the tree reconstructed by BEAST (Figure 18). Notably, whereas the most likely tree reconstructed using BEAST suggests that the avian-specific and the turtle-specific clades are sister clades, the tree predicted using phyML suggests that turtle-specific and non-specific beta-keratin clades are sister clades. We observed, however, that the bootstrap values on the phyML tree were generally low (<0.95) near the splits between avian-specific, turtle-specific and non-specific clades. Furthermore, the phyML tree was reconstructed using amino acid alignments, while BEAST used nucleotide alignments for reconstruction. As such, we believe that the tree predicted by BEAST yielded a more accurate representation of beta-keratin evolution.

3.1.3 Discussion

The feathers of birds and the shell of turtles are unique morphological traits that set apart the two groups from the other reptilian species. Several studies

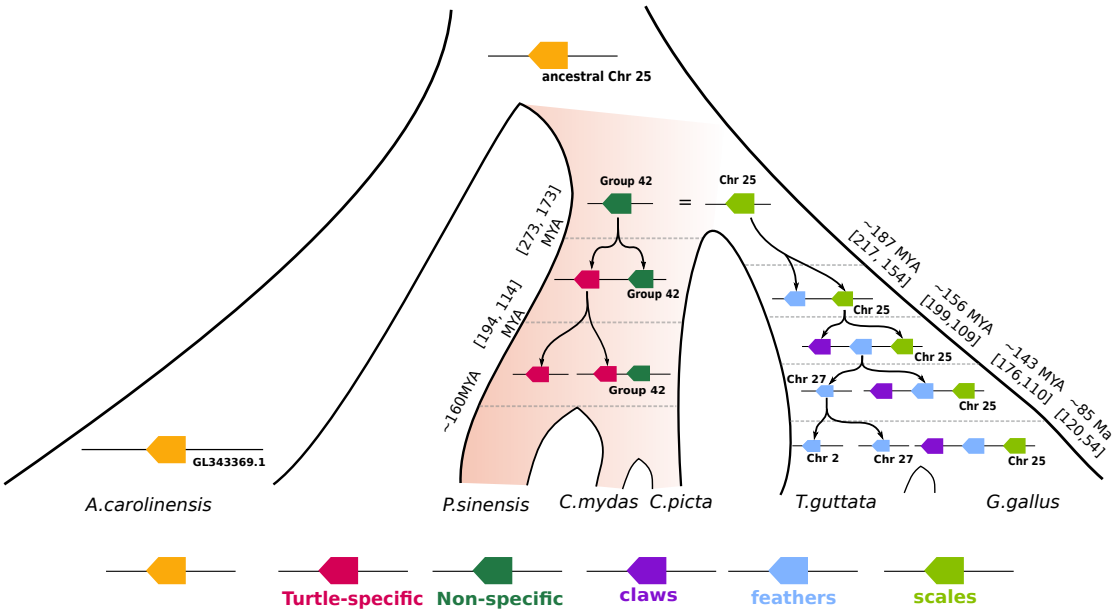

Figure 18: Evolutionary model of the beta-keratin genes in the Sauropsids. Evolutionary scenario for the diversification of the beta-keratin clusters on the common ancestor of the chicken microchromosome 25 and *C. picta* Group42 scaffold. Dates on the turtles lineages were estimated using BEAST (Figure 16), dates for the bird lineage were estimated in Greenwold and Sawyer (2011). "Turtle-specific" and "non-specific" annotations are based on the phylogenetic affinity.

suggest that the feathers originated from modifications of the scales (Greenwold and Sawyer, 2010). For turtles, research has been focused on the development of the shell owing to the extensive transformation of the skeleton and muscles during turtle evolution (Gilbert et al., 2001), while the evolutionary origin of the scutes remains poorly studied. Our study reveals that, concomitant with the formation of the shell, the beta-keratin gene family underwent repeated duplications in the turtle lineage from beta-keratin genes involved in the formation of the claws and scales in the ancestor of the Sauropsids. Previous immunological studies of beta-keratins within the epidermis of various reptilian species suggested a correlation between the type and amount of beta-keratin expressed and epidermis hardness. Consequently, it is possible that duplicated beta-keratin genes were retained in Chelonians because they add rigidity to the shell by increasing the amount of

beta-keratins (Alibardi et al., 2007).

Beta-keratins in the scuteless softshell turtles.

It is believed that the ancestors of softshell turtles possessed a hardshell which subsequently became scaly and soft through the loss of their scutes (Reisz and Head, 2008). According to our beta-keratin annotations, 43 beta-keratin-like genes (out of 122) in the genome of the Chinese soft-shelled turtle, *Pelodiscus sinensis*, were either truncated or disrupted by a premature stop codon compared to only 11 (out of 103) in the painted turtle, *Chrysemys picta*. The higher number of disrupted beta-keratin genes in *P. sinensis* compared to *C. picta* is not expected to be caused by the differences in genome assembly quality since the *P. sinensis* assembly was assembled using both long and short reads, while the assembly of *C. picta* used long reads (the N50 of the *P. sinensis* assembly was also higher: 22.2kb vs 12.2kb). On the other hand, the *C. mydas* genome assembly used short reads only, and comparison is questionable because short-read *de novo* assembly is known to be more problematic in tandemly duplicated regions such as the beta-keratin loci. The elevated number of disrupted beta-keratin genes in the *P. sinensis* genome compared to the *C. picta* genome suggests that the loss of scutes in the softshell turtles was caused by a relaxation of purifying selection on several beta-keratin genes. Retained beta-keratin genes lying in the turtle-specific clade could be more broadly expressed in the turtle epidermis.

An alternative hypothesis to the softshell turtles' loss of scutes was put forward by Toni and colleagues (2007). Their study of the spiny softshell turtle (*Apalone spinifera*) revealed that beta-keratin bundles generally present in turtles were absent in their soft epidermis (Toni et al., 2007). *In situ* analyses based

on immunostaining showed the presence of beta-keratins without a core-box in the epidermis of *A. spinifera*. They suggested that the absence – and possibly loss – of the core-box, which is associated with two beta-strands, in beta-keratin genes of *A. spinifera* could explain the soft epidermis of the softshell turtles. However, we identified numerous beta-keratins that possessed the core's two predicted beta-strands within the genome of the Chinese softshell turtle *P. sinensis* (Figure 13). To further validate the presence of at least some beta-keratins with the core-box in softshell turtles, we searched a preliminary assembly of the *Apalone spinifera* genome for beta-keratin genes. We identified 15 beta-keratins with conserved cores, which confirmed the existence of beta-keratins with a very well conserved core-box in *Apalone spinifera* (data not shown) considering the > 160 My of divergence from *C. picta* and *C. mydas*.

Segmental duplications drive the evolution of the beta-keratin gene family

Due to duplication processes such as non-allelic homologous recombination, members of a single gene family are likely to be clustered together (Lynch, 2007). The gene family consisting of the scale, feather, beak, and claw beta-keratin genes is no exception, and was found to be clustered on both microchromosomes 25 and 27 in birds (Greenwold and Sawyer, 2010). We found that beta-keratin genes in the painted turtle genome and in anole lizard genome also clustered together and shared syntenic position with microchromosome 25 of the birds. Although we can not exclude the possibility that beta-keratin genes were translocated early in reptile evolution, it is likely that the first beta-keratin genes originated on the ancestor of the chicken microchromosome 25. Our phylogenetic analysis revealed that beta-keratin genes with affinity with bird beta-keratin genes are located within regions that

are syntenic with Group42 which shares synteny with microchromosome 25 of birds. This further supports the hypothesis that non-specific beta-keratin genes were located on the ancestral microchromosome 25 which subsequently acted as a hotspot for tandem segmental duplication of turtle beta-keratin genes.

We have determined that, compared to the non-specific beta-keratins, the turtle-specific beta-keratins are more similar to one another. This can be explained by a larger number of more recent turtle-specific beta-keratin gene duplications. In addition, gene conversion between paralogous genes within a genomic region can also lead to reduced estimates of divergence and evolutionary rates (Innan and Kondrashov, 2010). Since mismatch repair during gene conversion tends to increase G+C content in the affected genomic region (Meunier and Duret, 2004), we searched and found a small but significant increase of GC3 within turtle-specific beta-keratin genes relative to non-specific beta-keratin genes. This suggests that biased gene conversion could have been stronger in the shell-related beta-keratin genes of *C. picta*.

Our analysis revealed positive selection to have occurred in the beta-keratin genes of both *C. picta* and *P. sinensis*. Although the sites detected that have evolved under adaptive evolution differ between the two species, they are found in the C-terminal portion of the proteins, outside of the two beta-strands. This result was also reported in the analysis of the green anole genome, where three sites located after the beta-strands were found to have been under positive selection (Alföldi et al., 2011). Beta-keratins have been described to form filaments by interacting in an anti-parallel manner at their beta-strands, and the N- and C- termini have been proposed to interact with other proteins, e.g. through disulfide bonds (Alibardi et al., 2009). Such potential interactions

might explain the observation of selected sites in the C-terminal part of the beta-keratins in multiple Sauropsid lineages.

We further found that beta-keratins within the turtle-specific clade had a glycine- and tyrosine-rich tail, while the beta-keratins in the clade shared with avians had a serine-rich tail (Figure 17). These results fit well with previous findings that turtle shell beta-keratins tended to be glycine-proline-tyrosine-rich (Dalla Valle et al., 2009b).

Turtle-specific beta-keratin genes may have contributed to the formation of the modern shell

It is generally agreed that beta-keratins are important in the evolution of hard skin appendages in reptiles (see (Alibardi et al., 2009) for a review). Our analyses are the first to reveal that independent expansions occurred early (> 160My ago) in the turtle lineage. Additionally, we estimated the duplication of the feather beta-keratins to be 111–184 My old which agrees with previous estimates (Greenwold and Sawyer, 2011). Although it is possible that ancestral beta-keratin-like genes were deleted on both turtle and bird lineages, the presence of only one small clade that contains both turtle and bird non-specific beta-keratin genes from our phylogenetic analyses suggests that the ancestor of turtles and birds likely had few beta-keratin genes. Fewer indeed than the current number of non-specific beta-keratin genes (< 30). In contrast, we found a large turtle-specific clade containing all 16 beta-keratin genes that are expressed in the precursor shell tissue in *P. nelsoni*, suggesting that this clade predominantly contains beta-keratin genes expressed in the shell tissue and whose functions are thus related to the shell (Figures 15 and 16). The only other beta-keratin gene found in *P. nelsoni* was clearly different in terms of sequence and corresponds

to a beta-keratin that shows expression in claws and digit-tips only (Dalla Valle et al., 2009b).

Divergence time of turtle-specific beta-keratin clade coincides with the appearance of the first turtles.

We obtained an approximate estimate of 173-273 My for the divergence time between turtle-specific beta-keratins and the other beta-keratins. Gene conversion can lead to an underestimation of the age of duplication events (Teshima and Innan, 2004; Innan and Kondrashov, 2010). Nevertheless, our results are in line with previous estimates of over 160 My for the divergence time between *P. sinensis* and the common ancestor of *C. picta* and *C. mydas* (Near et al., 2005), and estimates of the appearance of the first turtles, some 230-270 Mya (Li et al., 2008). Therefore, our evidence suggests that the beta-keratin expansion in turtles coincided with the emergence of turtles and the innovation of the turtle shell. Furthermore, we found that the turtle-specific beta-keratin clade was divided into two subclades which diverged some time between 114 and 194 Million years ago (Mya). Again, gene conversion might have led to an underestimation of the divergence age, which means that these two subclades may have diverged earlier. Because *P. sinensis* beta-keratins were found in both subclades, we were able to further narrow the divergence time to at least 160 Mya, which suggests that this duplication happened early in turtle evolution.

Previously, our understanding of turtle shell evolution has been limited by the scarcity of intermediate turtle forms. The lack of intermediate forms has reinforced the de novo model of shell evolution, and some have described the shell to have appeared within a short geological time frame through the differentiation of dermal bones (Gilbert et al., 2001). On the other hand,

the composite model posits that the rigid armoured body of turtles evolved gradually, in multiple steps (Lee, 1996; Cebra-Thomas et al., 2005). The discovery of rapid beta-keratin gene evolution early in turtle history supports the idea that intermediate forms existed for a short period (20My to 90My) after the first Chelionans appeared and rapidly evolved thereafter into the modern turtles. In fact, the recent discovery of *Odontochelys* (Li et al., 2008), the oldest turtle fossils described thus far (220 My), which lack a carapace but possess a fully formed plastron, accompanied by comparative anatomy work between the Chinese soft-shelled turtle and other amniotes (Nagashima et al., 2009), support a two-step scenario for the evolution of the shell. It is thus possible that the beta-keratin divergence early in turtle evolution corresponded to a subfunctionalization event in which duplicated plastron scute beta-keratins acted as a substrate for the origin of the beta-keratins in the scutes of the modern turtle carapace.

By determining the phylogeny of beta-keratin genes in *Chrysemys picta*, *Chelonia mydas* and *Pelodiscus sinensis*, together with beta-keratin sequences from chicken, zebra finch, anole lizard and *P. nelsoni*, we identified monophyletic turtle-specific genes that show evidence of expression in turtle shell skin. Many of these genes lie in a region of these turtles' genomes that is syntenic to microchromosome 25 of the chicken. This primordial cluster of beta-keratin genes is thus the building block that conceivably allowed two independent phenotypic innovations – innovations that differentiate turtles and birds from other reptiles. We provide the sequence of 211 manually curated turtle beta-keratin genes which can be used to study the evolution of the turtle shell. Our data along with previous analyses pertaining to the evolution of the beta-keratins in birds (Greenwold and Sawyer, 2010, 2011), provide us with

a better understanding of the evolutionary trajectory of the beta-keratins in the Sauropsids and how it relates to different morphological features (Figure 18).

This study is the first, to our knowledge, to report a large-scale expansion of beta-keratin genes in turtles, and to propose an association between this expansion and the innovation of the turtle shell. Although further functional studies are needed to determine the role of turtle-specific beta-keratins in the formation of the turtle shell, this study is the first to suggest that large-scale independent expansions of a single gene family contributed to the evolution of two different synapomorphies. Additionally, we also envisage that our characterisation of turtle beta-keratin genes will allow researchers to investigate the role of beta-keratin evolution on turtle scales and claws by comparing smaller turtle-specific clades with their sister clades, e.g. within the non-specific beta-keratins clade.

3.1.4 Materials and Methods

Genome assemblies and identification of an initial set of turtle beta-keratin

The genome assemblies of *Chrysemys picta* (GenBank assembly ID: GCA_000241765.1), *Chelonia mydas* (GenBank assembly ID: XM_007068653.1) and *Pelodiscus sinensis* (GenBank assembly ID: GCA_000230535.1) were obtained as part of a collaborative effort between two turtle consortia (Shaffer et al., 2013; Wang et al., 2013). All remaining genomes and annotations were downloaded from Ensembl (release 66, `http://www.ensembl.org/`. Gene prediction by homology was used to predict genes in the *Chrysemys picta* genome with GPIPE (Heger and Ponting, 2007) using human, anole lizard, chicken and zebra fish. Orthology relationships were then assigned by OPTIC (Heger and

Ponting, 2007, 2008) between *C. picta*, human, mouse, opossum, anole lizard, chicken, zebra finch, zebrafish, and pufferfish. Beta-keratins from chicken and lizard were then identified and used to identify beta-keratins in *C. picta*.

Identifying additional beta-keratins in *C. picta*, *C. mydas* and *P. sinensis*

In order to identify additional beta-keratins, we selected beta-keratins from OPTIC in *C. picta*, anole lizard, chicken, and zebra finch with a core box of 20 residues of a central filament region of beta-keratins, which is highly conserved throughout all reptiles and birds (Alibardi et al., 2009; Greenwold and Sawyer, 2010). These beta-keratins were then mapped onto the *C. picta*, *C. mydas* and *P. sinensis* genomes using both TBLASTN and BLASTN. Since beta-keratins are encoded within a single exon, the hits were then extended on both flanks to find the longest open reading frame. Finally, all proteins were visually inspected and discarded when containing a premature stop codon or undetermined bases.

Identifying the syntenic location of *C. picta* and *A. carolinensis* beta-keratin genes in relation to the birds

In order to identify the syntenic positions in chicken and zebra finch of the *C. picta* beta-keratin genes, one-to-one orthology predictions between *C. picta* and chicken or zebra finch for all non-keratin genes flanking the beta-keratin gene clusters in *C. picta* were used. The same was done for *A. carolinensis*.

To expand our synteny analysis to *P. sinensis* and *C. mydas*, we used a phylogenetic reconstruction of the beta-keratin genes from the three turtles. Scaffolds containing two or more beta-keratin genes were considered for this reconstruction.

Phylogenies of beta-keratins in Reptiles and Estimation of the age of beta-keratin clusters in turtles

Maximum likelihood trees were built using phyML (Guindon et al., 2010) and the JTT model based on the amino acid alignments of beta-keratins from *C. picta*, *C. mydas*, *P. sinensis* and *P. nelsoni* (GenBank ID: AM765814–AM765818, FM163386–FM163397) along with chicken, zebra finch, crocodile (*Crocodylus niloticus*, GenBank ID: AM765851, AM909650, AM765850, (Dalla Valle et al., 2009a), and anole lizard beta-keratins. Using the same alignments, we also reconstructed the phylogeny using MrBayes 3.2 (Ronquist et al., 2012), Jones model). Additionally we used the nucleotide alignments of the beta-keratins in the three turtles and *P. nelsoni* to assess the phylogenetic relationships of these genes in Chelonians. The tree was built under the GTR+I + Γ model implemented in MrBayes.

We also used the protein alignments to guide a nucleotide sequence alignment using tranalign from the EMBOSS package (Rice et al., 2000). The BEAST software (Drummond and Rambaut, 2007) was then used to date beta-keratin duplication events. In order to find the appropriate model of evolution, we used jModelTest (Posada, 2008) and found that the HKY+Γ model best fit the beta-keratin alignment in terms of Akaike Information criterion (AIC) and number of parameters. The uncorrelated lognormal relaxed-clock mode was employed to allow evolutionary rates to vary along branches. We also used priors for the divergence time between birds and turtles (278.4 million years) from Pereira and Baker (2006) and the Yule model as in (Greenwold and Sawyer, 2011). A MCMC run of 10 million episodes was used to determine the tree with the highest likelihood.

Positive selection analysis

Using the guided nucleotide alignment and species tree, positive selection associated with gene duplication was tested using PAML (Yang, 2007) after removal of gaps and ambiguous sites. Five models were tested to identify residues under positive selection: M0 (one ratio), M1a (neutral), M2a (selection), M3 (discrete), M7 (beta), and M8 (beta and ω). We applied a maximum likelihood ratio test to assess the significance of the difference between the models.

3.2 Gene duplications in the bowhead-whale genome

Bowhead whales are marine mammals with extraordinary lifespans; some specimens are thought to have lived over 200 years of age. Despite their larger sizes and number of cells, they do not suffer an increased rate of cancer incidence compared to smaller mammals, an observation known as Peto's paradox (Peto et al., 1975). Bowhead whales are therefore thought to possess cellular mechanisms conducive to long life and cancer resistance. The goal of this project was to identify bowhead whale longevity assurance mechanisms. As discussed previously in the introduction, gene duplication is a major mechanism through which phenotypic innovations can evolve (Holland et al., 1994; Kaessmann, 2010). I surveyed the bowhead whale genome for expanded gene families which may underlie bowhead whale-specific phenotypic traits and adaptations.

3.2.1 Results and Discussion

To predict gene family expansions, I established the orthology relationships of all bowhead whales genes with minke whale, dolphin, cow, platypus, dog, mouse and human genes using OPTIC (Heger and Ponting, 2008), and employed a tree reconciliation method (Materials and Methods). In the bowhead whale lineage, 575 gene families were predicted to have expanded (Figure 19a). We must note here, however, that the number of expanded families in the bowhead whale lineage is not significantly different than that in the minke whale lineage.

Because gene expansion predictions are susceptible to false positives owing to pseudogenes and annotation artefact among other biases, I applied a stringent filter based on percent identity (Materials & Methods) which reduced the number of candidate expansions to 41. The remaining 531 gene families are

predicted to be have duplicated either very recently (high percent identity; >99%) or to have underwent substantial changes (low percent identity; <90%). These were discarded to lower the likelihood that identified duplicates are assembly artefact or are undergoing pseudogenization. Among these and upon manual inspection, I identified several duplicates of interest. For instance, Proliferating Cell Nuclear Antigen (PCNA) is duplicated in bowhead whales with one copy harbouring four lineage-specific residue changes (Figure 19b). Both PCNA copies are expressed in bowhead whale's muscle and kidney (Materials and Methods). By mapping the lineage-specific residues onto the structure of PCNA in complex with FEN-1, I uncovered one amino acid substitution (Q38H) which may affect the interaction between PCNA and FEN-1 (Figure 19c), allowing the duplicated PCNA to acquire a novel function through neofunctionalization. The duplication of PCNA during bowhead-whale evolution is of particular interest due to its involvement in DNA damage repair (Hoege et al., 2002) and association to ageing (Rodriguez-Lopez et al., 2003).

Another notable duplicated gene is Late Endosomal/Lysosomal Adaptor, MAPK And MTOR Activator (LAMTOR1), in which six lineage-specific amino acid changes were identified (Figure 20). LAMTOR1 is involved in amino acid sensing and activation of mTORC1, a gene well studied in the context of ageing (Johnson et al., 2013). Although I could only detect the expression of the original LAMTOR1 copy in both bowhead whale muscle and kidney, this does not preclude the LAMTOR1 copy to possess novel functions in other tissues. Other gene duplications of potential interest include cAMP-Regulated Phosphoprotein (ARPP19), Stomatin-like 2 (STOML2), heat shock factor binding protein 1 (HSBP1), spermine synthase (SMS) and Suppression Of Tumorigenicity 13 (ST13). In this analysis, I identified candidate duplicated genes which have

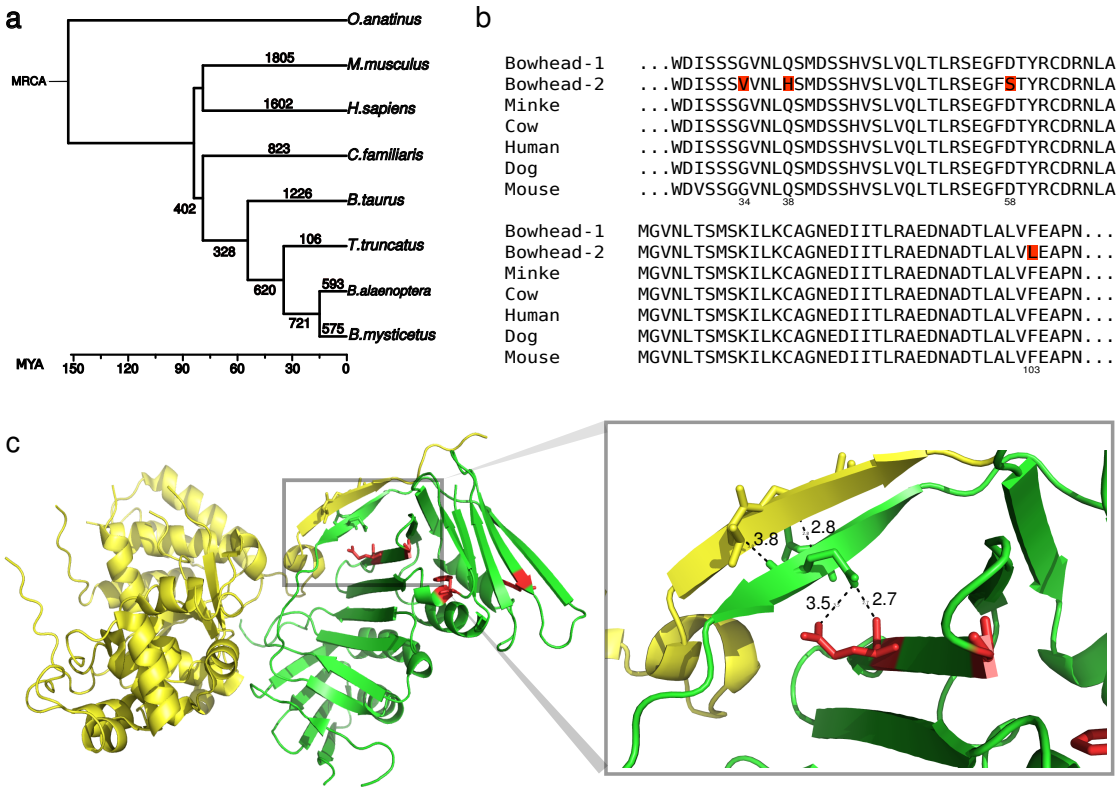

Figure 19: Gene family expansion and PCNA in bowhead whales
(a) Gene family expansion. Numbers in red correspond to the predicted number of gene expansion events during mammalian evolution (b) Multiple sequence alignment of PCNA residues 28-107, showing bowhead whale-specific duplication (gene IDs: bmy_16007 & bmy_21945). Lineage-specific amino acids in the duplicated PCNA of bowhead whales are highlighted in red. (c) Crystal structure of the PCNA (green) and FEN-1 (yellow) complex. Lineage-specific residues on the PCNA structure are coloured in red. A zoom in on the structures reveal a putative interaction between two beta-strands, one within PCNA and another within FEN-1. This interaction may be altered through a second interaction between the PCNA beta-strands and a lineage-specific change from Glutamine to Histidine within PCNA. Distance measurements between pairs of atoms are marked in black. PDB accession: 1UL1.

been retained in the bowhead whale lineage. Although these genes have previously been associated to longevity and may therefore have contributed to the exceptional longevity of bowhead whales, we must point out that our analysis does not support this interpretation. However, a study comparing the number of duplicated genes related to ageing in both minke and bowhead whales may allow us to further investigate this question.

Figure 20: Multiple alignment of LAMTOR1.
Lineage-specific amino acid substitutions are highlighted in red.

3.2.2 Materials & Methods

Orthology prediction

Human, mouse, dog, cow, dolphin and platypus genomes and gene annotations were obtained from Ensembl. The genome and gene annotation of minke whale were obtained from (Yim et al., 2014). In total, 21 069, 22 275, 19 292, 19 988, 15 769, 17 936, 20 496 and 22 733 human, mouse, dog, cow, dolphin, platypus, minke whale and bowhead whale genes respectively, were used to

construct orthology mappings using OPTIC (Heger and Ponting, 2008). Briefly, OPTIC builds phylogenetic trees for gene families by first assigning orthology relationships based on pairwise orthologs computed using PhyOP (Goodstadt and Ponting, 2006). Then, PhyOP is used to cluster genes into orthologous groups and, lastly, gene members are aligned and phylogenetic trees built with TreeBeST (Vilella et al., 2009). Predicted orthology groups can be accessed at `http://genserv.anat.ox.ac.uk/clades/vertebrates_bowhead`.

Gene expansion analysis, filtering and expression

To identify gene families that underwent expansion, gene trees were reconciled with the consensus species tree and duplicated nodes were identified. The following algorithm was used to reconcile gene and species trees:

1. Let S and G denote the set of nodes of the species tree and gene tree. With $g \in G$, define $\sigma(g)$ to be the set of species contained in the subtree that begins at node g. For $s \in S$ define $\sigma(s)$ similarly.

2. Map from G to S: for each $g \in G$, let $M(g)$ be the most recent $s \in S$ for which $\sigma(g) \subseteq \sigma(s)$.

3. For any internal $g \in G$, with child nodes g_1 and g_2, g was inferred to represent a duplication event if and only if $M(g)$ is equal to either $M(g_1)$ or $M(g_2)$.

As stringent filter, gene duplicates in bowhead whales were required to differ by at most 10% in protein sequence from a cognate copy, but were also required to differ by at least 1% to avoid assembly artefact and to remove recently duplicated copies with no function. RNA-seq reads from bowhead whale muscle and kidney were mapped to bowhead whale annotated transcripts using BWA to confirm the

expression of both PCNA copies by observing reads that were uniquely mapped to divergent regions.

4 Chapter 4: Genome evolution in the recently speciated East African cichlids

East African cichlid fishes are celebrated models for recent adaptive radiations both because of their large numbers of species and their diverse phenotypic traits. Over two thousand cichlid species occupy rivers, lakes and swamps in East Africa. Hundreds of distinct cichlid species share the same habitat, and some are thought to have diverged only several tens of thousand years ago (Schluter, 2000; Kocher, 2004).

Cichlids vary significantly in their behavior, body morphology, colouration, and dietary habits. They are therefore excellent models to study phenotypic variation in terms of genomic changes. To better understand the genomic context of speciation, collaborators have sequenced the genomes and transcriptomes of five cichlids: *Oreochromis niloticus* (Nile river), *Neolamprologus brichardi* (Lake Tanganyika), *Astatotilapia burtoni* (swamps near lake Victoria), *Metriaclima zebra* (Lake Malawi), and *Pundamilia nyererei* (Lake Victoria). These five lineages diverged from each other primarily through geographical isolation, subsequent to the formation of the great East African lakes (Figure 21; Kocher (2004)). My aim was to study genome evolution in these cichlids and to identify genetic changes that may have played a role in phenotypic evolution.

The following section presents my work in collaboration with the East African cichlids genome consortium. I study how adaptive and non-adaptive processes shape the genomes and transcriptomes of the closely-related cichlids. First, I describe some technical aspects of the genome project. I then describe my characterisation of incomplete lineage sorting in the genomes of the three

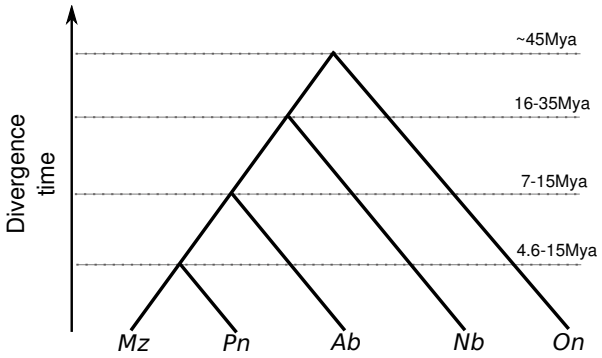

Figure 21: Divergence time between the five sequenced cichlids
Times were derived from Figure 6a of (Genner et al., 2007) and personal communication with Professor Ole Seehausen. *Mz*: *Metriaclima zebra*, *Pn*: *Pundamilia nyererei*, *Ab*: *Astatotilapia burtoni*, *Nb*: *Neolamprologus brichardi*, and *On*: *Oreochromis niloticus*.

most closely-related cichlids. This yielded insights into the population history of the cichlids and how non-adaptive processes impact cichlid genomes. Next, I studied the evolution of selected genes to complement a genome-wide screen for positive selection. Lastly, I studied gene architecture evolution in the context of novel exonic regions and alternative splicing. The first and second parts of this section are included in the cichlids genome paper on which I am a co-first author (Brawand et al., under review); my study on gene architecture evolution is unpublished.

All analyses presented here are my own, except those described in "Samples and Data processing". David Brawand annotated the five cichlid genomes (with my contribution). David also performed the genome-wide positive selection analysis and computed gene expression levels using the generated RNA-seq data. Luis Sanchez-Pulido helped with structural analyses of EDNRB1 substitutions.

4.1 Samples and Data processing

Sampled cichlids and tissues

The cichlids genome consortium sequenced DNA and RNA from five East African cichlids which are described below:

- *Oreochromis niloticus*. DNA-sequencing: High molecular weight (HMW) DNA was extracted from a female individual from an inbred clonal line (Sarder et al., 1999). RNA-sequencing: 11 tissues (blood, brain, eye, embryo, heart, kidney, liver, muscle, ovary, skin and testis) were isolated from several individuals from a Swansea stock.

- *Metriaclima zebra*. DNA-sequencing: HMW DNA was extracted from a single female M. zebra (wild caught). RNA-sequencing: 11 tissues (blood, brain, eye, embryo, heart, kidney, liver, muscle, ovary, skin and testis) were isolated from several wild individuals.

- *Astatotilapia burtoni*. DNA-sequencing: HMW DNA was extracted from a single female individual inbred for ∼60 generations. RNA-sequencing: 11 tissues (blood, brain, eye, embryo, heart, kidney, liver, muscle, ovary, skin and testis) were isolated from several individuals inbred for ∼60 generations.

- *Pundamilia nyererei*. DNA-sequencing: HMW DNA was extracted from a single male partially inbred for ∼5 generations. RNA-sequencing: 9 tissues (brain, eye, gills, heart, kidney, muscle, ovary, skin and testis) were isolated from three individuals inbred for ∼5 generations.

- *Neolamprologus brichardi*. DNA-sequencing: HMW DNA was extracted from a single female inbred for ∼10 generations. RNA-sequencing: 8 tissues (blood, brain, eye, heart, kidney, muscle, skin and testis) were isolated from several individuals inbred for ∼10 generations.

Genome assembly

All five cichlid assemblies were assembled from libraries with 4 different insert lengths prepared from HMW DNA samples: (1) 180 bp paired end fragment libraries (45X coverage), (2) 3 kb jumping libraries (45X coverage), (3) 6-14 kb jumping libraries (2X coveage) and (4) 40 kb fosmid libraries (1X) (Williams et al., 2012). These libraries were sequenced using Hi-Seq Illumina machines and sequence reads were assembled using ALLPATHS-LG (Gnerre et al., 2011).

Table 2: Genome assembly statistics. Adapted from (Brawand et al., under review).

	O. niloticus	*N. brichardi*	*P. nyererei*	*A. burtoni*	*M. zebra*
Est. Genome Size (Gb)	1.0	0.98	0.99	0.92	0.95
Contig N50 (kb)	29.3	13.2	22.6	21.9	20.0
Scaffold N50 (Mb)	2.8	4.4	2.5	1.2	3.7
Sequence coverage (X)	269	171	126	131	128
Heterozygosity (X)	$\frac{1}{4365}$	$\frac{1}{365}$	$\frac{1}{729}$	$\frac{1}{976}$	$\frac{1}{1029}$
Protein-coding genes	24,559	20,119	20,611	23,436	21,673

The resulting assemblies are of high quality (Contig N50 >20Kb, Table 2), except for the *N. brichardi* assembly (N50 of 13.2kb) for which the level of heterozygosity is twice the other cichlid assemblies. This is surprising given that the sequenced individual was inbred for ∼10 generations, whilst the wild caught *M. zebra* genome showed a level of heterozygosity that is nearly three times lower. However, this could be the consequence of a large difference in genetic diversity between the two cichlid species. The ancestors of *M. zebra* and *P. nyererei* may have undergone severe bottlenecks (Founder effect; (Mayr, 1942)) during the colonization of Lake Malawi and Lake Victoria, respectively (∼5Mya) (Genner et al., 2007). This would have caused a sharp decrease in genetic variation in these founding populations. In comparison, the *N. brichardi* lineage is predicted to have colonized Lake Tanganyika over 16Mya (Genner et al., 2007). Although all three populations are predicted to have undergone reduced genetic variation, the *N. brichardi* lineage colonized Lake Tanganyika

earlier, which allowed genetic variation to accumulate and increase to a higher level than in the other two cichlids' lineages.

The protein-coding gene annotations are expected to be of high quality (Table 2). Briefly, to predict protein coding genes, David merged four independent annotation approaches that include *ab initio* gene predictions with Augustus (Stanke et al., 2006), prediction by homology by GPIPE (Heger and Ponting, 2007), projections from annotated protein coding genes from Ensembl (for *O. niloticus* only), and gene models constructed from mapping assembled RNA-seq data using Trinity (Grabherr et al., 2011) and PASA (Haas et al., 2008). A whole-genome alignment was constructed from the five cichlid assemblies (see Methods section from next section), which allowed me to project features, e.g. genes, from one genome onto another. While comparing gene architecture across cichlids (see last section of this chapter), I observed that many exons annotated in one genome could be projected onto genomic regions of another genome that appear to be unannotated. To quantify the number of exons that our annotation approach missed, I projected all annotated exons onto the *O. niloticus* genome, and then projected these exons back onto all four remaining assemblies. I found that a large number of exons were previously missed, even using our exhaustive approach that combined four annotation sources. In particular, our previous annotation comprised of 189,039 exons in *N. brichardi* (Table 3), while 259,005 exons could be identified by projection. We therefore updated our gene models and required projected exons to maintain coding frame and to possess canonical splice sites when applicable.

Next, I wished to verify that information from the whole-genome alignment could recapitulate the known phylogenetic relationships between the five cichlids.

Table 3: Evaluation of gene prediction from annotated exon.

	# of Exons pre-projection	# Exons post-projection
A. burtoni	221,424	267,924
M. zebra	220,518	267,603
N. brichardi	189,039	259,005
O. niloticus	235,460	282,110
P. nyererei	209,317	264,675

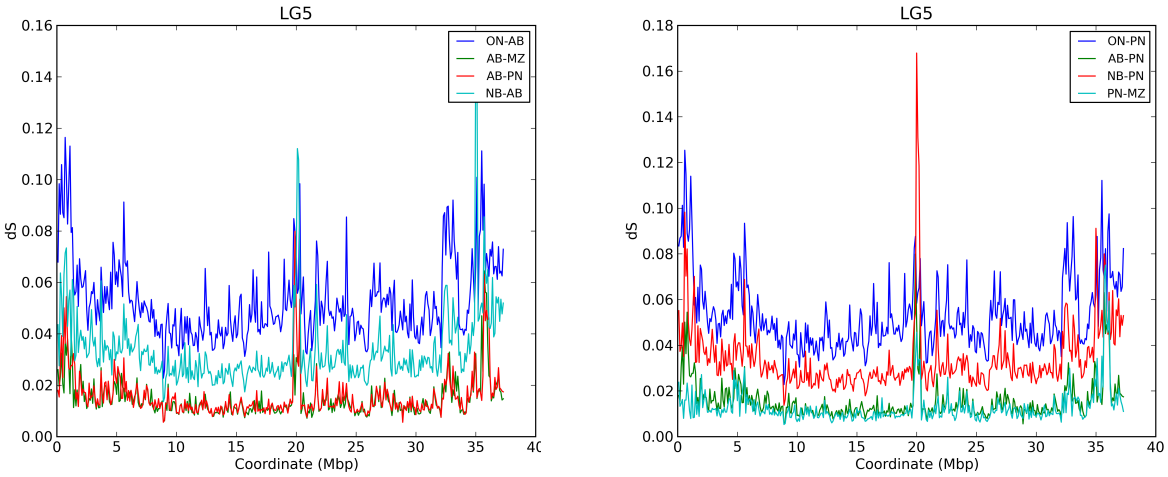

Figure 22: dS over genome alignment.
Pairwise distance (in this case dS is the number of substitutions per site) *Ab* vs other cichlids (left) and *Pn* vs other cichlids (right) on a linkage group (LG5). Right: the divergences *Ab-Mz* and *Ab-Pn* are nearly identical since *Mz* and *Pn* diverged more recently. Left: the divergence *Pn-Mz* is slightly, but consistently, lower than *Ab-Pn*, which supports a short period between successive speciation.

To do this, I computed the pairwise substitution rates between 200kb aligned sequences of all cichlids pairs, and anchored all regions using the *O niloticus* scaffolds. I observed that the substitution rates indeed support the known cichlid phylogeny (Figure 21 and 22). Furthermore, the short evolutionary period between the *Mz-Pn* split and *Mz-Pn-Ab* split is reflected in small differences in substitution rates between the *Pn-Mz* and *Ab-Pn* pairs (Figure 22).

4.2 Incomplete lineage sorting in East African cichlids

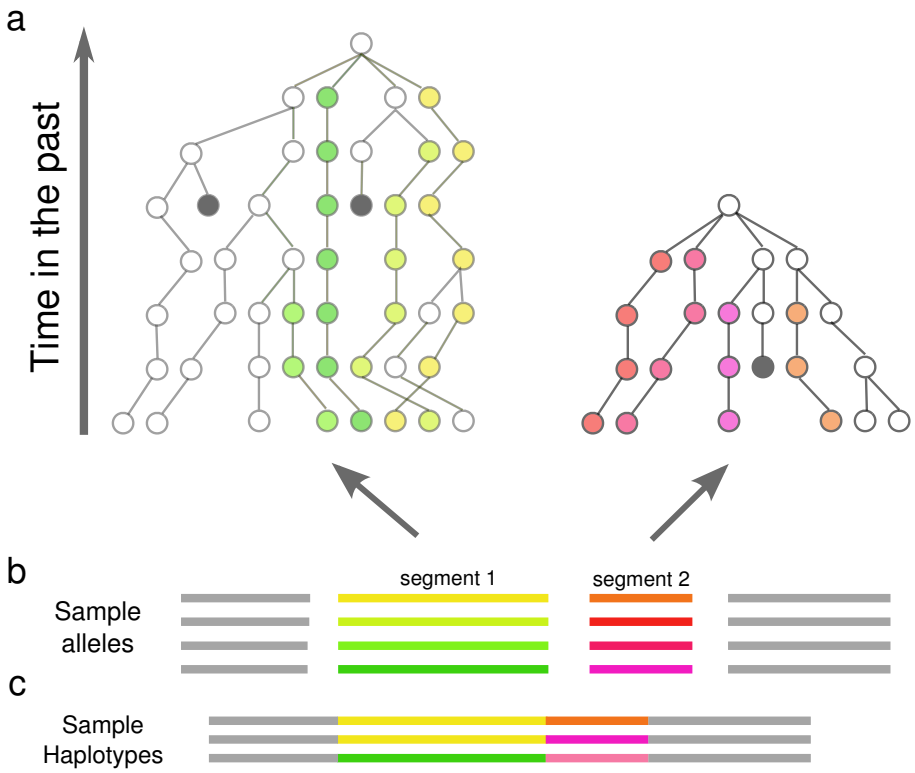

Figure 23: Coalescence of alleles
(a) Different alleles of representative genome segments (e.g. segment 1 and 2) in a contemporary population (leaves) can be traced back to a most recent common ancestor. Each colour represent one allele (apart from white which represent all other alleles). Black alleles are lost from the population. Splits represent the introduction of a new allele by mutation or recombination. (b) Segment 1 and 2 have different versions (alleles) in the contemporary population. (c) Examples of possible haplotypes of contemporary individuals.

The genome is a mosaic of discrete segments, each with its own evolutionary history. This is owing to sexual recombination, which randomly generates the offspring's genome based on the genomes of its parents. All alleles within individuals of the same population can be traced back to a most recent common ancestor (MRCA). However, due to the stochasticity of recombination, the age of the MRCA can vary widely (Figure 23). When the MRCA of alleles precedes a speciation event, an allele from one species may be more closely-related to alleles from another species than another allele from the same species (Note however

that this can also happen when the MRCA is followed by introgression events, e.g. see Neanderthal introgression Sankararaman et al. (2014)). Analogously, MRCAs of fixed alleles from two closely-related species (A and B) can predate a speciation event between their common ancestor (of A and B) and another species (Figure 24). In this case, there is a probability (theoretically $\frac{2}{3}$) that the alleles in species A or B coalesce with the alleles of species C first. For some genomic regions, the segment in species A (or B) will therefore be more closely related (and appear more similar) to the orthologous segment in species C than to the orthologous segment in species B (or A), despite species A and B being more closely related to each other than they are to species C. This phenomenon is called incomplete lineage sorting (ILS) and has been studied mostly in primates.

In 1% of the genome, orangutan sequences are evolutionarily closer to the sequences of human or chimpanzee than the latter are to each other (Hobolth et al., 2007). In the orangutan-human-chimpanzee trio, Scally et al. (2012) estimated that 30% of the genome was incompletely sorted. Because the amount of ILS and the size of incompletely sorted segments depend on several population parameters (including population size, population history, recombination rates and speciation times), studying ILS can provide insights into the evolutionary history of closely-related species (Hobolth et al., 2007; Dutheil et al., 2009). For instance, the speciation event of human-chimpanzee and human-chimpanzee-gorilla were estimated to have occured 3.7 and 5.95 million years ago (Scally et al., 2012). Furthermore, the finding of a 1% ILS in the genomes of the orangutan-human-chimpanzee trio contributed to the belief that the ancestors of human and chimpanzees never experienced a severe population bottleneck (Hobolth et al., 2007). Indeed, population genetic theory

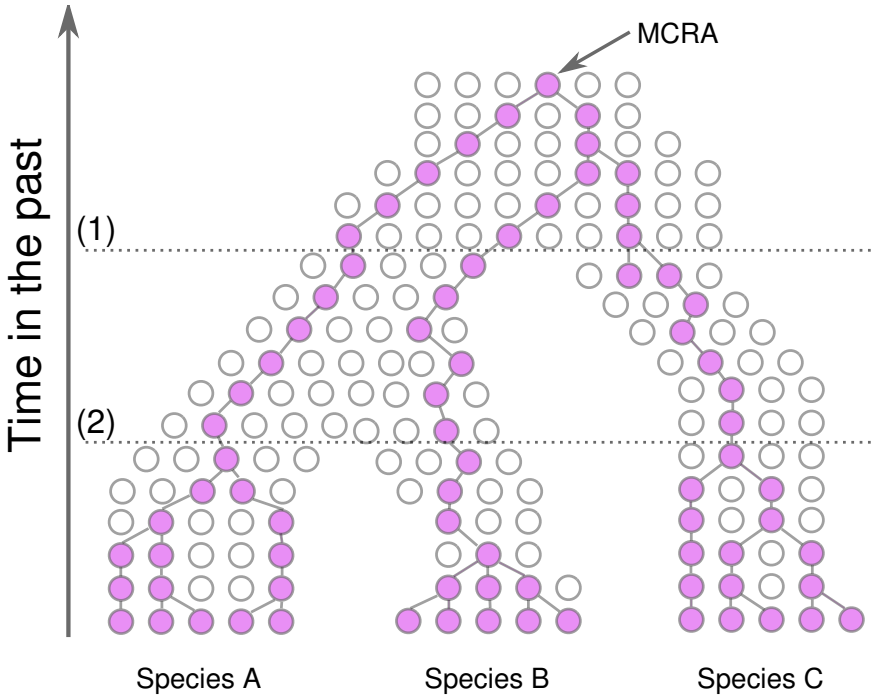

Figure 24: Example of incomplete lineage sorting.
Alleles derived from an ancestral allele (most recent common ancestor: MRCA) in the common ancestor of species A, B, and C are coloured in purple. These alleles were subsequently fixed in the population of all three species (while other alleles, coloured in white, disappeared). The alleles from species B are therefore more closely related to the alleles from species C than to alleles from species A, despite species A and B being more closely related. (1) Speciation time between the common ancestor of species A and B, and species C, (2) speciation time between species A and species B.

predicts that a severe population bottleneck would dramatically reduce the genetic diversity in the human-chimp ancestors which is inconsistent with a predicted ILS level as high as 1%. I therefore wanted to study the evolutionary history of cichlids by analysing ILS within their genomes.

Owing to their relatively recent divergence time (Joyce et al., 2011; Loh et al., 2013), we predicted that incomplete lineage sorting (ILS) may have played a major role in shaping the genomes of haplochromine cichlids (i.e. *M. zebra*, *A. burtoni* and *P nyererei*). Previous studies estimated that *A. burtoni* (*Ab*) diverged from the common ancestor of *M. zebra* (*Mz*) and *P. nyererei* (*Pn*)

7–15 million years ago and that the three lineages split in short succession. Consistent with this, I found that nearly half (43%) of the three haplochromine genomes sequenced are incompletely sorted (Figure 25). Furthermore, assuming a constant mutation rate, and a *Ab-Mz-Pn* speciation event 10My ago (based on (Genner et al., 2007)), I predicted the subsequent speciation event between *Mz* and *Pn* to be ∼8.5My ago (Table 4; see Methods). Therefore, only ∼1.5My is predicted to have separated the *Ab-Mz-Pn* and *Mz-Pn* splits. A large amount of genetic variation in the *Ab-Mz-Pn* common ancestor likely survived in the ancestors of *Pn* and *Mz* during this colonization period that lasted 1.5My. Following the *Mz-Pn* divergence ∼8.5My ago, different alleles became fixed in the two lineages, resulting in a genetic landscape with high levels of ILS.

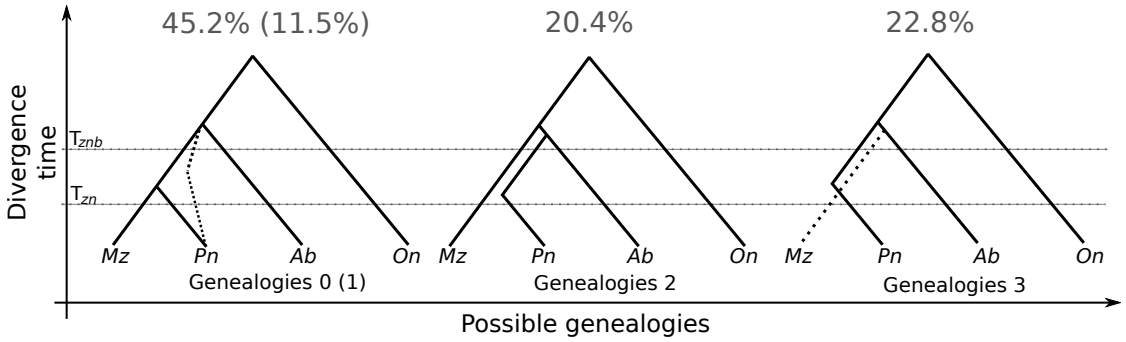

Figure 25: Representation of possible coalescence times and trees predicted by coalHMM.
In genealogy 0, the *M. zebra-P. nyererei* divergence falls between the two speciation times, T_{zn} and T_{znb}. In genealogies 1 (dashed line), 2 and 3, all coalescence events are ancient and occur before T_{znb}.

Consistent with studies in primates (Scally et al., 2012), the degree of ILS is highly variable across chromosomal location (Figure 26). Additionally, coding regions were found to be slightly, yet significantly, depleted in ILS compared with intergenic regions (43.5% vs 41.0%, $P < 0.001$). This depletion is likely due to background selection: the amount of genetic variation in regions under background selection (or positive selection) is lower than in regions evolving

neutrally. This means that there are fewer alleles to be incompletely sorted in these regions. In their analysis of the gorilla genome, Scally et al. (2012) found that the proportion of ILS within coding regions was reduced to $\sim$ 23% compared to $\sim$30% at the genome-wide level. The smaller reduction in the extent of ILS within coding and intergenic regions (2.5% vs $\sim$7%) may reflect weaker purifying selection in cichlids than in mammals (and weaker background selection).

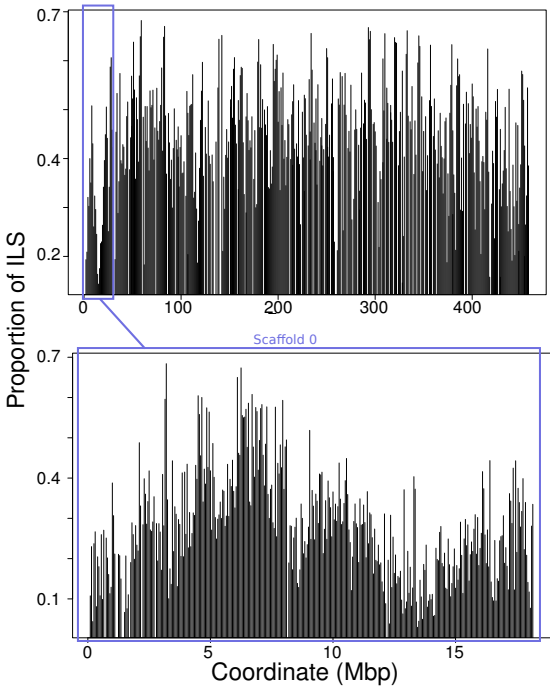

Figure 26: Levels of incomplete lineage sorting varies across the genome. ILS across scaffold 0–100 of *M. zebra* with 1Mb windows and zoom in across scaffold 0 of *M. zebra* with 50kb windows

4.2.1 Methods

Estimating incomplete lineage sorting

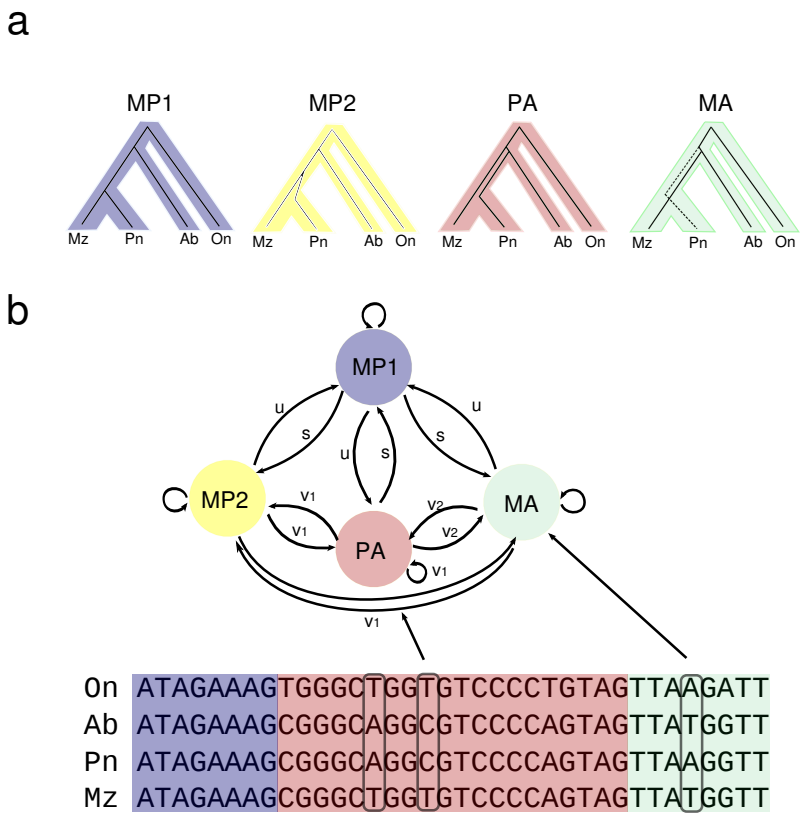

Figure 27: Illustration of CoalHMM
(a) The different coalescence history represented by CoalHMM. (b) A hidden markov model represent the coalescence history and estimates transition probabilities (u, s, v_1, v_2) from data (Adapted from Hobolth et al. (2007)).

I constructed a 5-way whole genome alignment among *O. niloticus*, *N. brichardi*, *A. burtoni*, *M. zebra* and *P. nyererei*, centred around *M. zebra* using MULTIZ (Blanchette et al., 2004). After re-aligning alignment blocks from MULTIZ with muscle (Edgar, 2004), low-quality regions of the alignment were filtered out by requiring the following criteria to be met:

- Genomic segments from the haplochromime species (*A. burtoni*, *M. zebra* and *P. nyererei*) must be present, as well as the genomic segments from the outgroup, *O. niloticus*.

- Contiguous aligned blocks must be at least 50 nucleotides long.
- 10-nt sliding windows must be aligned with a maximum number of variation and gaps.

Contiguous alignment blocks were then concatenated into longer segments if they were separated by fewer than 30nt within the *M. zebra* reference genome. To estimate incomplete lineage sorting, coalHMM (Dutheil et al., 2009) was used in the same way as (Scally et al., 2012) and (Prufer et al., 2012): scaffolds were divided into chunks of 1Mbp alignment blocks based on the *M. zebra* reference genome and coalHMM estimated coalescence parameters on each of these blocks separately (Figure 27). Only scaffolds of length greater than 500kb were assessed for ILS, which together make up 785Mb out of the 849Mb of the *M. zebra* assembly. Overall, 418Mb (out of 785Mb) of aligned regions passed the filtering steps and had an assigned genealogy. Out of these, 189Mb (45.2%), 48Mb (11.5%), 85Mb (20.4%) and 95Mb (22.8%) were assigned to genealogies 0, 1, 2 and 3, respectively (Figure 25). Posterior decoding was used to infer the genealogy with highest likelihood for all considered alignment blocks (Scally et al., 2012). These blocks were finally used to estimate the proportion of ILS within coding and intergenic regions using genomic association tester (GAT; Heger et al. (2013)). Gene annotations produced by the consortium was used to define coding regions and used as tracks for GAT. Briefly, GAT tests whether two sets of genomic intervals are associated more than expected by chance from simulations. Using a thousand simulations, intergenic regions and coding regions were found to overlap ILS regions in 43.5% (0.4% enrichment compared to neutral expectations, $P < 0.001$) and 41.0% (7.2% depletion, $P < 0.001$) of their total lengths, respectively.

Estimating divergence times in the haplochromine cichlids

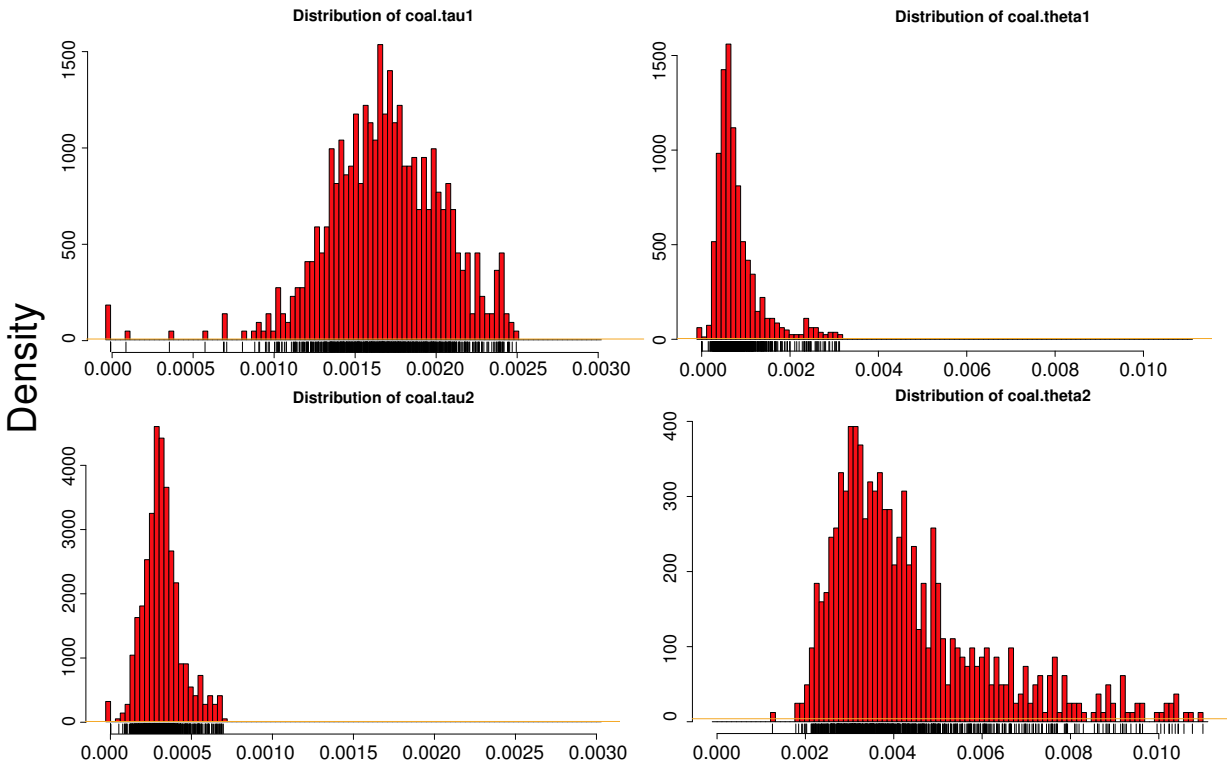

Figure 28: Distribution of parameters estimated by coalHMM
Parameters were estimated on 1Mb windows of alignments centred on the *M. zebra* reference genome. See Table 4 for the description of parameters.

CoalHMM also estimated (unscaled) population parameters for each of the 1Mb blocks (Figure 28), and these can be found in Table 4. The parameters can be used to estimate divergence time and population size when true population parameters such as mutation rates are known (Scally et al., 2012). Although these parameters are unknown in cichlids, it is possible to use the unscaled divergence τ to estimate differences in speciation time, T, because their relationship is linear (Scally et al., 2012):

$$T = \frac{\tau}{2\mu} \tag{1}$$

And therefore, if a constant mutation rate μ is assumed,

$$T_{zn} = T_{znb} \times \frac{\tau_{zn}}{\tau_{znb}} \tag{2}$$

Table 4: CoalHMM: Initial and estimated parameters

Parameters	Description	Initial	Estimated mean (median)
τ_1	Unscaled div. *Mz-Pn*	0.004	$1.74 \times 10^{-3}(1.71 \times 10^{-3})$
τ_2	Unscaled div. *Mz-Pn* and *Ab*	0.0015	$3.31 \times 10^{-4}(3.10 \times 10^{-4})$
θ_1	Unscaled pop. size *Mz-Pn*	0.002	$8.84 \times 10^{-4}(6.72 \times 10^{-4})$
θ_2	Unscaled pop. size *Mz-Pn-Ab*	0.002	$4.47 \times 10^{-3}(3.95 \times 10^{-3})$
ρ	recombination rate	0.2	1.26(1.24)
rd	rate distribution	$\Gamma(n = 4, \alpha = 1.0)$	$\Gamma(n = 4, \alpha = 1.9)$
τ_{min}	minimum τ	10^{-6}	NA
θ_{min}	minimum θ	10^{-6}	NA

Assuming a speciation time of 10My (T_{znb}) between *A. burtoni* and the common ancestor of *M. zebra* and *P. nyererei*, we obtain the following equation:

$$T_{zn} = 10 \times \frac{\tau_{zn}}{\tau_{znb}} \tag{3}$$

Here τ_{zn} and τ_{znb} correspond to the unscaled divergence of the *A. burtoni-M. zebra-P. nyererei* split and *M. zebra-P. nyererei* split, respectively. Using median unscaled divergence estimated by coalHMM (Table 1), we obtain $\tau_{zn} = \tau_1 = 1.71 \times 10^{-3}$, $\tau_{znb} = \tau_1 + \tau_2 = 2.02 \times 10^{-3}$, and a divergence time T_{zn} of 8.47 My. The fraction $s = \tau_{znb}/\tau_{znb}$ measures the time between speciation events and approaches 1 when speciation events occur in rapid succession. In the investigated haplochromine cichlids, s equals 1.18. Using this pipeline on chromosome 19 of human, chimpanzee and gorilla genome alignments with the Sumatran orangutan as outgroup, I found a fraction s of 1.51, similar to a genome-wide fraction of 1.63 reported by (Scally et al., 2012). The low s estimated for the three haplochromine cichlids therefore suggests that the three lineages arose in rapid succession.

4.3 Selection in morphogenesis, vision and pigmentation genes in East African cichlids

Adaptive radiations of East African cichlids partly depend on strength of sexual selection and ecological opportunities existing in different lakes in East Africa (Wagner et al., 2012). One hypothesis is that cichlid-specific evolutionary innovations have played an important role in these radiations by allowing cichlids to exploit ecological opportunities or quickly respond to changing natural and sexual selection pressures.

We selected 30 genes known to be involved in morphogenesis, vision or pigmentation with previous evidence of rapid molecular evolution in cichlids to increase power (personal communication, Ole Seehausen and Todd Streelman). Of these 30 genes, 22 were found (Table 5; 8 were missing in one more more cichlids/teleosts: Bmp9, Bmp3b, Bmpr1a, B-catenin, Sws2b, Wnt4, Edar) in all five cichlids and I sought to identify positively selected genes that might have played a role in cichlids-specific innovations in the ancestor of haplochromines and in the ancestor of haplochromines and Lamprologini (i.e. ancestor of haplochromines and *N. brichardi*). Of these 22 genes, I identified three, *ednrb1*, *kfh-g* and *rho*, that have undergone accelerated evolution in the ancestors of haplochromines and Lamprologini (Table 5). All three proteins belong to the G protein–coupled receptor gene family. Green-sensitive opsin (kfh-g) and Rhodopsin (rho) are proteins important in vision and may have undergone accelerated evolution owing to changing ecological environment or evolutionary pressures related to sexual selection. (Yang and dos Reis, 2011)

Because EDNRB1 is known to affect colour patterning in both fishes and mammals (Parichy et al., 2000; Shin et al., 1997), I hypothesised that

Table 5: dN/dS analysis for selected genes. M0, one ratio model versus M2[a], branch test with one ratio.

	Lamprologini		Haplochromine	
Gene Name	$2\Delta L$	corrected p-value	$2\Delta L$	corrected p-value
csf1ra	8.2	n.s.	1.22	n.s.
runx2b	0.25	n.s.	0.17	n.s.
bmp4	0.25	n.s.	0.52	n.s.
irx1b	0.79	n.s.	2.81	n.s.
dlx2a	0.6	n.s.	0.89	n.s.
wnt4a	-0.14	n.s.	0.0	n.s.
eda	5.09	n.s.	4.91	n.s.
cntn4	1.23	n.s.	0.04	n.s.
wnt1	0.54	n.s.	0.0	n.s.
bmp2a	-0.05	n.s.	0.73	n.s.
rho	14.58	0.016	4.72	n.s.
ctnnb2	1.0	n.s.	0.45	n.s.
edar	0.0	n.s.	3.28	n.s.
foxp2	2.4	n.s.	0.01	n.s.
rh2-b	2.4	n.s.	0.01	n.s.
opn1sw1	0.12	n.s.	9.02	n.s.
kfh-g	15.59	0.009	0.25	n.s.
foxg1a	0.26	n.s.	0.56	n.s.
bmpr1aa	0.0	n.s.	1.32	n.s.
fgf10a	1.64	n.s.	0.42	n.s.
gdf10b	1.68	n.s.	0.04	n.s.
ednrb1	13.69	0.049	1.57	n.s.

changes in the EDNRB1 protein might have functional consequences on the colour patterning in East African cichlids. Previously, the endothelin pathways (ET1/EDNRA and ET3/EDNRB1) were suggested to play a role in the development of the pharyngeal jaw apparatus in all cichlids and in anal fin egg-spots in haplochromine cichlids (Diepeveen and Salzburger, 2011). In their study, Diepeveen and Salzburger (2011) studied the partial DNA sequence of 26 East African cichlid *ednrb1* sequences and concluded that *ednrb1* was under strong purifying selection. The complete genomic sequences of the five cichlid species allowed us to determine four additional amino acid substitutions (Figure 29) in ancestral East African cichlids after their split from *O. niloticus* (blue in

Figure 29), and helped to confirm one substitution that occurred in the ancestor of haplochromines (red in Figure 29).

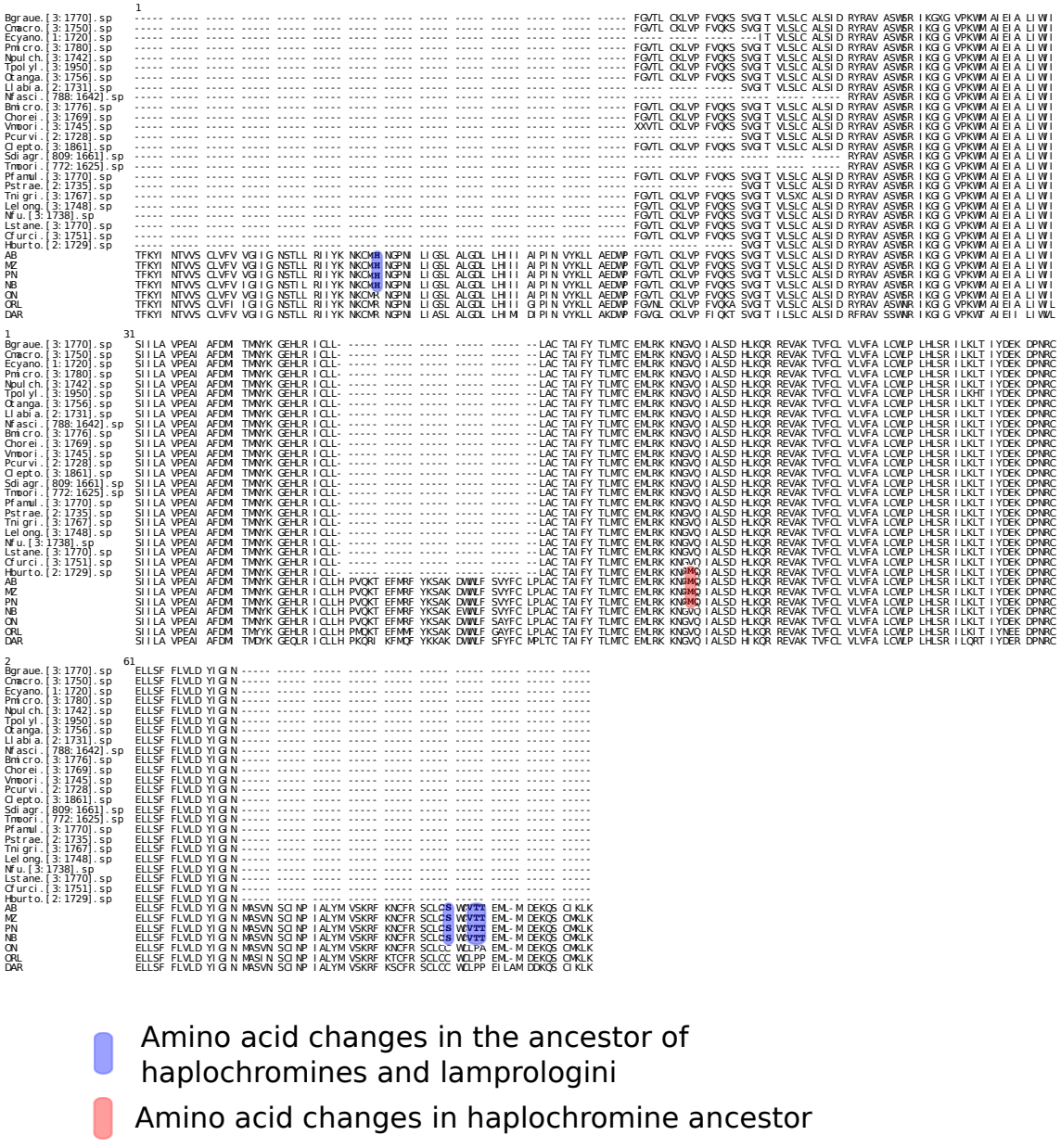

Figure 29: Comparison with 24 partial *ednrb1* sequences Sequences from East African cichlids from Diepeveen and Salzburger (2011). ORL: medaka, DAR: zebrafish.

The endothelin system consists of G protein–coupled endothelin receptors that are activated by endothelin signaling peptides. Previous work on endothelin duplications in chordates showed that divergent sites between paralogs were generally located in the extracellular regions of the 7 transmembrane regions, which were associated with variation in ligand peptides (Braasch et al., 2009). In East African cichlids, I found that the sequence of the mature endothelin, ET3, was completely conserved across cichlids (data not shown) and no variation was apparent from the receptor regions of EDNRB1, suggesting no variation in ligand-receptor interaction. However, all five amino acid changes are located in intracellular regions of EDNRB1 (Figure 30), which can interact with G proteins. The site in the first intracellular loop of EDNRB is among three sites (Arg to Histidine in Met-Arg-Asn) that were shown to be required to activate a downstream target, SRF, in human by interacting with the G protein $G_{\alpha 13}$ (Liu and Wu, 2003). Additionally, the cluster of 3 sites overlaps with three cysteines known to be important for palmitoylation of human EDNRB1 (Doi et al., 1997; Okamoto et al., 1998). The second of the three cysteines (Cys-Cys-Trp-Cys) is substituted by a serine (Cys-Ser-Trp-Cys) and may affect the anchoring of the C-terminus of EDNRB1 to the transmembrane domain. Proper palmitoylation of EDNRB1 C-terminus is thought to be required for G protein coupling in several G protein coupled receptors including the EDNRB1 paralog *Ednra* (Moffett et al., 1993; Horstmeyer et al., 1996). Overall, these results indicate functional changes in the interaction between EDNRB1 and downsteam G protein signalling. Thus, EDNRB1 is a notable candidate gene for which amino acid substitutions may have contributed to cichlid-specific traits important for their explosive radiation.

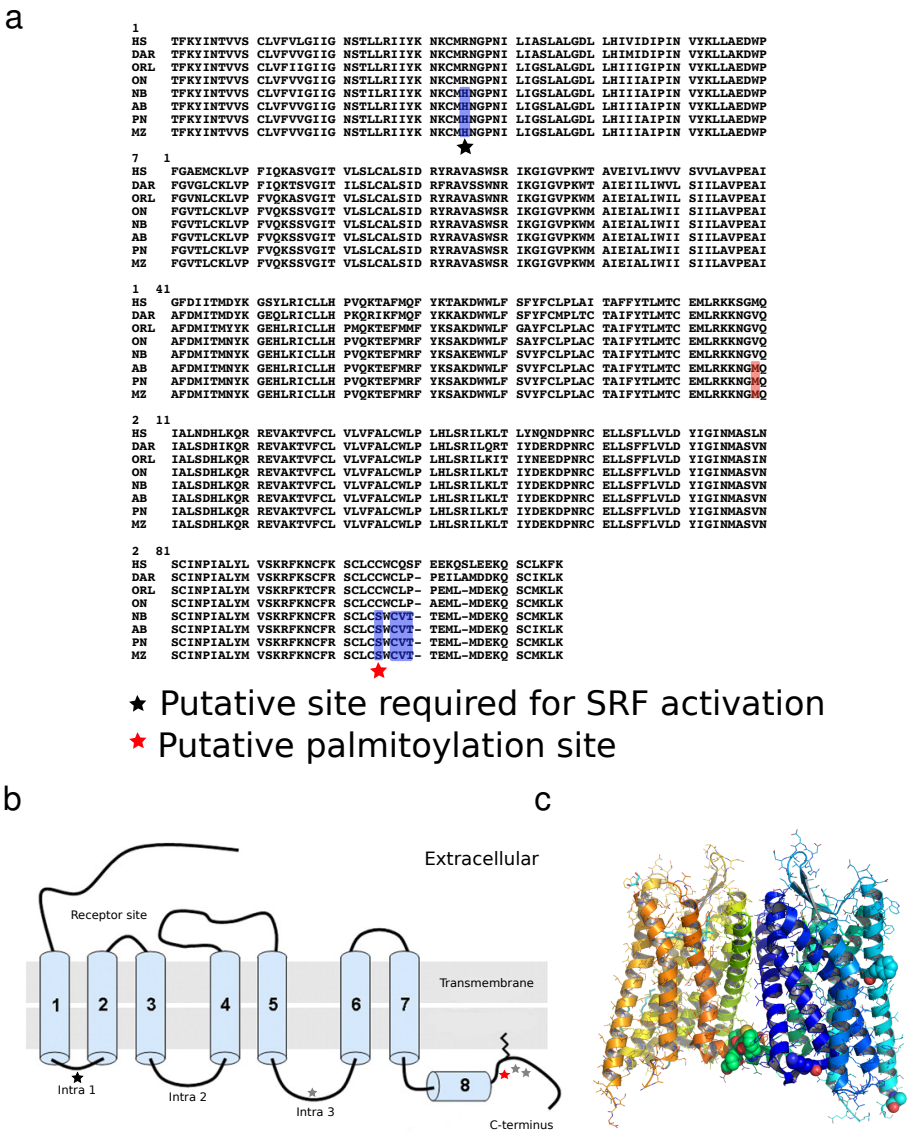

Figure 30: EDNRB1 variable sites and G proteins interaction.
(a) Alignments of EDNRB1 in cichlids with human (HS), zebrafish (DAR) and medaka (ORL). Black star denotes site shown to be required to activate SRF in human by interacting with the G protein $G_{\alpha 13}$ (Liu and Wu, 2003). Red star denotes site that may affect the anchoring of the C-terminus of EDNRB1 to the transmembrane domain (Okamoto et al., 1998). Highlighted are amino acid substitutions in the common ancestor of haplochromine and lamprologini (blue) and in the ancestor of haplochromine (red) (b) Location of substitutions on 7 transmembrane domain representation (Adapted from (Lalueza-Fox et al., 2007)) (C) Sites (red spheres) on the structure of the human kappa opioid receptor in complex (4DJH). Only the right homodimer is annotated.

4.3.1 Materials and Methods

Selection analysis

Sequences from 30 genes were obtained for zebrafish or medaka from Ensembl 71 and were used as seed sequences to search, using Blastp, the medaka protein database (also from Ensembl 71) and our tilapia protein database. Once the tilapia ortholog was identified, we projected the corresponding sequence onto the four other cichlids' genome to retrieve the remaining cichlid orthologs. Multiple sequence alignments including sequences from zebrafish, medaka and the five cichlids were then produced using muscle (Edgar, 2004). All alignments were curated manually and truncated to remove poorly aligned sequences (22 genes remained after manual curation). An in-house peptide sensitive approach was used to align the cDNA into codons and used codeml/PAML to test M0, a one-rate model which assumes the same rate of evolution in all branches against $M2^a$, a branch site test with one rate for the background and one rate for the specified branch (Yang, 2007). Two branches, the one leading to the lamprologini and haplochromines and another leading to the haplochromines, were tested. All tests were performed in triplicate and were checked for convergence. The p-values from the likelihood ratio tests were corrected using Bonferroni correction assuming 44 independent tests (22 genes and two branches).

Retrieving missing *N. brichardi ednrb1* exons from RNA-seq reads

EDNRB1 was located in a region of the *N. brichardi* genome assembly with gaps which led to 4 missing exons in the *N. brichardi* EDNRB1 gene model. The full *ednrb1* transcript was also missing from the Trinity *de novo* transcriptome assembly. Consequently, to recover the transcript, RNA-seq data from eight *N. brichardi* tissues was mapped to the EDNRB1 cDNA transcript of the four other cichlids using stampy (Lunter and Goodson, 2011). Reads that were mapped with less than 5 edit distance (# of mutations and gaps) away from the reference transcript were then assembled using cortex (Iqbal et al., 2012).

Comparison of partial *ednrb1* genomic sequences from Diepeveen and Salzburger (2011)

The genomic sequence of all 26 cichlids *ednrb1* were retrieved from Supplementary Table 1 of (Diepeveen and Salzburger, 2011). Genewise (Birney et al., 2004) was then used to map the *A. burtoni* EDNRB1 peptide sequence to the genomic sequences and to extract the corresponding cDNA. Two truncated sequences, from *Gnathochromis permaxillaris* and *Variabilichromis moori*, were removed. The remaining 24 *ednrb1* cdna sequences were then aligned with the sequences from zebrafish (DAR), medaka (ORL), and the five sequenced East African cichlids (ON, NB, AB, MZ, and PN).

Structural analysis of nonsynonymous sites on EDNRB1

To find a proper structure to model the location of the nonsynonymous sites, both BLASTP and HHpred were used with the medaka EDNRB1 peptide sequence to identify homologous proteins for which structure was available. A concordant hit was found to be the human kappa opioid receptor (4DJH). Variable sites were then mapped onto this structure using PyMol.

4.4 Gene architecture evolution in East African cichlids

Divergence in protein sequences among closely-related species is thought to generate phenotypic variation. A large number of studies have contributed in the compilation of a comprehensive list of evolutionary mechanisms that transform protein coding genes and/or drive protein diversity. For example, the analysis of gene coding sequences in multiple species has shown that simple mutations generate protein differences through non-synonymous substitutions and result in phenotypic variation. However, researchers have only recently started to study the contribution of alternative splicing in the rapid generation of protein diversity and/or differences between closely related species.

Here, I wanted to exploit the transcriptome and genomic data generated in East African cichlids to investigate gene architecture evolution at a fine scale by studying differences in transcripts expressed in closely-related species.

4.4.1 Results and Discussion

Obtaining a list of orthologous exons.

To study changes in gene structures that occurred during East African cichlid evolution, I wished to identify a high-confidence set of orthologous exons. This is because my aim was to study changes in exon splice sites and splicing ratios. Using whole genome multiple alignments and various conservative filters (Materials and Methods), I constructed a list of highly confident orthologous exons and exon clusters. Exon clusters represent clusters of overlapping exons of which 221,861 (69.1%) were found to be present in all five cichlids. The remaining 30.9% of exon clusters were either found to have ambiguous mapping in one or more lineages, to be lineage-specific, or to be unmapped (Materials

and Methods). All downstream analyses used a subset of these 221,861 exon clusters to allow cross-species comparisons, unless otherwise stated.

Discovery of alternative spliced exons and alternative splice sites

To annotate alternative splicing events, I mapped RNA-seq reads from each of the five cichlids onto their respective genomes (Materials and Methods). Reads mapping across putative introns were collected and checked for canonical splice sites. Introns were then used to define alternative splicing events (Materials and Methods). To avoid errors in downstream splice ratio estimation due to incomplete transcript annotations, I considered only exons which showed simple patterns of alternative splicing, i.e. skipped exons that are flanked by constitutive exons and exons with only one alternative splice site (Materials and Methods). Overall, I discovered 6,095 exons showing evidence of alternative skipping within at least one sample. This number decreases to 2,890 and 2,070 when counting exons showing evidence of alternative skipping in at least 3 different species and 3 different organs, respectively. Moreover, 1,874 exons were predicted to have at least one alternative splice site in at least one tissue and one species.

Conservation profile of gene expression levels and alternative splicing ratios

To investigate global patterns of gene expression and alternative splicing evolution, I obtained gene expression levels of 16,009 orthologous genes from (Brawand et al., under review) and used MISO (Katz et al., 2010) to quantify the percent spliced in (PSI) ratios between exons that are alternatively skipped (Materials and Methods). The PSI of 6,095 alternative skipped exons (ASEs) were computed across all samples. When comparing PSI profiles of all ASE,

samples clustered according to species, while samples clustered by tissue-type when comparing gene expression profiles (Figure 31a and b).

However, when restricting our analysis to exons showing evidence of skipping in at least 2 of the 5 cichlids, inclusion ratios of ASE started to show clustering according to tissue-type. This by-tissue clustering is stronger when restricting to exons showing evidence of skipping in at least 3 cichlids (Figure 31c).

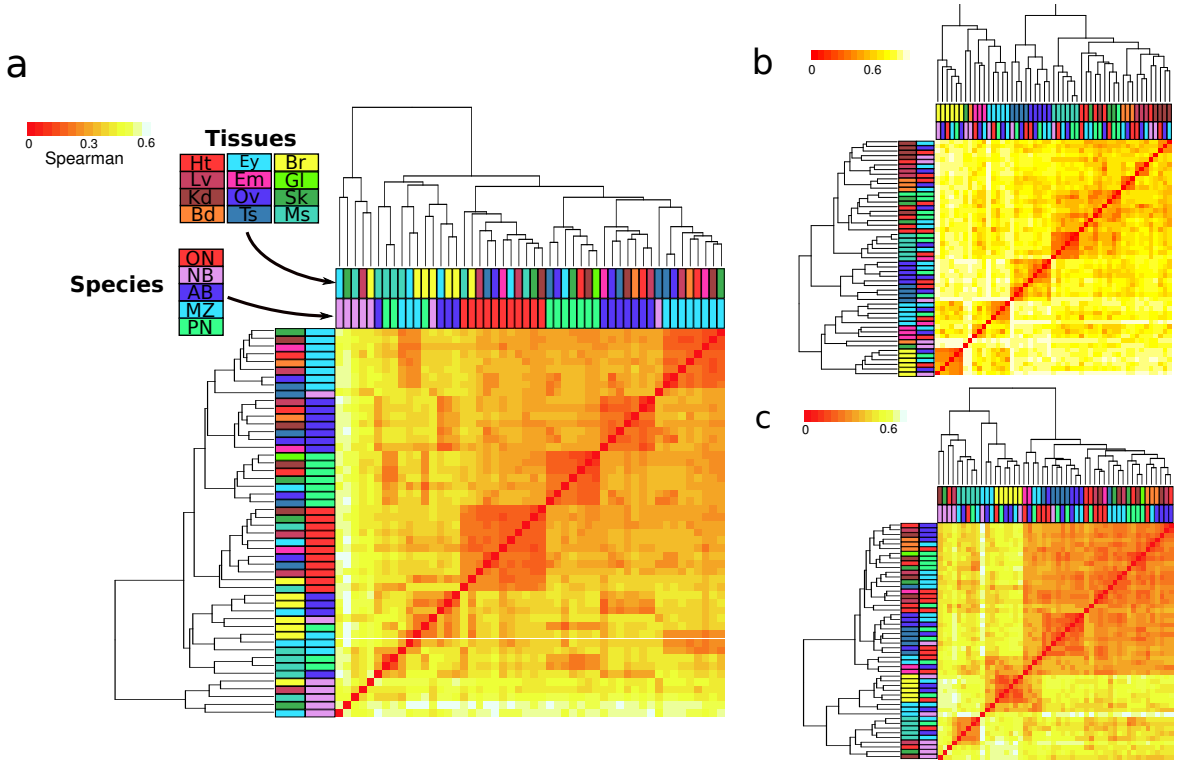

Figure 31: Heatmaps of gene expression and alternative splicing levels in cichlids. (a) Heatmap of Spearman correlation according to PSI for all skipped exons (b) Pearson correlation according to expression and (c) Spearman correlation according to PSI for exons alternatively skipped in 3 or more species. Ht: heart, Ey: eye, Br: brain, Lv: liver, Em: Embryo, Gl: gills, Kd: kidney, Ov: ovary, Sk: skin, Bd: blood, Ts: testis, Ms: muscle.

The clustering of sample by organ in terms of gene expression profiles was previously reported in mammals (Brawand et al., 2011). However, at the time of this analysis (in 2012), no published study reported a clustering of sample by

species in terms of exon usage. In late 2012, two studies (Merkin et al., 2012; Barbosa-Morais et al., 2012) identified a similar clustering of sample by species when they studied exon usage in mammals. In particular, Barbosa-Morais et al. (2012) showed that this clustering by species holds even when comparing exon usage levels in tissues from two closely-related primates (human and chimp) that diverged only ∼4Mya. Interestingly, I found that when considering exons that show evidence of skipping in at least 3 of the 5 cichlids, the clustering seemed to shift back to a by-organ clustering such as observed for gene expression levels. The by-organ clustering appears to be strongest for muscle, brain and testes – an observation that has also been made by Merkin et al. (2012). This may reflect a higher number of functional differential splicing events in brain, muscle and testes. This also suggests that while the alternative splicing of some exons is important for cellular function (e.g. in brain tissues), the usage of a large number of exons evolves neutrally (lineage-specific AS). This hypothesis is consistent with the idea that there exists a large amount of noisy splicing events that are poorly conserved, but result from an imperfect splicing machinery rather than selection for an expanded protein repertoire (Pickrell et al., 2010b). Additionally, Reyes et al. (2013) found that, in primates, most exon usage shows small inter-species differences that they suggest may reflect neutral drift.

Evolution of gene architecture

I next attempted to survey gene architectural changes in terms of indels and changes in 5' or 3' splice sites. I compiled a list of 79,974 coding exons with unambiguous orthologs for which both 5' and 3' splice sites were supported by reads mapping across splice sites in all five species (Materials and Methods). I found that almost all exons (70,471, 89.2%) had preserved splice sites in all

five species and showed no change in length. However, 2,320 (2.9%) exons had at least one indel and 6,129 (7.8%) exons had variation in splice sites. As expected, the distribution of size change is highly biased towards a multiple of 3 for indels (ratio 5.2 while the expected ratio is 0.5), but less so for newly gained or lost splice sites (ratio 0.9; Figure 32). Compared to indels which only modestly affect the length of exons (and thus gene), the gains and losses of a splice site can often generate much more variable isoforms. However, given the large number of changes that are expected to disrupt coding frame, many of the gained or lost splice sites may be/have been the product of noisy splicing.

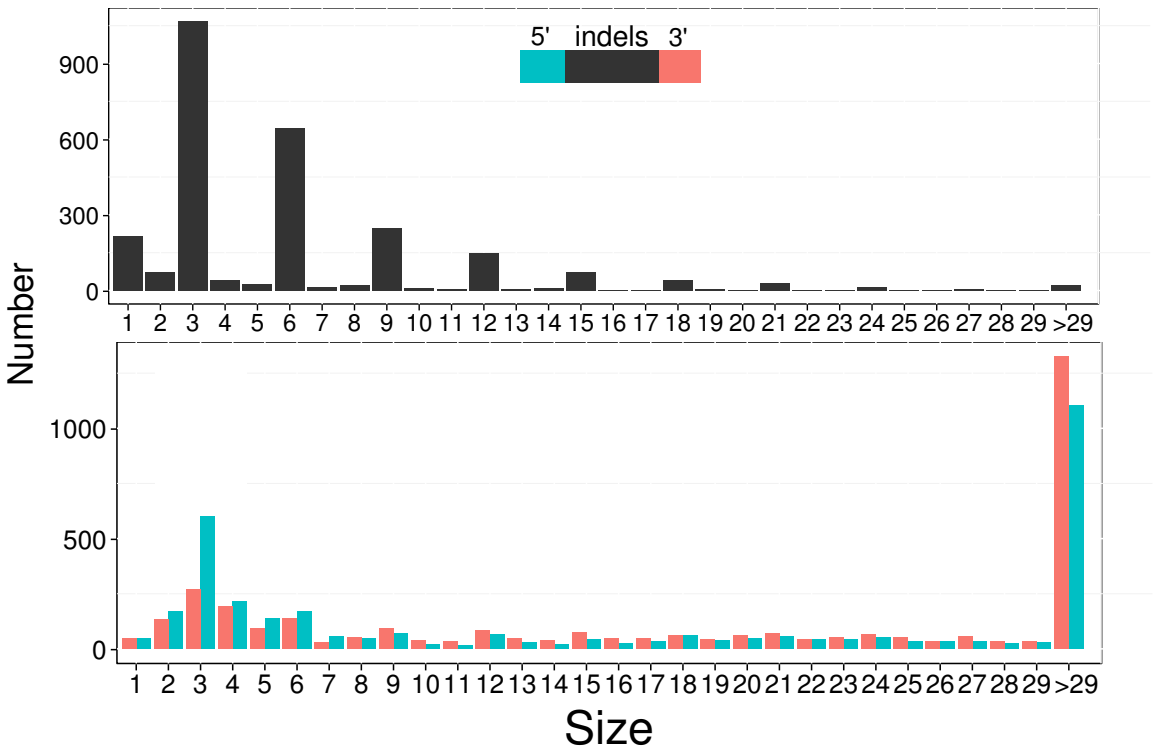

Figure 32: Exon length differences caused by genome variation
Distribution of length differences of exons from indels and from 5' and 3' splice site variation. The bias for exonic changes that are a multiple of 3 is much stronger for indels compared to alternative splice sites.

I then hypothesized that many of the splice site gains or losses are lineage-specific and might not represent variation that evolve under selective pressures.

Using maximum parsimony, I predicted the lineages in which gains and losses occurred during the evolution of the five cichlids (Materials and Methods). In total, I found that 576 gains and 509 losses were lineage-specific, while 147 gains and 65 losses were predicted to have occurred in the ancestors of at least 2 species (Figure 33a). These numbers, when normalised by the estimated divergence time, are similar (assuming that the *Nb-Mz-Pn-Ab* divergence occurred ∼10My prior to the *Mz-Pn-Ab* divergence). This suggests that there is a constant turnover of splice sites during cichlid evolution and these gains and losses can be retained for over ∼25My.

Previous studies suggest that lowly used splice sites evolve under weaker selection compared to splice sites that are used at higher levels (Ermakova et al., 2006). I therefore expected that splice sites that were lost during evolution tend to be used at low levels. I found that a surprisingly large proportion of splice sites lost were constitutively spliced in the ancestral state (Figure 33b). After inspecting several examples of lost splice sites, I found that a common mode of loss is a disruption of the splice site (mutation) in the presence of a near compensatory AG or GT dinucleotide (Figure 33c). Though, a genome-wide analysis is required to establish exactly how widespread this mechanism is, there is evidence in *Arabidopsis* that a majority of transcript model differences can be traced back to the disruption of ancestral splice sites that are rescued by compensatory splice sites (Gan et al., 2011). In these cases, the compensatory splice sites become the new splice sites that are constitutively used (Figure 33b). However, in line with our expectations, a large number of newly acquired splice sites are used at low levels (Figure 33b), and are not expected to play major functional roles.

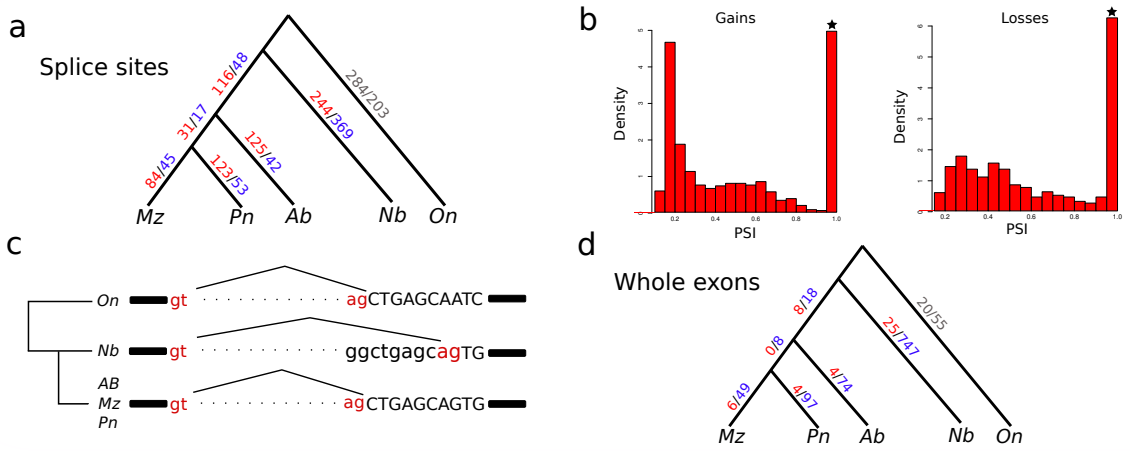

Figure 33: Gains and losses of splice sites and whole exons.
(a) Mapping the gains (red) and losses (blue) on each branch using a maximum parsimony approach and *O. niloticus* as outgroup. (b) The inclusion ratios (PSI) of gained splice sites and the ancestral PSI of lost splice sites. (c) a class of exon (stars) have changed splice site completely after the ancestral splice site is disrupted. (d) A similar analysis as (a) using inclusion ratios show that the majority of exon losses are lineage-specific (Materials and Methods).

I next identified exons with evidence of gain or loss in expression across four cichlids and using *O. niloticus* as an outgroup (Figure 33d). These exons were also required to be constitutively skipped in the species without expression. As with the splice site turnover analysis, the lineages in which the exons were gained or lost were inferred using maximum parsimony. In contrast to the splice site losses, I found that a disproportionate number of exonic losses were predicted to be lineage-specific rather than in inner branches (26 in inner branch versus 49, 97, and 49 during haplochromine evolution). Only 8 gains were inferred to have occurred in ancestral cichlids and are shared in their haplochromine descendants. These results, though preliminary, suggest that exon gains and losses were not a major mode of gene evolution in cichlids. However, it should be noted here that this analysis was very conservative and it is likely that several more exons were gained over cichlid evolution but were undetected. Interestingly, I found one exon gain in the 5' UTR of Ednrb1, a gene that has been under accelerated evolution during cichlid evolution (see previous section). I also identified a putative causal mutation in the ancestors

of haplochromines (Figure 34a): a single A to T mutation created a canonical splice site (GA to GT) which is used in the haplochromines. Mapping RNA-seq data from eye transcriptomes of these cichlids revealed that while the novel exon in haplochromines is spliced constitutively in Ednrb1 transcripts, the orthologous position in *O. niloticus* and *N. brichardi* are always skipped (Figure 34b). Because this exon is located in the 5' UTR and not within the coding body of Ednrb1, it is difficult to assign a function to this gain. Nevertheless, 5' UTRs are known to be important for post-transcriptional regulation (Dvir et al., 2013; Lawless et al., 2009). The novel exon in Ednrb1 could therefore have contributed to changes in its regulation in the haplochromines (also see Discussion).

The interest in understanding gene structure evolution at a higher resolution was appreciated by (Perry et al., 2012), who investigated gene structure differences in primates. To this end, they searched transcriptome alignments for internal gaps bigger than 50bp in at least one species, but with assembled sequence in at least one other species. Contrary to previous expectation, Perry and colleagues found that most gene structures show strong signs of conservation across long evolutionary time (they found only 308 potential exon structure changes, 304 of which could be explained by alignment artefact or alternative splicing). My analysis of gene structure in cichlids revealed that $\sim$10% of all exons experienced structural changes, either owing to indels or to variation in splice sites. These structural changes are generally small and tend not to alter coding frame. A large fraction of splice sites gained are predicted to be used at low levels and may represent noisy splicing. These observations suggest therefore that structural changes in genes have not been a primary source of phenotypic variation in cichlids. Nevertheless, isolated cases such as the novel exon in the

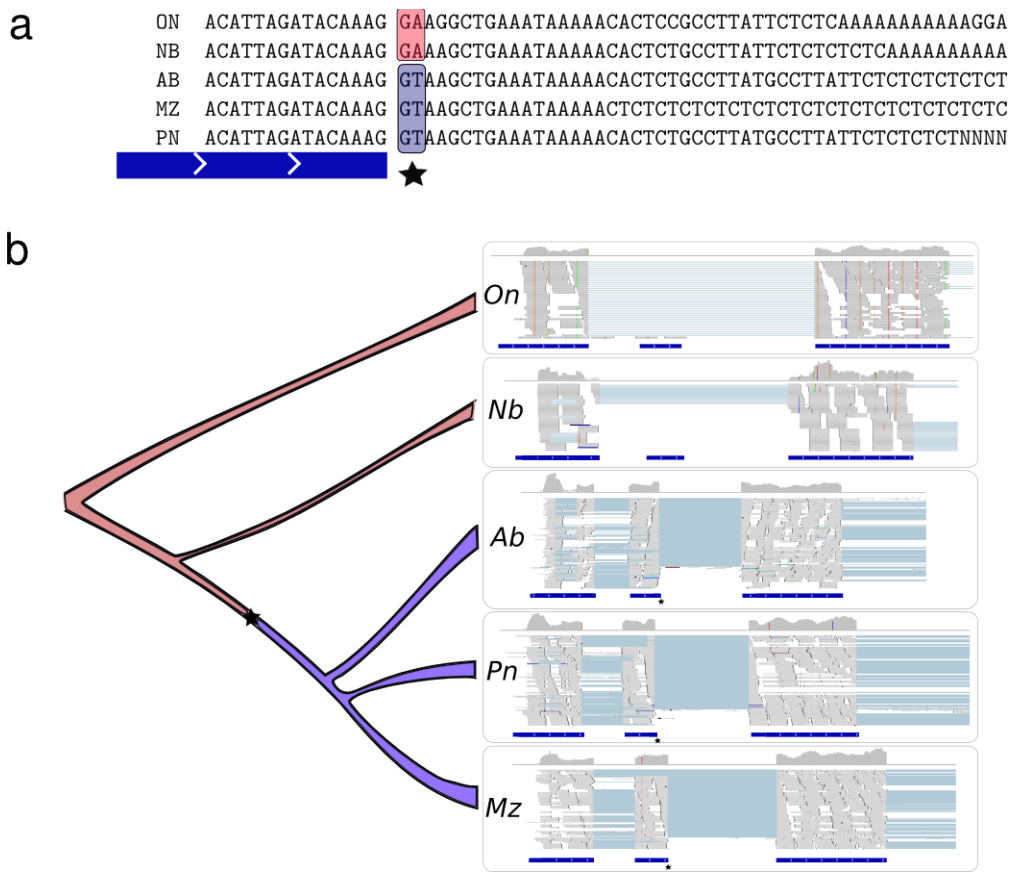

Figure 34: Gain of exonic region.
(a) Sequence of Ednrb1 5' UTR region, the putative causal SNP is highlighted and is predicted to have arisen in the ancestors of haplochromines (*A. burtoni*, *M. zebra*, and *P. nyererei*). (b) Expression and splicing in eye tissues from the five cichlids suggest an exonic gain in haplochromines.

5' UTR of Ednrb1 may have played a role in functional diversification (see Discussion).

4.4.2 Materials and Methods

Mapping and discovery of splicing events

The trimmed RNA-seq read libraries were mapped using GSNAP (Wu and Nacu, 2010) to their respective genomes with standard parameters without annotation. For each library, spliced reads, i.e. reads mapping across putative

introns, were collected and checked for canonical splice sites. Only splice reads harbouring introns with canonical splice sites were kept and used to define spliced introns.

Projection of gene models and orthologous exons

MAF alignments were obtained from the Broad Institute and used to project the gene models of *M.zebra*, *N.brichardi*, *A.burtoni* and *P.nyererei* onto the genome of *O.niloticus*. Subsequently, all projected and native gene models from *O.niloticus* were projected back onto each of the four other cichlids using liftOver. The sizes of projected exons were required to be within a factor of 1.5 compared to the native exon. This resulted in mappings between exons across the annotations of different cichlids. MAF alignments centred around *M. zebra* were also produced using MULTIZ (Blanchette et al., 2004). This second alignment was used to validate the projections from the *O. niloticus* centred alignment and to map exons from unaligned regions in the *O. niloticus* centred alignment. A breakdown of the number of exons mapped can be found in Table 3.

All projected exons were then clustered together in each species separately to form exonic regions. These clusters represent exons that overlap or are connected by exons that overlap. Next, the orthology between the clusters from each species was determined by using the orthology of their constituent exons.

Predicting alternative splicing events and their orthologous events

To predict alternative splicing events, I produced an in-house exon-centric pipeline. The cichlid annotations were used to identify internal exons. An internal exon, e_M, was considered to be skipped if there is an exon triplet (e_U, e_M, e_D) such that there is at least one sequenced sample in which (1) there

are at least 3 mapped spliced reads spanning the junction of an exon upstream, e_U, and an exon downstream, e_D and if (2) there are 3 mapped spliced reads spanning the junction between the exon upstream and the internal exon, and between the internal exon and the exon downstream. Similarly, an exon e_{as} was considered to have an alternative splice site if there is an exon pair (e_{as}, e_c) such that there are at least 3 reads in at least one sample spanning two distinct e_{as} junctions and e_c.

Once all alternative skipped exons and alternative splice sites were predicted in individual samples, all alternative splicing events were merged and projected across species. To compare alternative splicing patterns across species, only those alternative splicing events involving exons that were projected successfully in all species were used.

Characterising alternative splicing differences

To compute relative inclusion frequency (or percent spliced in) of alternative splicing events in each separate library, MISO (Katz et al., 2010) was run on each RNA-seq dataset separately. To detect signatures of conserved profile of splicing events, Spearman correlations were computed using percent spliced in (PSI). Only events which show alternative splicing, e.g. $5\% < \text{PSI} < 95\%$, in at least one library in every species were included in the profile comparison. Because complex alternative splicing events can bias the estimation of PSI, our analysis focused only on alternative splicing with two variants.

Inferring gains and losses of splice sites and exons

Every splice site (and exon) from *M.zebra*, *N.brichardi*, *A.burtoni* and *P.nyererei* were mapped to the *O.niloticus* genome. A splice site (exon) is con-

sidered absent if it does not have support from any spliced reads from any of the sequenced tissues. The *pars* program of PhyML was used to infer the presence/absence of splice sites (exons) at nodes of the phylogenetic tree.

5 Chapter 5: Analysis of small internal exons and their function in human brains

Eukaryotic genes often consist of multiple exons that must be recognised to allow efficient production of functional proteins. Because internal exons in mammals are located between introns which are much longer (typically 10-100 times longer), the spliceosome must detect small exonic islands within a large intronic sea. How the spliceosome recognises exons remains an active field of investigation. A widely held belief is that signal sequences such as exonic enhancers (i.e. motifs within RNA exons that can be bound by splicing factors) are crucial for proper splicing of the pre-mRNA. I was therefore surprised that collaborators at the Lieber institute for Brain Development (Baltimore, MD) identified several exons as small as 6nt that are spliced in human brain transcripts. However, it was unclear at that point whether these small exons were functional, noisy splicing or simply evolutionary oddities.

For the study described in this chapter, I use comparative genomics to analyse small exons ($\leq$51nt) in human transcriptomes and show that they are likely to play important roles in cellular functions, notably in the brain. This chapter consists of work in preparation for submission (Li et al., in preparation).

All analyses presented here are my own, except for the analyses of protein tertiarty structures for which Luis Sanchez-Pulido made major contributions. Furthermore, Wilfried Haerty and Chris Ponting substantially edited the manuscript in preparation.

5.1 Background

Most vertebrate genes are divided into exonic sequences separated by large intronic stretches. Although alternative splicing of exons allows multiple protein isoforms to be produced from the same gene, many isoforms are expected to be nonfunctional (Pickrell et al., 2010b; Reyes et al., 2013). Consequently, the distinction of functional isoforms from those that serve no protein-encoded function represents considerable challenges both for genomics researchers and perhaps also for the cellular splicing machinery.

Exons are diverse in their characteristics and functions. They differ greatly in their nucleotide composition (Amit et al., 2012), inclusion patterns (Keren et al., 2010) and lengths (Sorek et al., 2004), all of which can affect their biological roles and how they are recognized during splicing. The length of an exon is often assumed to follow a symmetric distribution centered around an optimal size. In mammals, this optimal size appears to be approximately 140nt (Zhu et al., 2009; Gelfman et al., 2012), which is proposed to relate to the amount of DNA wrapped around single nucleosomes (Schwartz et al., 2009). The preferential positioning of nucleosomes within exons is hypothesised to aid exon recognition by acting as a "speed bump" to RNA polymerase II, thereby allowing the splicing machinery to catch up and splice out the intron immediately upstream (Schwartz et al., 2009).

Ultra-short exons are uncommon, difficult to detect, and have unknown functional roles. They are thus often ignored in transcriptomics studies. Popular RNA-seq aligners such as TOPHAT offer users an option to search for small exons, but still fail to identify most exons shorter than 21nt (Wu et al., 2013a). The number of micro-exons (exons of length $\leq$ 51bp) annotated in human shows

a sharp decrease as exon size decreases (Figure 36b). Our choice to study exons 51nt or shorter was based on their tendency to be skipped in mature transcripts, possibly because steric hindrance due to molecular crowding between large multimeric complexes inhibits spliceosome assembly (Dominski and Kole, 1991; Black, 1991; Simpson et al., 2000; Carlo et al., 2000). It was also proposed that to be accurately spliced, exons must possess a minimum number of exonic splicing enhancers that promote splicing (Blencowe, 2000; Fairbrother et al., 2002; Caceres and Hurst, 2013), something that is obviously challenging for short exons. These observations lend support to the hypothesis that micro-exons are rare because of significant evolutionary pressure on exon lengths to remain longer than 51nt. Annotated micro-exons by Ensembl which are sometimes as small as 1nt are therefore at odds with this expectation.

A priori, it is expected that many of these annotated micro-exons are the product of noisy splicing (Pickrell et al., 2010b) or are annotation errors. However, functional micro-exons have been previously identified (Carlo et al., 2000; Zibetti et al., 2010). For example, the inclusion of a 12nt micro-exon in *LSD1* is regulated by PTB (polypyrimidine-tract-binding protein) in a brain-specific manner (Xue et al., 2013) and contributes to neurite morphogenesis in mammals (Zibetti et al., 2010). Yet, the total number of functionally important micro-exons is as yet unknown. Therefore, not only is the study of micro-exons of interest in the context of splicing mechanisms, but mutations controlling splicing of functional micro-exons is also of relevance to disease genetics.

Here, we provide a comprehensive characterisation of micro-exons at the DNA conservation, RNA splicing, and protein tertiary structure levels. We first identified functional micro-exons by examining the vast transcriptome datasets

currently available, and by detecting their signatures of vertebrate conservation. Next, we uncovered patterns of micro-exon usage by studying the inclusion of micro-exons across a large number of tissues. Furthermore, because the splicing of short exons is thought to present particular difficulties to the cellular machinery (Dominski and Kole, 1991), we sought to identify regulatory signals that enhance the recognition and usage of micro-exons. Finally, we studied how the inclusion of alternatively spliced micro-exons impacts on the tertiary structure of a protein.

In this study, we analysed over 57 billion reads from 876 human RNA-seq libraries including 25 post-mortem brain samples across development (Mazin et al., 2013), 345 samples from post-morterm prefrontal cortices (Lonsdale et al., 2013) and 506 samples from diverse other tissues (Lonsdale et al., 2013; Illumina Human Body Map 2.0, 2011). We show that, contrary to our expectations, most micro-exons are highly conserved across vertebrates. Moreover, we found that alternatively spliced micro-exons possess signatures of tissue-specific usage, and can alter protein-protein interactions. Notably, inclusion of micro-exons can be regulated by splicing factors belonging to the Rbfox family in a brain-specific manner. To the best of our knowledge, we are also the first to provide examples of alternatively spliced micro-exons that can regulate protein-protein interactions by altering protein tertiary structure. These results should renew our appreciation of smaller sized exons and should encourage the use of micro-exon aware algorithms, such as ATMap that we describe below or others such as OLEGO (Wu et al., 2013a), for the study of transcriptomes in the contexts of brain development, disease and beyond.

5.2 Results

Discovery of micro-exons and quantification of their usage

To complement current Ensembl annotations, we first sought to identify novel micro-exons from a large number of available RNA-seq data sets from human brain (307 samples), muscle (74 samples) and nerves (47 samples) using a discovery pipeline (Materials and Methods; Figure 36a). In total, we identified 234 putative novel micro-exons, 230 of which were included in transcripts expressed in these brain samples. In comparison, 7,085 of 12,835 Ensembl-annotated micro-exons were included in transcripts expressed in the brain (Figure 36c). This indicates that most human micro-exons are Ensembl-annotated and our analysis is representative of micro-exons in general.

Among several novel micro-exons we identified was a 6nt micro-exon which we could map to a highly conserved region of the *CACNA1A* gene (Figure 36d). This encodes a calcium channel, voltage-dependent, P/Q type, alpha 1A subunit which is mutated in spinocerebellar ataxia type 6, a familial hemiplegic migraine and episodic ataxia type 2 (MIM 108500, 141500 and 183086). Interestingly, we were able to map the two amino acids (NP) that are encoded by this micro-exon to a loop linking the S3 and S4 regions of CACNA1A (Payandeh et al., 2012). According to Payandeh et al. (2012), this loop has a dynamic connection to S4, which moves during channel gating. The alternative inclusion of the micro-exon may therefore generate two *CACNA1A* isoforms with different gating kinetics.

We also verified that our predictions of chromosomal positions of micro-exons corresponded to sequence that has been well-conserved across vertebrate evolution, as well conserved indeed as Ensembl-annotated micro-exons (Figure

35a). This observation justified the consideration in subsequent analyses of both novel and previously annotated micro-exons.

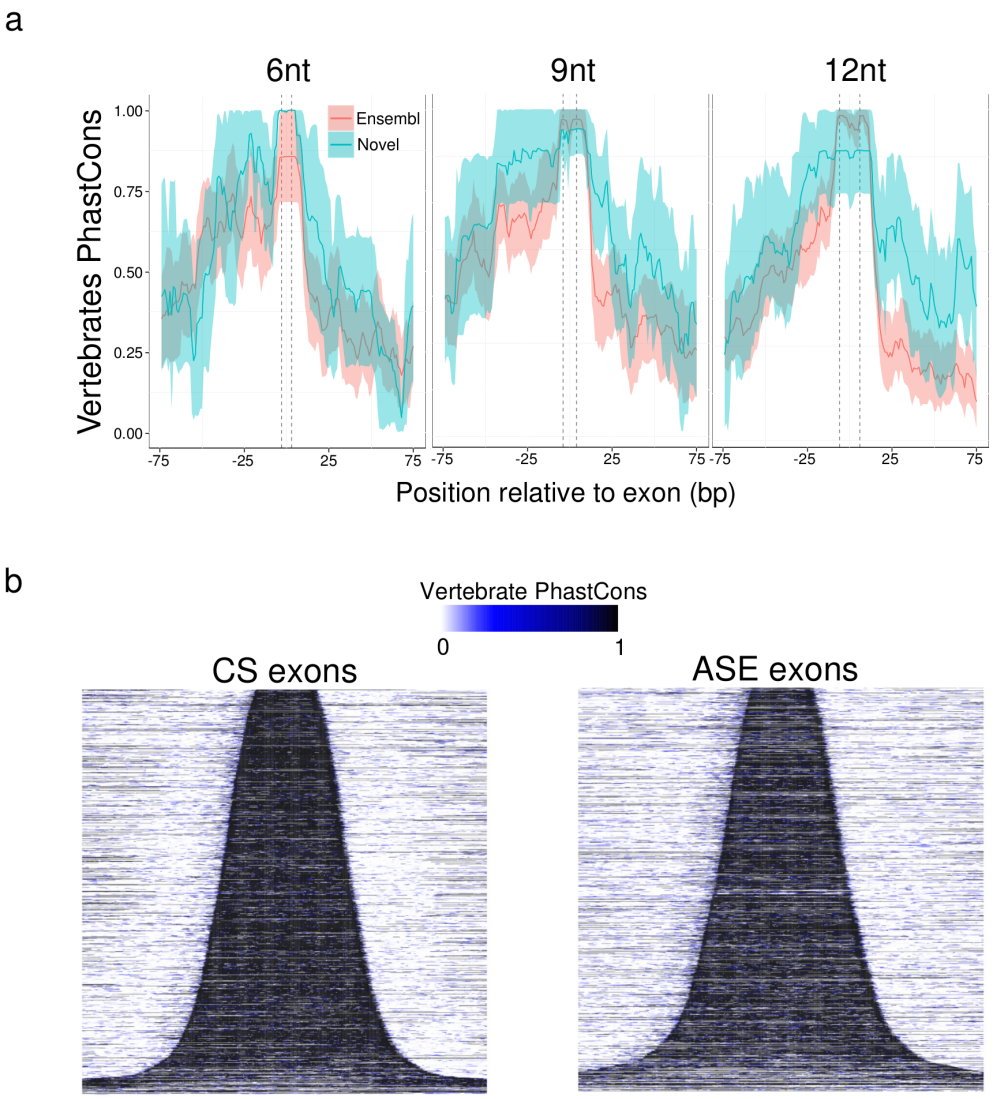

Figure 35: Vertebrate conservation of micro-exons and exons in general. (a) Vertebrate conservation profiles of 6, 9 and 12nt Ensembl-annotated versus novel predicted micro-exons. (b) Vertebrate conservation heatmap of 1,500 randomly sampled constitutively spliced (CS) and alternatively spliced exons (ASE). Each exon was centred within a 300nt window.

Interestingly, we noted that nearly half (114 or 49.6%) of our novel micro-exons were 6 to 21nt in length compared to only 8.2% (581) of Ensembl-annotated micro-exons. This prompted us to investigate the ability of algorithms to detect

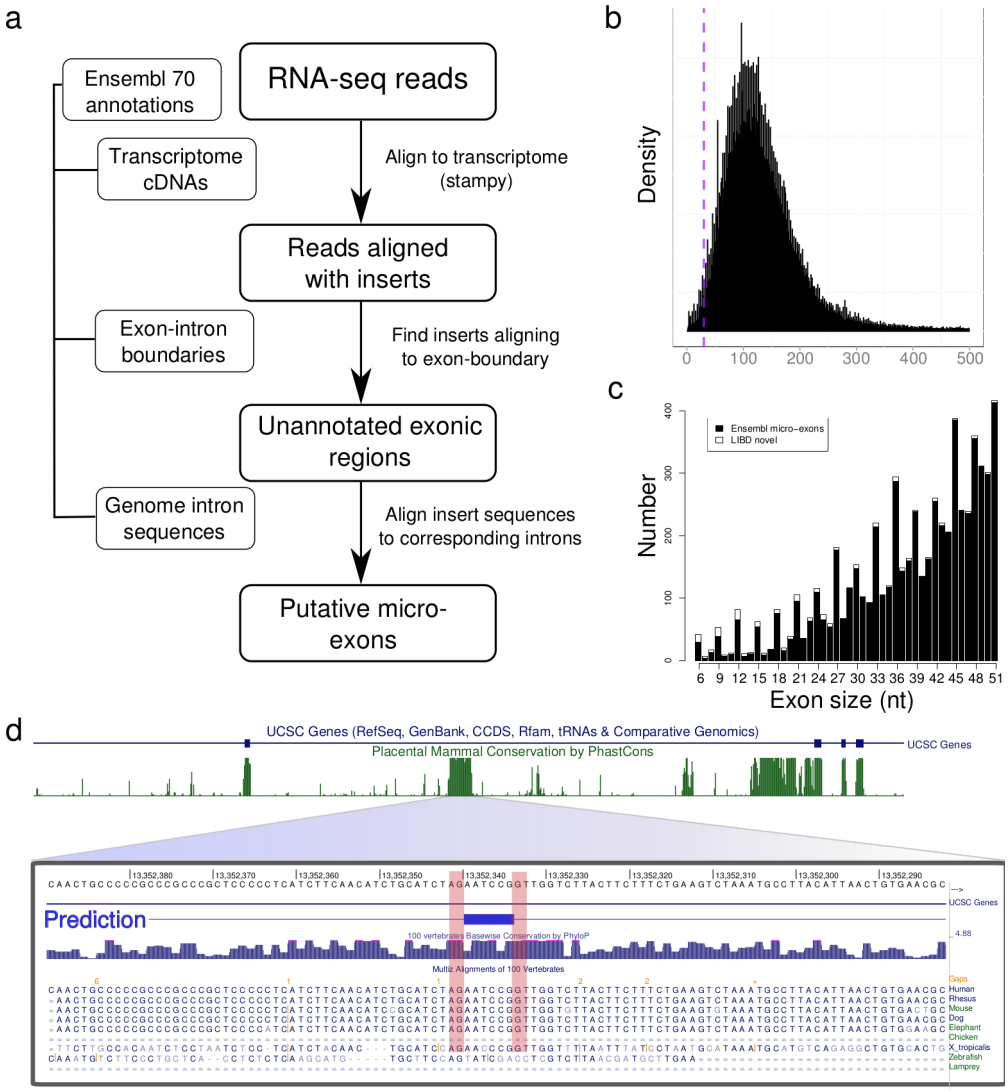

Figure 36: Identification of novel micro-exons

(a) Flow chart of our micro-exon discovery pipeline. Ensembl release 70 annotation was first used to build all cDNA transcripts on which RNA-seq reads were mapped to using stampy. Reads aligned with insertions of up to 51nt in length were then scanned to identify insertions aligning to exon-exon boundaries. Subsequently, the inserted sequences were aligned to the intronic sequences separating the corresponding exons. Putative novel micro-exons were then defined as exons which were flanked by canonical splice sites and were supported in at least 15% of all samples. (b) The density of internal exon sizes shows that the majority are distributed around 140nt in length, while there is a sharp decrease in number of exons shorter than 51nt (purple dashed line) as exon size decreases. (c) Previously annotated micro-exons from Ensembl release 70 that show evidence in brain samples (black) compared to novel predicted micro-exons which show evidence in brain samples (white). Although the annotation of internal exons of sizes 22 to 51nt appears to be nearly complete, we identified a large number of novel micro-exons between 6 and 21nt in length. (d) Example of a novel predicted micro-exon. This micro-exon is only 6nt in length and lies within a conserved region of the *CACNA1* gene. The splice sites of this micro-exon are conserved in mammals and Xenopus.

and accurately map RNA-seq reads onto micro-exons. Mapping RNA-seq reads directly onto the genome is known to be computationally difficult owing to a large search space for small exons. We expected reads spanning micro-exons (and thus three or more exons) to further complicate the mapping procedure. We therefore compared the ability of several RNA-seq aligners, including STAR (Dobin et al., 2013), TOPHAT (Kim et al., 2013), and OLEGO (Wu et al., 2013a), to map reads onto micro-exons of decreasing sizes (Materials and Methods). Compared to a micro-exon mapping method we developed (ATMap or Augmented Transcriptome Mapping; see Materials and Methods), both TOPHAT and STAR aligners mapped fewer reads onto short micro-exons, whilst all four methods mapped similar numbers of reads to larger exons (Figure 37). In particular, ATMap mapped more reads to micro-exons of sizes 9 to 21bp (median $\log_2$ fold differences of 0.55–4.00) compared to TOPHAT and STAR (Figure 37). OLEGO's performance was similar to ATMap's (median $\log_2$ fold differences of 0.48–0.92). However, due to its improved performance on short exons, we used ATMap to quantify the usage of novel and previously annotated micro-exons across all 876 RNA-seq samples (Materials and Methods).

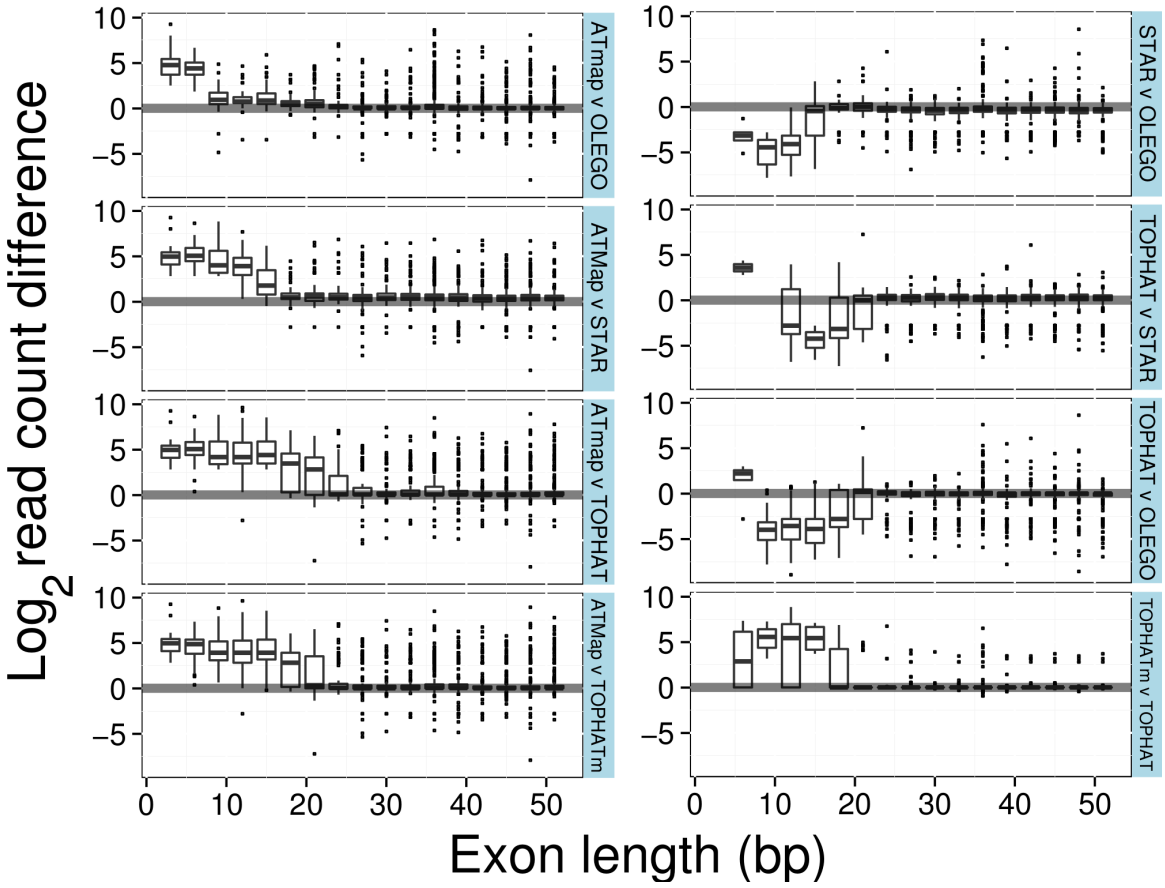

Figure 37: Mapping RNA-seq reads onto micro-exons
The short pair-end reads (76bp) from a human brain sample (SRR112675) were mapped onto the Ensembl (release 70) annotated human genome with ATMap, STAR, TOPHAT, TOPHAT with "micro-exon-search" activated (TOPHATm), and OLEGO. For all exons of sizes 3 to 51bp, we computed the pairwise differences in inclusion predictions and binned them by exon sizes. Only symmetric exons (i.e. exons of sizes that are a multiple of 3) are displayed as they consist of a larger number of cases.

Most micro-exons are well conserved in vertebrates

Sequence conservation has been widely used as a proxy for functionality (Hardison, 2003). We therefore hypothesized that most micro-exons, if functional, should also show evidence for increased sequence conservation relative to neutrally evolving sequences. We found that 60% of all 13,065 micro-exons assessed have a PhastCons score of 0.78 or higher and are thus likely to be functional. This is because while 2.3-10% of the bases in the human genome are thought to be under functional constraint (Chiaromonte et al., 2003;

Meader et al., 2010), less than 5% possess a PhastCons score higher than 0.78 (Siepel et al., 2005). We also found that the 7,315 micro-exons with evidence of inclusion in the brain were far better conserved than the 5,750 Ensembl-annotated micro-exons with no evidence of brain usage. Over 75% of all brain-expressed micro-exons had an average PhastCons score of 0.8 or higher (Figure 38c). In contrast, 40% of Ensembl-annotated micro-exons with no evidence of brain-expression had an average PhastCons score of 0.23 or lower (less than 5% of RefSeq CDS bases score lower than 0.23; Siepel et al. (2005)). This suggests that a considerable number of Ensembl-annotated micro-exons are either annotation errors or annotated products of noisy splicing.

These observations motivated us to focus on the 7,315 brain-expressed micro-exons, which we compared to brain-expressed exons of longer sizes. Exons were also separated by whether they were alternatively spliced (AS) or constitutively spliced (CS) in the brain. They were considered to be AS if their median percent spliced-in (PSI) values were between 5% and 95% in GTEx brain samples and CS if their PSI was larger than 95%. The remaining 5,750 exons with no evidence of brain usage or with lower than 5% PSI are likely not used in the brain and were therefore discarded. Altogether, 5,744 micro-exons were CS and the remaining 1,571 were AS. For comparison, we identified 96,429 and 11,624 longer exons that were CS and AS in the brain, respectively. Both AS and CS micro-exons were highly conserved across vertebrates and possess similar signatures of conservation compared to longer exons (Figure 38a,b and Supplementary Figure 35b).

We also characterised the conservation of exons differing in sizes according to sequence identity. This provided us with a more intuitive comparison

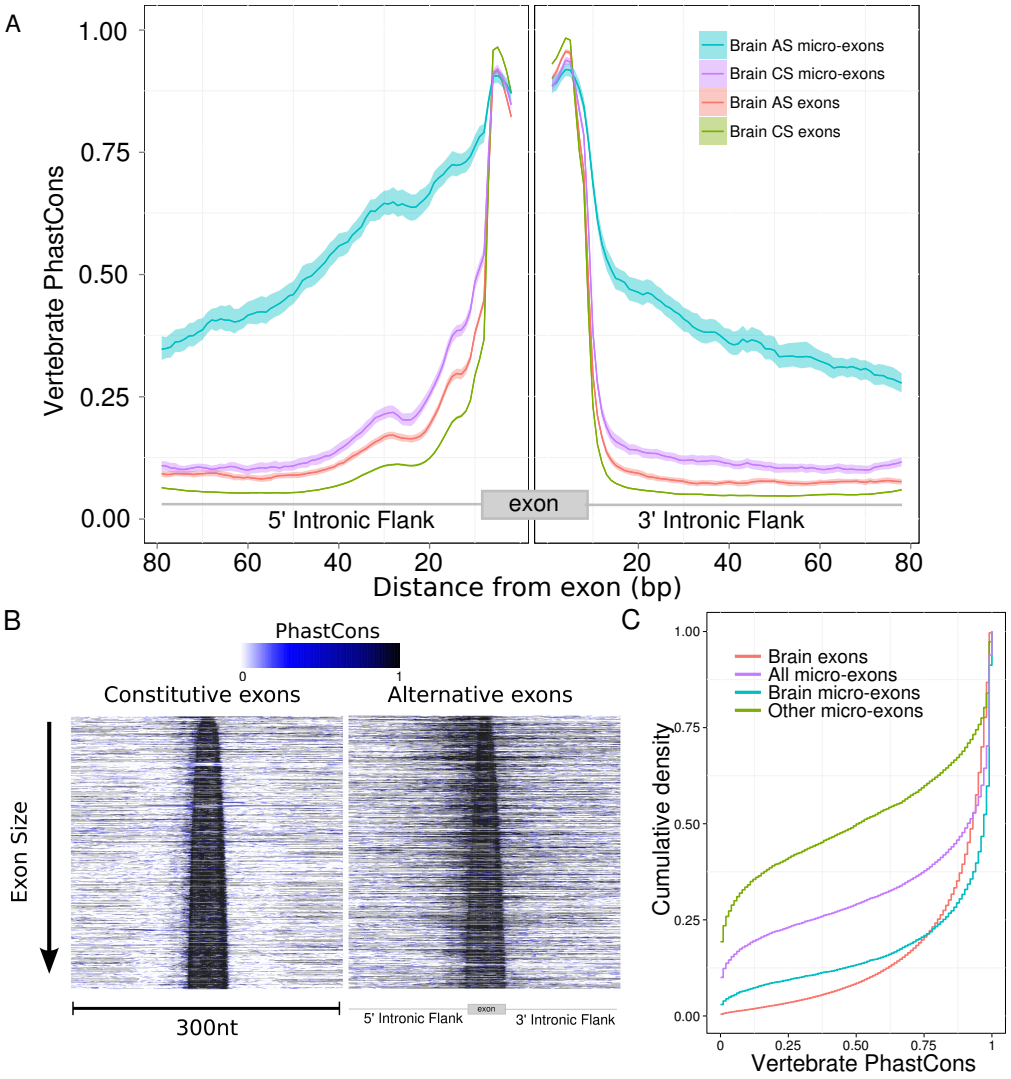

Figure 38: Conservation of micro-exons across vertebrates
(a) Mean vertebrate PhastCons scores and 95% confidence interval in the intronic flanks of symmetric alternatively spliced (AS) micro-exons, constitutively spliced (CS) micro-exons, AS exons, and CS exons. The intronic conservation is significantly higher for AS micro-exons compared to other classes of exons. (b) Vertebrate PhastCons scores of 1,500 randomly chosen CS and AS micro-exons sorted by size and centered within a 300nt window. A large proportion of AS micro-exons show conservation in their intronic flank in addition to strong conservation within exonic sequences. (c) Cumulative density of average PhastCons score of all exons expressed in the brain (red), all annotated micro-exons (purple), all annotated micro-exons with brain usage (teal), and all annotated micro-exons without evidence of brain usage (pale green).

between different exon types than using PhastCons scores. To do this, we computed the pairwise percent identity between human exons and orthologs in four species of increasing divergence (Materials and Methods), altogether

representing over 300 million years of evolution (Pereira and Baker, 2006a). Overall, the distributions of percent identity are shifted upwards for AS micro-exons compared to longer exons and CS exons (Figure 39). This trend also holds true for smaller AS micro-exons (exons less than 25bp) compared to AS micro-exons (exons less than 51bp) and CS micro-exons compared to CS exons in general. Therefore, short AS exons tend to be better conserved than longer exons, possibly owing to stronger functional constraint at the protein level or to a higher density of exonic enhancers compared to longer exons.

The intronic flanks of micro-exons harbour conserved splicing signals

We observed that intronic flanks of symmetric AS micro-exons, i.e. exons of length divisible by 3, were highly conserved, while the flanks of non-symmetric AS micro-exons were conserved at lower levels (Supplementary Figure 40), and in contrast to the ones of other exon classes which showed nearly no vertebrate conservation (PhastCons score < 0.2, Figure 38a). The elevated conservation flanking AS micro-exon extends to over 75bp within each intronic flank, although 5' intronic flanks tend to be better conserved than 3' intronic flanks (Figure 38a).

We therefore hypothesised that intronic flanking AS micro-exons harbour conserved splicing motifs that facilitate their recognition by the splicing machinery. To detect such splicing signal, we aligned human exons and their intronic flanking sequences to the genomes of four other mammalian species: rhesus macaque, cow, dog and mouse (Materials and Methods). Splicing motifs are generally 4 to 10nt in length (Fairbrother et al., 2002). We therefore searched for conserved 6nt motifs that are over-represented near AS micro-exons. To this end, we computed the entropy for each gapless 6nt sliding window (Materials

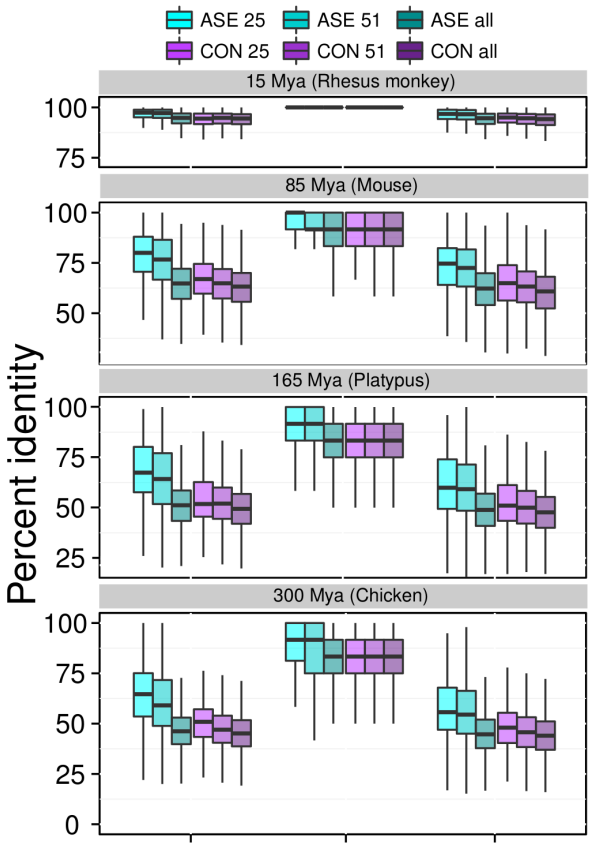

Figure 39: Sequence percent identity of exons and their flanking introns. Percent identity of the pairwise aligned intronic sequences and exonic sequences of different classes of exons between human, and rhesus monkey, mouse, platypus, and chicken. The nucleotide percent identity follows an increasing trend with decreasing exon sizes for both AS and CS exons in intronic flanking regions. However, only AS micro-exon exonic sequences show a difference in percent identity compared to other classes of exons.

and Methods) and considered 6-mers with entropy in the lowest 10-percentile of the empirical distribution to be conserved (entropy $<$ 1.0). In the 5' intronic flank of AS micro-exons, CS micro-exons and all AS exons, several pyrimidine-rich motifs were found to be over-represented (Figure 41a). Only one motif, TGCATG, was found to be highly over-represented in the 3' intronic flanks of AS micro-exons. Notably, this over-representation is absent in the intronic flanks of CS micro-exons and longer exons and is therefore unique to the 3' intronic flanks of AS micro-exons.

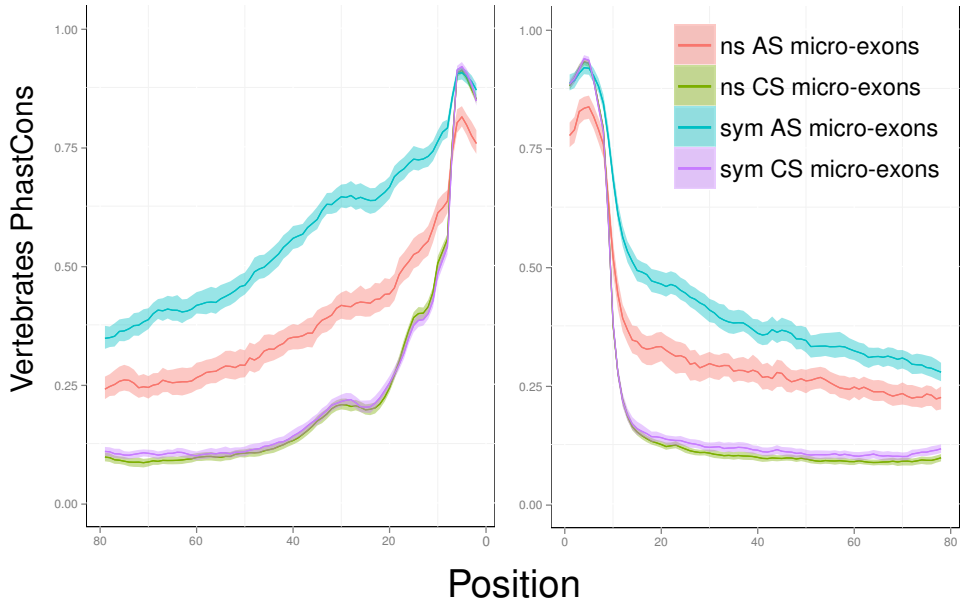

Figure 40: Conservation of symmetric versus non-symmetric micro-exons.

TGCATG is a well characterised motif that is bound by Rbfox splicing factors (Zhang et al., 2008; Lovci et al., 2013). We therefore sought to determine the spatial distribution of the motif in the human genome, without requiring conservation. We observed up to two-fold enrichments in the number of known Rbfox binding motifs in the immediate 3' flanks of AS micro-exons compared to other classes of exons (Figure 41b). This further suggests a role of Rbfox proteins in the splicing of AS micro-exons.

Next we speculated that AS micro-exons are regulated by Rbfox binding events in their intronic flanks. Indeed, we showed an unexpectedly higher density of Rbfox binding events near AS micro-exons, most prominently on their 3' sides. We also showed that this density is even higher for AS micro-exons that are brain-specific (Materials and Methods). To do this, we obtained CLIP-seq replicates for all three Rbfox family members (Rbfox1, Rbfox2, Rbfox3) from mouse brains (Weyn-Vanhentenryck et al., 2014). We projected all exons from

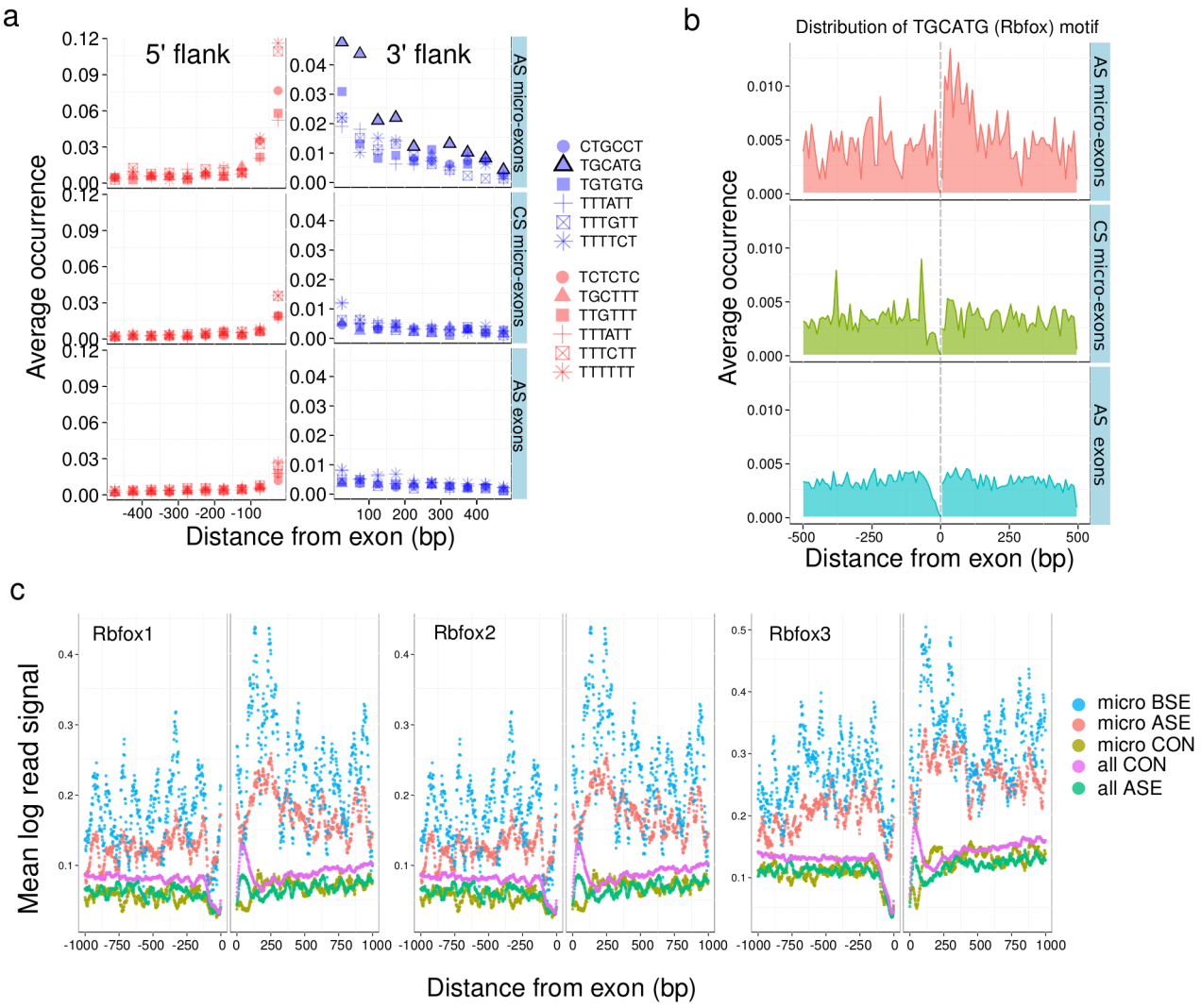

Figure 41: Conserved motifs and splice factor binding sites

(a) Average occurrence of diverse conserved 6-mers in the intronic flanks of exons. Conserved 6-mers were computed according to an entropy threshold based on multiple sequence alignments including human, rhesus monkey, mouse, cow and dog sequences. Of all conserved 6-mers, pyrimidine-rich 6-mers were found to be enriched in the intronic sequences immediately upstream of exons belonging to all classes, with a higher enrichment upstream of alternatively spliced (AS) micro-exons. The 6-mer corresponding to the Rbfox protein family motif (TGCATG) is highly overrepresented in introns downstream of AS micro-exons compared to other classes of exons. (b) In human, the Rbfox binding motif is overrepresented in the intronic sequences downstream of AS micro-exons. (c) Analysis of Rbfox protein CLIP-seq datasets in mouse brains show that AS micro-exons (micro ASE) and AS micro-exons that are brain-specific (mini BSE) possess a higher number of Rbfox binding events in their intronic flanks compared to other types of exons.

the different exon classes to the mouse genome and studied read density near exons from each class (Figure 41c; Materials and Methods). All three Rbfox

family members exhibit the same binding patterns: a higher density of reads in the intronic flanks of 167 brain-specific AS micro-exons (Materials and Methods) and, to a lesser extent, in the intronic flanks of all AS micro-exons compared to other classes of exons. We also observed that the read density is higher in the 3' intronic flanks compared to the 5' intronic flanks of brain-regulated and normal AS micro-exons.

A large number of micro-exons are used exclusively in mammalian brains

We next sought to characterise the tissue and organ usage of micro-exons by quantifying micro-exon inclusion rates in the large array of 876 human RNA-seq datasets described previously. For each sample, the PSI values of annotated micro-exons were computed by mapping the corresponding RNA-seq reads onto a micro-exon augmented transcriptome using ATMap (Materials and Methods). We observed that the splicing ratios of a vast majority of micro-exons were consistent across human tissues (Figure 42a and 42b), which suggests they possess widespread and basic molecular functions. Reassuringly, the inclusion proportions of micro-exons in the brain were replicated in GTEx brain samples, samples from developing and ageing brains (Mazin et al., 2013) and the Human Body Map 2.0 brain samples (Supplementary Figure 43a,b). This excludes potential technical artefact such as batch effects and protocol-specific biases, and suggests a constant usage of most micro-exons across brain development and ageing.

Of the 1,571 alternatively spliced micro-exons, 167 were included in brain-expressed transcripts, but skipped in virtually all other tissue samples (Materials and Methods; Supplementary Figure 43c), suggesting they have brain-specific functions. We next found evidence that brain-specific patterns of micro-exon

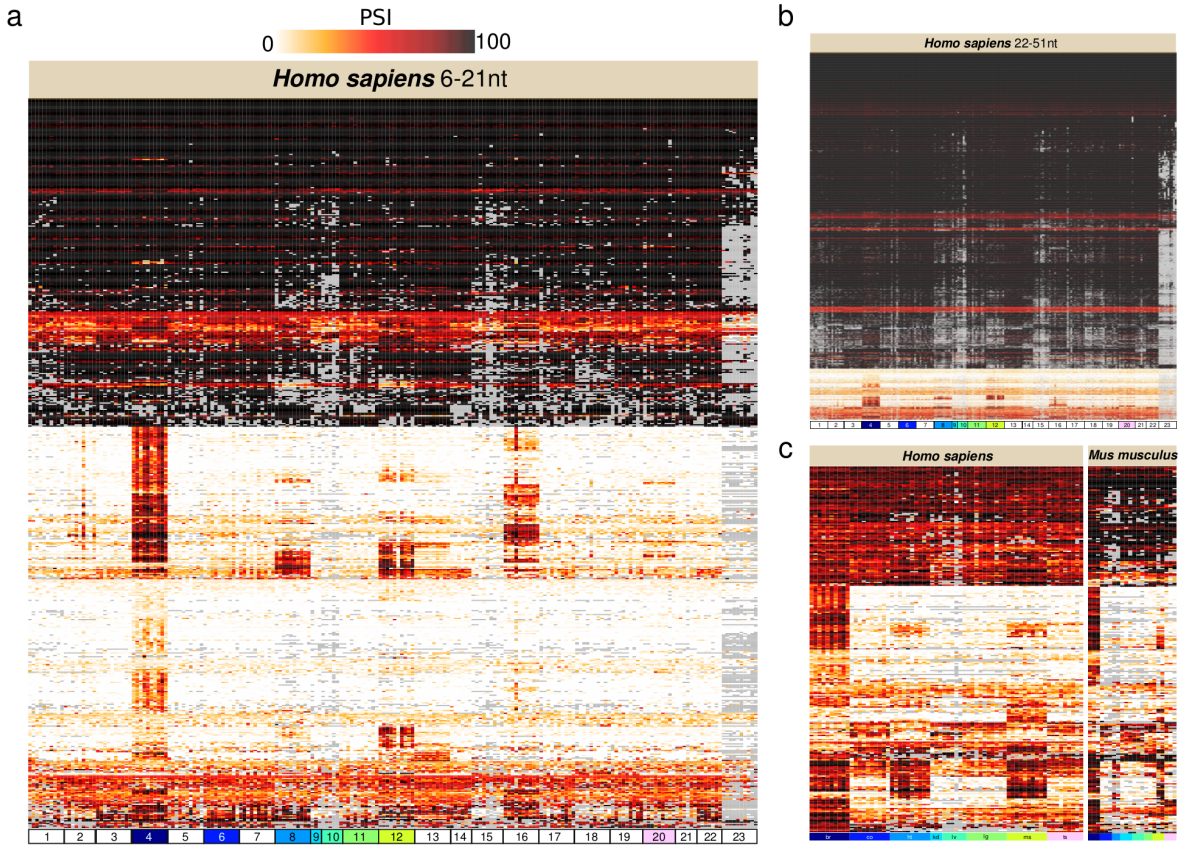

Figure 42: Tissue-dependent inclusion of micro-exons
Inclusion rates of 570 micro-exons of length 6 to 21nt (a), and 5,827 micro-exons of length 22 to 51nt (b) in 23 tissues. 1: adipose, 2: adrenal 3: artery 4: brain 5: breast 6: colon 7: esophagus 8: heart 9: kidney 10: liver 11: lung 12: muscle 13: nerves 14: ovary 15: pancreas 16: pituitary 17: prostate 18: skin 19: stomach 20: testis 21: thyroid 22: uterus 23: whole blood. Coloured samples are compared to matched mouse samples (c). The inclusion rates of 271 human alternatively spliced micro-exons and their mouse orthologs in 8 tissues. Left to right in (c) br: brain, co: colon, ht: heart, kd: kidney, lg: lung lv: liver, ms: muscle, ts: testis. In gray: micro-exons for which PSI could not be computed because of insufficient number of reads ($\leq$5) spanning splice junctions.

usage have been largely preserved across the ~90 million years that separate human and mouse. To ascertain whether the brain-specific usage of micro-exons we observed in human was conserved in mouse or not, we retrieved 24 mouse RNA-seq samples from brain and seven other organs (Merkin et al., 2012) and quantified the inclusion ratio of micro-exons annotated within the mouse genome. We then compared the inclusion ratios of these mouse micro-exons to

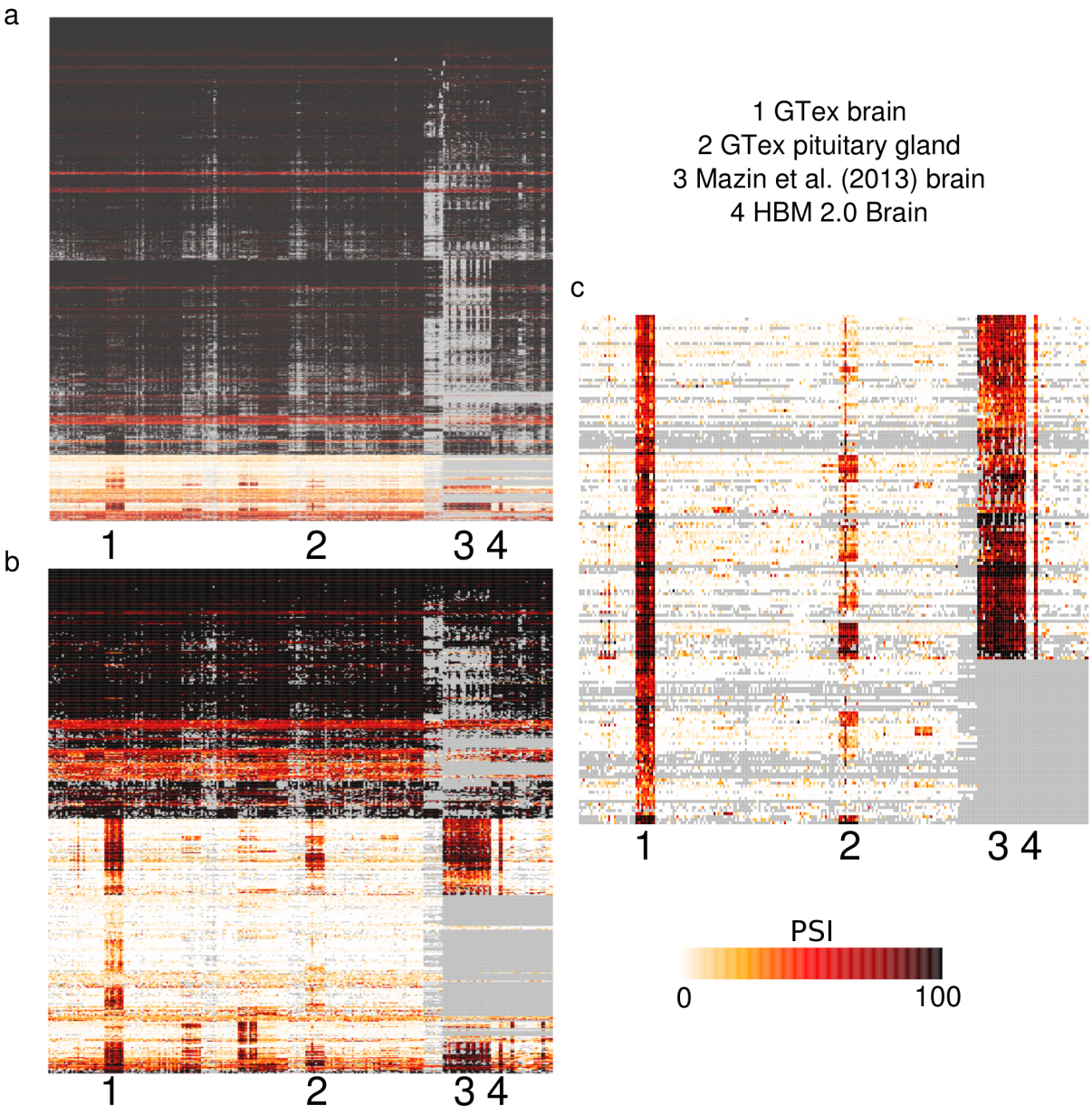

Figure 43: Inclusion of micro-exons in all analysed tissues.
PSI of 22–51nt micro-exons (a) and 6nt–21nt micro-exons (b) in all analysed tissues, including GTEx samples, Human Body Map 2.0 samples, and samples from developing and ageing postmortem human brains (Mazin et al., 2013). (c) Inclusion ratios of 167 AS micro-exons predicted to be brain-specific. In gray: micro-exons for which PSI could not be computed because of insufficient number of reads ($\leq$5) spanning splice junctions.

the inclusion ratios of the orthologous micro-exons in human. Owing to the small number of micro-exons annotated in the mouse genome and the shallower

depth of mouse RNA-seq data, we were only able to quantify the inclusion ratios of 271 micro-exon orthologs (out of 1,571) that are alternatively spliced in human brains (Materials and Methods). Nevertheless, we observed a clear similarity between the micro-exon inclusion patterns of human and mouse brains (Figure 42c). Numerically, micro-exon usage are highly similar among human brain samples (Pearson correlation = 0.72–0.88; Supplementary Figure 44), but also between human and mouse brain samples (Pearson correlation = 0.58–0.75). In contrast, the correlations of micro-exon usage between human brain samples and samples from other human tissues, or the ones between mouse brain samples and samples from other mouse tissues are significantly lower (0.19–0.59 and 0.27–0.49, respectively).

To identify potential functions of AS micro-exons in the brain, we searched for over-represented gene ontologies (GO) in the set of genes possessing micro-exons that are AS or CS in the brain compared to the set of all brain-expressed genes. We found that genes that possess AS micro-exons were preferentially involved in axon guidance and neuron migration (Table 6) even after correcting for multiple testing (p-values 1.0×10^{-10} and 3.8×10^{-4} respectively). These findings further support the importance of AS micro-exon for increasing transcriptome diversity in the brain and, in particular, during brain development.

Constitutively and alternatively spliced micro-exons possess distinct genomic attributes

We next considered whether micro-exons require greater precision by the splicing machinery, as determined by RNA splicing signal motifs. 5' and 3' splice sites for CS micro-exons (but not longer exons) were found to have the greatest signal strength as measured by MAXENT (Yeo and Burge, 2004)

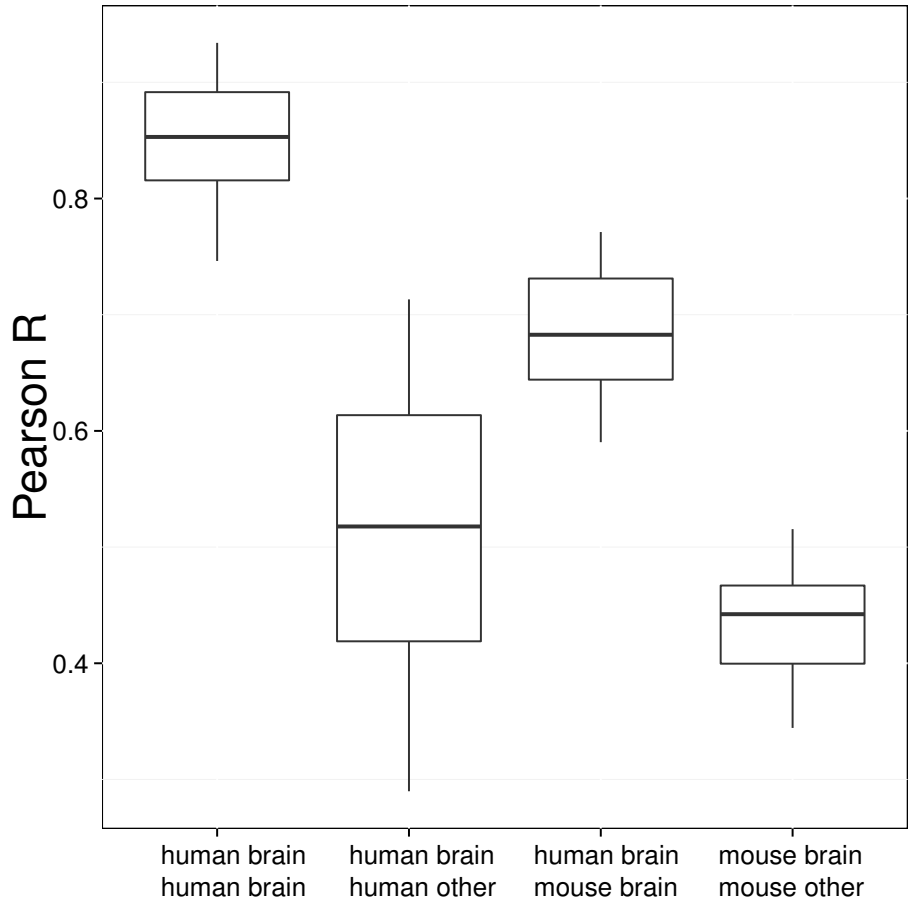

Figure 44: Pearson correlation of micro-exon usage levels
Pearson correlation of micro-exon PSI among human brain samples, between human brain and other human samples, between human and mouse brain samples and between mouse brain and other mouse samples.

score (5': median 9.21 vs 8.60–8.73, $p < 6.4 \times 10^{-36}$; 3': median 9.49 vs 7.83–8.81, $p < 6.9 \times 10^{-56}$ Mann-Whitney test; Figure 45a). Introns flanking AS micro-exons tend to be shorter than longer AS exons (median 1,418 vs 1,852 and 1,434nt vs 1,770nt for 5' and 3' introns, respectively; Figure 45b). We also observed an increased thymine content 10–100bp upstream of AS micro-exons compared to other classes of exons (median 34.1% vs 29.5–31.1%; Supplementary Figure 45c). The higher thymine content can be explained by elongated poly-pyrimidine tracts immediately upstream of AS micro-exons that guide the splicing machinery to the AS micro-exons.

GO.ID	Term	Annotated	Significant	Expected	elimFisher
GO:0007411	axon guidance	162	71	29.9	3.7×10^{-14}
GO:0001764	neuron migration	45	24	8.3	1.4×10^{-7}
GO:0048814	regulation of dendrite morphogenesis	21	12	3.9	8.3×10^{-5}
GO:0048011	nerve growth factor receptor	131	41	24.2	0.00023
GO:0001952	regulation of cell-matrix adhesion	23	12	4.2	0.00027
GO:0016044	cellular membrane organization	214	66	39.5	0.00030
GO:0019886	antigen processing and presentation	44	18	8.1	0.00043
GO:0006929	substrate-dependent cell migration	10	7	1.8	0.00051
GO:0006468	protein phosphorylation	538	128	99.3	0.00067
GO:0008089	anterograde axon cargo transport	13	8	2.4	0.00070

Table 6: Gene ontology analysis.
Top 10 over-represented gene ontologies in genes possessing micro-exons that are alternatively spliced in the brain. The elimFisher p-value statistic is uncorrected. Using the stringent Bonferroni correction, the two most significant GO terms were still significant (2756 GO terms were tested).

The differences in attributes between exons of different sizes, independent of their splicing patterns, suggest that cellular mechanisms controlling splicing of micro-exons may be different from those regulating general exon splicing.

Alternative inclusion of micro-exons can alter protein-protein interactions

A widely believed role of alternatively spliced exons is to contribute to protein diversity (Black, 2000; Romero et al., 2006). We therefore sought to characterise the impact of micro-exon inclusion or exclusion on protein structure. We first attempted to quantify the proportion of micro-exons that have coding potential. Overall, we found that at least 93.8% (5416 out of 5744) and 78.7% (1236 out of 1571) of CS and AS micro-exons, respectively, have the potential to encode for amino acids, i.e. they do not introduce in-frame stop codons (Materials and Methods). AS micro-exon sequence contained an unusually high amount of intrinsically disordered regions[1] (54.0–60.2%, 95% confidence interval) compared to CS micro-exons (45.6–48.4%) and longer exons (32.5-41.2%; Materials and Methods; Figure 47a). We also observed

[1]As predicted by DISOPRED, see Materials and Methods.

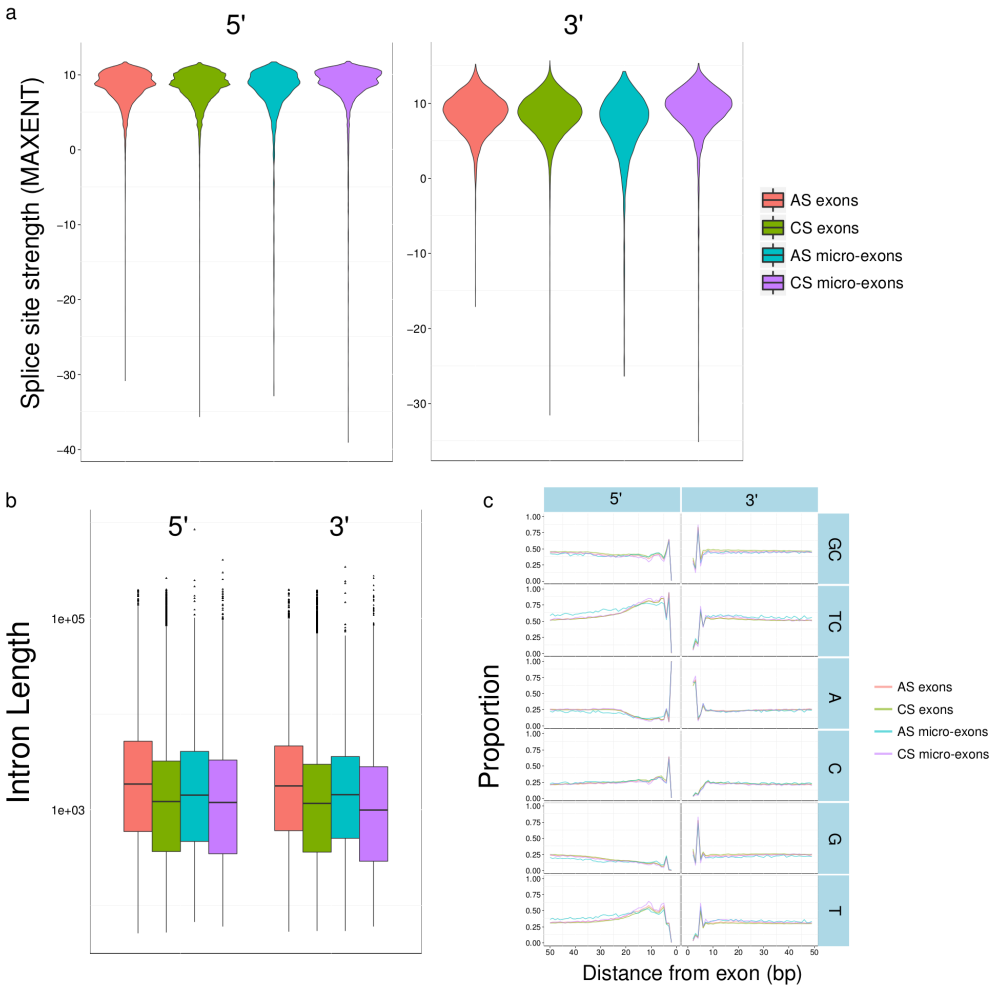

Figure 45: Attributes of different classes of exons.
(a) Strength of 5' and 3' splice sites according to MAXENT scores in alternatively spliced, constitutively spliced exons and in alternatively spliced and constitutively spliced micro-exons. (b) length distribution of intron lengths flanking different classes of exons. (c) GC content, TC content, Adenine composition, Cytosine composition, Guanine composition and Thymine composition in intronic flanks of different classes of exons.

that residues encoded within AS micro-exons were slightly but significantly over-represented within unstructured sequence compared to other classes of exons (61.7–62.0% vs 52.6–57.5%; Supplementary Figure 46). Since alternative splicing of disordered regions are known to rewire interaction networks in a tissue-specific manner (Romero et al., 2006; Buljan et al., 2012), an important role of AS micro-exons might be to alter protein-protein interactions.

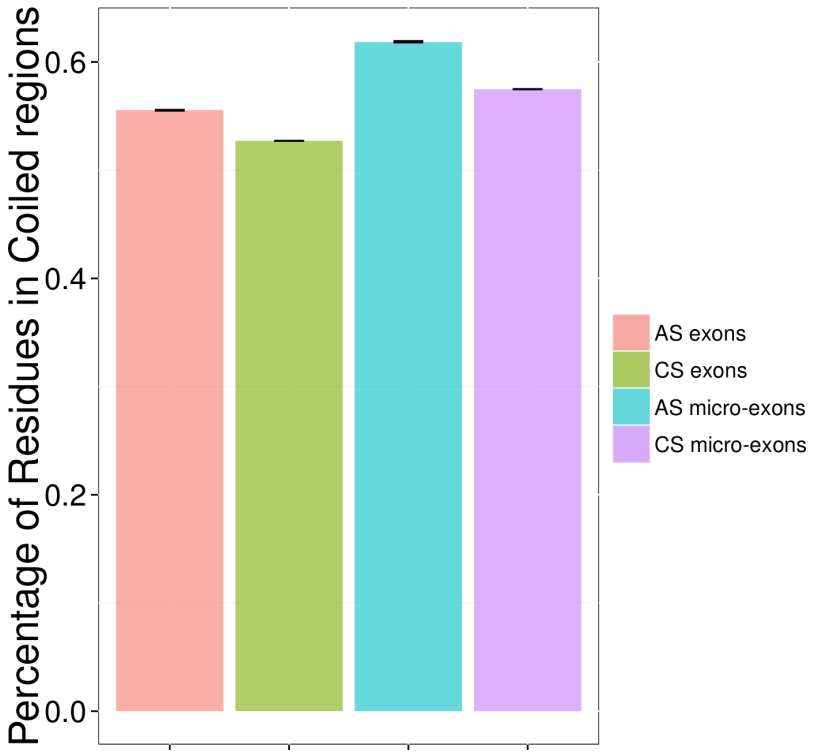

Figure 46: Proportion of residues in coiled coils.
Alpha helices and beta strands in different classes of exons were also estimated using PSIPRED.

To explore whether micro-exons within intrinsically disordered regions could alter protein-protein interactions, we searched for micro-exons that encode residues from domains known to interact with other proteins. Among genes with conserved AS micro-exons with coding potential, we identified a paralogous gene family consisting of three amyloid binding proteins: APBB1, APBB2, and APBB3. All three APBB (amyloid-beta (A4) precursor protein-binding, family B) genes possess 6nt micro-exons encoding two amino acids located within the first of two phosphotyrosine-binding (PTB) domains. Searching cDNA databases revealed that the micro-exons within APBB are also present and alternatively spliced in fishes, chicken and diverse mammals (Figure 47b). These micro-exons are located at paralogous positions within APBB genes, indicating that these micro-exons survived over 400My of evolution since these

genes' duplications in early vertebrate evolution (Figure 47b).

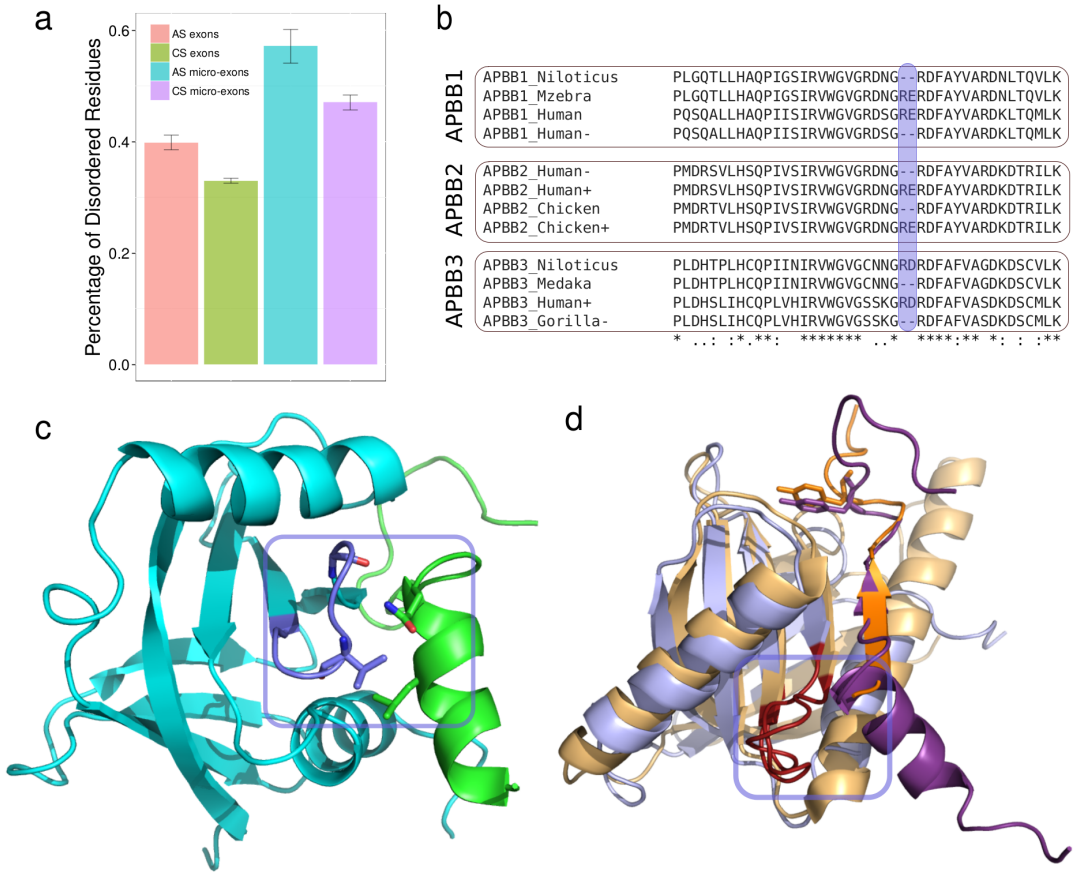

Figure 47: Alternatively spliced micro-exons and protein structure.
(a) Alternatively spliced micro-exons tend to fall within intrinsically disordered regions more than other classes of exons (bar represent 95% confidence intervals). (b) Translated cDNA sequences from proteins belonging to the APBB protein family. A 6nt micro-exon exists at a paralogous position in all three APBB proteins. Furthermore, all three proteins possess two isoforms, one with the micro-exon included and one without, in both mammals (human, gorilla), reptiles (chicken), and fish (Niloticus: *O. niloticus*, Mzebra: *M. zebra*). (c) 3D structure of the phosphotyrosine binding domain of APBB2 (cyan) in complex with the interacting amyloid-beta protein (green). The close proximity of the cytoplasmic tail of amyloid-beta protein to the loop containing the two amino acid residues in complex suggest they interact. (d) Superposition of the phosphotyrosine binding domain of APBB2 (blue) and of LDLR (orange). After TNS4 was mapped onto the structure of LDLR, we found that both APBB2 and TNS4 possess a micro-exon in homologous loops (red). The inclusion or exclusion of these amino acid residues are therefore expected to both alter the interactions of APBB2 and TNS4 with amyloid-beta protein and beta integrin, respectively.

To investigate the functional impact of these micro-exons on protein structure, we mapped the two amino acid residues encoded by the micro-exon onto the

known 3D structure of APPB2 (Figure 47c). This mapping showed that the two residues are located in a beta-turn loop of APPB2 which we predict to interact with the cytoplasmic tail of amyloid-beta protein from their close proximity in the complex (Figure 47c). Interestingly, all 4 protein members from the tensin (TNS) family also possess micro-exons (Supplementary Figure 48) which encode residues within their phosphotyrosine-binding domains. By mapping the TNS4 sequence onto a homology model (Materials and Methods), we found that tensin micro-exons are mapped to the same beta-turn as are the amino acids encoded by micro-exons of the APBB protein family (Figure 47d). This points towards a highly conserved mechanisms that control protein-protein interactions through the alternative inclusion of micro-exons.

```
TNS1 QRKLFFRRHYPLNTVTFCDLDPQERK WMKTEGGAPA KLFGFVARKQGSTTDNACHLFAELDPNQPA
TNS2 QRKLFFRRHYPVNSITFSSTDPQDRR WTNP-DGTTS KIFGFVAKKPGSPWENVCHLFAELDPDQPA
TNS3 QRKLFFRRHYPVNSVIFCALDPQDRK WIK--DGPSS KVFGFVARKQGSATDNVCHLFAEHDPEQPA
TNS4 QRKVFFRRHYPLTTLRFCGMDPEQRK WQK--YCKPS WIFGFVAKSQTEPQENVCHLFAEYDMVQPA
     ***:*******:.:: *.  **::*. * :     .: .:*****..  .. :*.****** *  ***
```

Figure 48: Tensin alignments
Regional alignment of human proteins belonging to the tensin family. Residues encoded by micro-exons are highlighted in red.

5.3 Discussion

About 6% of all annotated human internal exons are 51nt or shorter (micro-exons) and fewer than 0.5% are 21nt or shorter. Our aim was to study micro-exons at the genome-wide level because to the best of our knowledge no genome-wide functional study of micro-exon exists. Therefore, it has remained unknown whether micro-exons are generally functional, the product of noisy splicing (Pickrell et al., 2010b), or mostly annotation artefacts. Here, we presented data suggesting that over 60% of annotated micro-exons are well-conserved in vertebrates and are thus expected to be functional. An even higher percentage ($> 80\%$) of

micro-exons with evidence of usage in the human brain are well-conserved and expected to be functional. AS micro-exons were found to be different from other classes of exons in several ways. They possess shorter flanking introns, elongated poly-pyrimidine tract, and a higher number of conserved splicing motifs in their flank compared to other types of exons. These features likely help the splicing machinery in efficiently recognizing and processing AS micro-exons in a tissue-dependent manner.

Analysis of deep transcriptomic data reveals that expressed micro-exons are functional

By analysing 20,167,964,776 reads from 345 brain samples (Lonsdale et al., 2013), we were able to identify 7,315 micro-exons included in brain transcripts with a PSI 5% or higher. Of these, 255 micro-exons were novel, which suggests that our current annotation was appropriate for a genome-wide study of micro-exons. To estimate the proportion of functional micro-exons, we searched for signatures of conservation in the 7,315 micro-exons expressed in the brain and their flanking sequences, and observed that the vast majority of both AS and CS micro-exons are highly conserved according to both vertebrate PhastCons score and pairwise alignment percent identity, which strongly supports their functional potential.

We found that the intronic sequences flanking symmetric AS micro-exons tend to be conserved, which suggest that they too play a functional role. For all pairwise alignments assessed, both exonic and flanking sequences of AS micro-exons were more highly conserved than longer AS exons. This further indicates that AS micro-exons are functional and possess constrained sequences in their flanks that likely serve as splicing regulators in a spatial or temporal context.

A large number of alternatively spliced micro-exons is involved in brain function

There are 7,315 micro-exons with evidence of expression in the brain, 1,571 (21.5%) of which are alternatively spliced. The proportion of AS micro-exon increases to 48.5% when restricting micro-exons to 21bp or smaller. As many as 167 micro-exons were predicted to be included in transcripts expressed in the brain, but not in transcripts expressed in other tissues. The brain-specific inclusion of these micro-exons suggests a functional role which is exclusive to the brain. By comparing the inclusion rates of micro-exons annotated in the mouse genome across several organs to the ones of orthologous human micro-exons, we observed a strong conservation of brain-specific splicing patterns which further supports their involvement in mammalian brain function (Brawand et al., 2011).

To further explore the role of AS micro-exons, we searched for conserved motifs enriched in the flanks of AS micro-exons. We found conserved sequences in the 3' flanking regions of AS micro-exons to be enriched in Rbfox binding motifs, a signature that was absent for other classes of exons. Surveying the flanks of exons in human revealed a similar enrichment in Rbfox motifs in the 3' intronic sequences of AS micro-exons. Furthermore, by analysing Rbfox1, Rbfox2 and Rbfox3 CLIP-seq data from mouse brains, we confirmed that all three Rbfox proteins bind with higher affinity near AS micro-exons compared to other exons, and that they bind with even higher affinity near AS micro-exons with brain-specific usage patterns.

These findings are interesting for several reasons. Firstly, Rbfox proteins are well known splicing factors that play crucial roles in both brain development and function (Gehman et al., 2011, 2012; Weyn-Vanhentenryck et al., 2014).

The enrichment of conserved Rbfox motifs near AS micro-exons therefore strengthens our conjecture that micro-exons are involved in brain function. Secondly, earlier studies on Rbfox reported increased inclusion when Rbfox bound downstream to an exon, while the opposite effect was observed when Rbfox bound upstream. The increased number of Rbfox binding motifs in the 3' flanking regions compared to the 5' flanking regions of AS micro-exons is therefore expected to enhance the inclusion of these micro-exons, possibly by recruiting other factors involved in splicing. Lastly, the conservation of Rbfox motifs near AS micro-exons may explain how their brain-specific usage patterns have been preserved between human and mouse, and possibly in all mammalian lineages.

Our gene ontology analysis revealed that genes possessing AS micro-exons tend to be involved in axon guidance and neuron migration. Although these functions are somewhat expected from genes expressed in the brain, we must note here that the enriched gene ontologies were obtained by comparing our set of genes harbouring AS micro-exons to a set of brain-expressed genes. It is therefore likely that AS micro-exons increases the transcriptome (and proteome) diversity that are particularly important for these two processes.

Alternatively spliced micro-exons can alter protein structure and protein-protein interactions

Lastly, we investigated the role of micro-exons on protein structure. Because each micro-exon only encodes a small number of amino acid residues, we wished to understand if and how they could affect the structure of a protein. Our hypothesis was that alternatively spliced micro-exons can generate multiple protein isoforms with distinct functions. We estimated that over 78.7% of AS

micro-exons have coding potential. This estimate is conservative as it is based on previously annotated upstream and downstream coding exons. However, it is possible that several AS micro-exons function by means other than encoding amino acid residues, for instance by regulating gene expression through nonsense mediated decay (Weischenfeldt et al., 2012). The strong bias in symmetric AS micro-exons however suggest that a major role of AS micro-exons is to increase protein and, possibly, functional diversity.

By analysing micro-exons that fall within protein domains, we identified several proteins with micro-exons in their phosphotyrosine binding domains. For example, all three members of the APBB gene family possess an alternatively spliced 6nt micro-exon in a beta-turn loop which lies in close proximity to the amyloid-beta protein in complex. Furthermore, all four members of the tensin (TNS) family also possess a micro-exon encoding 8–10 residues in a beta-turn loop homologous to the one habouring the two residues encoded by micro-exons in proteins of the APBB family. The inclusion or exclusion of these micro-exons lengthens or shortens the beta-turn loop, and is therefore likely to alter protein-protein interactions: APBB with amyloid-beta protein and TNS with beta integrin. Although these are only a few examples, our genome-wide predictions of micro-exons location indicate that a higher proportion of AS micro-exons are found in intrinsically disordered regions compared to other types of exons. These disordered regions are well known to play important roles in protein-protein interactions (Romero et al., 2006; Vavouri et al., 2009; Babu et al., 2011), thus we hypothesise that many additional AS micro-exons encode residues whose inclusion or exclusion can alter protein-protein interactions. We also hypothesise that the use of micro-exons to regulate protein-protein interactions helps the protein maintain its conformation by slightly altering the local structure of a protein. In

contrast, the inclusion or exclusion of longer AS exons are expected in certain cases to significantly alter protein structure.

Concluding remarks

Overall, our analyses predict that thousands of micro-exons are functional in mammals and reveal at least 167 that are likely to have brain-specific functions. This study highlights the importance of alternative splicing to protein diversity and provides specific examples of alternatively spliced micro-exons that may alter protein-protein interactions. Despite their short sizes, AS micro-exons likely possess a large number of regulatory regions in their intronic flanks. This represents a large non-coding mutational target which can cause aberrant regulation of important AS micro-exons and subsequently lead to disease. Aberrant splicing induced by Rbfox dysregulation is associated with a variety of brain-related disorders including autism, mental retardation and epilepsy (Gehman et al., 2012; Weyn-Vanhentenryck et al., 2014). Therefore, it would be particularly interesting to study how aberrant splicing of AS micro-exons can affect protein-protein interactions during brain development and impact brain-related diseases.

5.4 Materials and Methods

Dataset retrieval

Overall, this study uses transcriptome data in the form of 76bp paired-end RNA-seq reads from 345 post-mortem prefrontal cortices, 74 muscle samples, 47 nerve samples from human, and up to 10 samples for other tissues within the GTEx (Lonsdale et al., 2013), 76bp paired-end from 25 post-mortem human brain across development and ageing (Mazin et al., 2013) and from a diverse human organs from the Illumina Human Body Map 2.0 (2011); mouse transcriptome

data was obtained from Merkin et al. (2012); Rbfox CLIP-seq data were obtained from mouse brains (Weyn-Vanhentenryck et al., 2014).

Discovery and mapping of splicing events

To identify novel micro-exons from RNA-seq data, 307 brain samples, 74 muscle and 47 nerve samples from GTEx were mapped onto Ensembl (release 70) cDNA transcripts using stampy allowing for multiple mapping locations (options *–xa-max=5 -t4 -v3*). In the filtering step, only reads mapping with an insertion of size 3 to 51nt which are flanked by at least 6nt matches on both sides were kept. Subsequently, insertions overlapping exon-exon boundaries were retrieved, and those supported by fewer than 10 reads were discarded. Introns separating exon-exon boundaries were then scanned for the inserted sequences (Figure 49). Sequences flanked by the canonical splice sites were then considered to be putative micro-exons. In case of ambiguous mapping, a location at random is chosen to represent the putative micro-exon. Of 7,575 inserted sequences, only four inserted sequences 6nt or longer were found to map to ambiguous sites ($< 0.1\%$), but 297 sequences of 3nt in length were found to be ambiguous (21.7%). We further required micro-exons to be expressed in at least 15% of all samples coming from each tissue (i.e. each micro-exon must have at least 5% PSI in 45 of 335 brain samples) in order to be considered to be novel. To allow comparison between novel predicted micro-exons and previously annotated micro-exons, we also required previously annotated micro-exons to have the same expression breadth as novel micro-exons.

An in-house pipeline was developed in order to quantify micro-exon usage. A micro-exon alternatively splicing augmented transcriptome was created in the following way: (1) identify all micro-exons annotated within an annotation file,

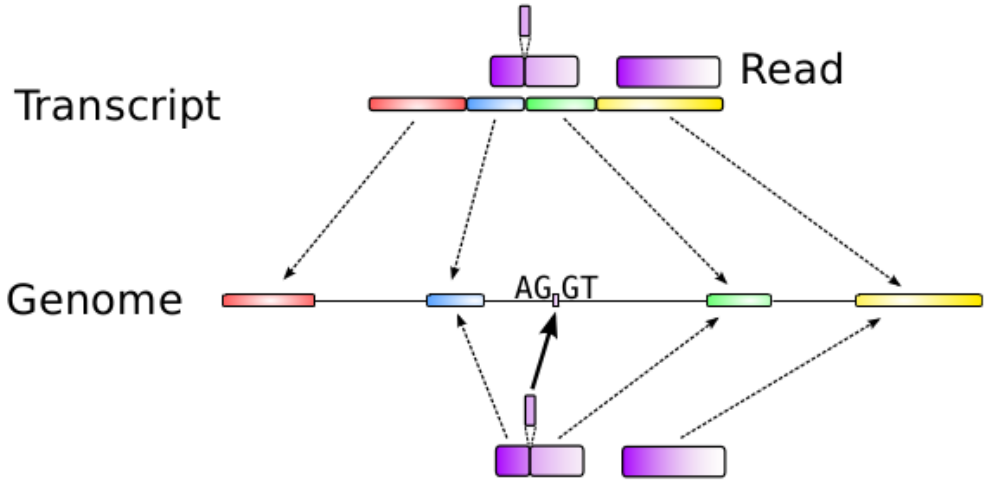

Figure 49: Discovery of novel micro-exons.
Inserted sequences within mapped reads are aligned to intronic sequences if they coincide with exon bondaries. We required canonical splice sites to flank the inserted sequences within the introns.

(2) for each transcript, construct a version with the micro-exon(s) included and one without, (3) in cases where there are multiple micro-exons in the same gene, construct transcripts representing all possible combination of micro-exons inclusion/exclusion for micro-exons that are 100nt within one another. Step (3) is important for quantifying alternatively spliced micro-exons located in tandem (several collagen genes harbour multiple micro-exons). RNA-seq datasets were then mapped to this augmented transcriptome using bwa (single-end; options *samse -n 100*), allowing at most 2 mismatches per read. The quantities R_L and R_R, representing the number of reads supporting the left and right junction respectively, can then be computed by counting the number of reads that span each junction. According to these quantities, the number of reads supporting each micro-exon, R_{tot}, can then be computed using the following equation:

$$R_{tot} = 2 \cdot \min\{R_L, R_R\} \tag{4}$$

In this case, taking the minimum of these two quantities avoids cases in which an alternative 5' or 3' splice site biases the estimated micro-exon usage.

Using R_{tot}, the percent spliced-in statistic can be computed for each micro-exon: $\frac{R_{tot}}{R_{tot} + R_{skipped}}$, where $R_{skipped}$ represents the number of reads supporting an exon skipping event.

To compare our in-house pipeline to STAR (Dobin et al., 2013), TOPHAT (Kim et al., 2013) and OLEGO (Wu et al., 2013a), paired-end reads from a 76nt post-mortem human brain sample (SRR112675; Mazin et al. (2013)) were mapped using STAR with standard options and maximum two mismatches (*-M 2*), TOPHAT with both standard options and *micro-exon-search*, and OLEGO with standard options. The numbers of reads supporting micro-exons were then computed according to the equation above.

Identifying brain-specific micro-exons

Brain-specific micro-exons were defined to be micro-exons that have a median PSI of at least 25 in GTEx brain samples, and an 80-percentile of at most 10 PSI in other samples, excluding samples from pituitary gland (due to their relatedness to brain). 167 brain-specific micro-exons were identified using these thresholds, but between 120–200 using different cut-offs.

Finding a set of constitutive and alternatively spliced exon

To identify a set of control exons to allow comparison with micro-exons analysed in this study, 100 samples from the GTEx brain were randomly chosen and analysed. Spliced reads were recovered from each of the samples and the PSI of each internal exon was computed in the same way as for internal micro-exons. Exons with median PSI at least 10% and at most 95% were classified as alternatively spliced, while exons with PSI higher than 95% were classified as constitutively spliced.

Conservation of micro-exons

To characterise the sequence conservation of micro-exons and their intronic flanks, micro-exons were centred within 300nt windows and PhastCons scores (from UCSC genome browser, version hg19) were retrieved for each position. PhastCons confidence intervals for different classes of exons were then computed by bootstrapping as such: (1) Let S denote a set containing vectors of size 300, each representing the PhastCons score of one exon at each of the 300 nucleotide positions, (2) draw randomly with replacement $|S|$ vectors from this set, (3) compute the average conservation profile $\bar{S}_i$, (4) repeat this process 1000 times, obtaining $\{\overline{S_i} : 1 \leq i \leq 1000\}$. The lower and upper values of the confidence interval for each position $1 \leq k \leq 300$ correspond to the the 5- and 95-quantiles of $\{\overline{S_i}(k) : 1 \leq i \leq 1000\}$, respectively.

To compute the percent identity between human brain-expressed exons and orthologs in dogs, mouse, platypus and chicken, the internal human exons and their intronic flanks were aligned to corresponding orthologous sequences. The percent identity between human and the other species were then computed for the 5' and 3' flanks of each exon. Additionally, the percent identity of 12nt from the exon (6nt from both the 5' end and 3' end) were also computed to remove biases from differing exon lengths.

Identification of conserved and over-represented k-mers

A diagram illustrating our approach can be seen in Figure 50. In order to identify conserved k-mers, all regions consisting of human internal exons plus 500nt intronic flanking sequences (250nt from each flanking intron) were collected. These regions were projected onto the genomes of rhesus monkey, cow, mouse and dog using liftOver, and were then aligned with muscle (Edgar, 2004).

Subsequently, the resulting multiple sequence alignments were divided into three regions: 5' intronic sequences, exonic sequences, and 3' intronic sequences. A sliding window of size 6nt was then used to scan for conserved regions, which correspond to gapless 6nt alignments with entropy scores over 1.0). This entropy threshold corresponds to the 10% most conserved 6-mers. Different entropy thresholds lead to very similar results.

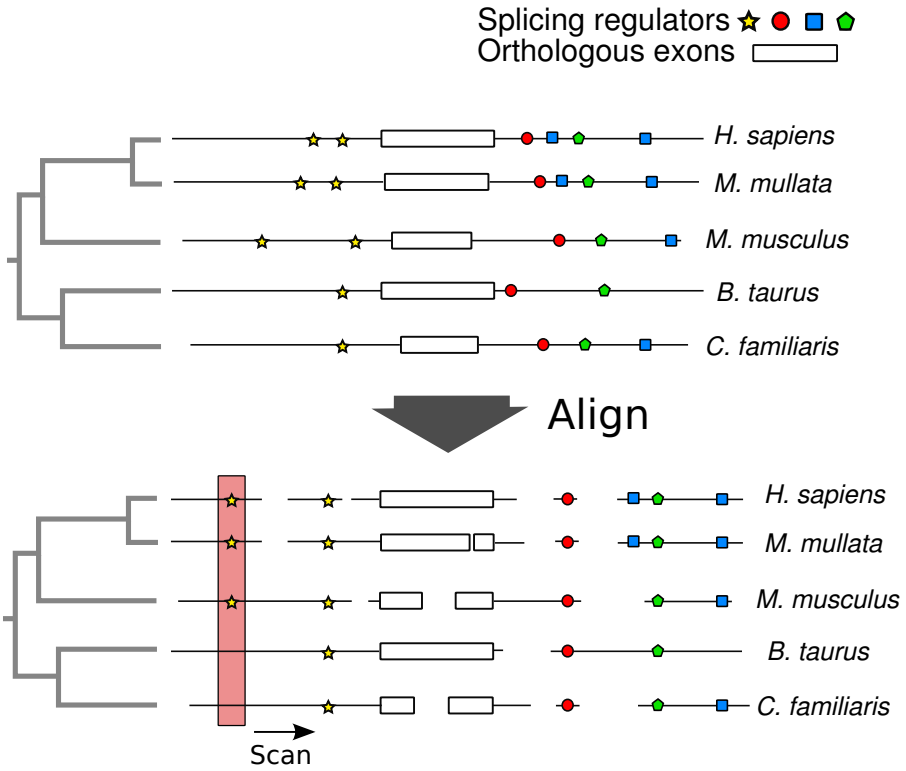

Figure 50: Identification of conserved k-mers.
Diagram representing our approach to identify conserved intronic k-mers among mammals.

Micro-exons, protein domains and structures

All 8,891 protein sequences (one per gene) containing a brain-expressed internal exon studied here were collected and disordered regions and secondary structures were predicted using DISOPRED and PSIPRED, respectively. The disorder and secondary structures (Coil, Helix or Strand) of 8,308 proteins were

successfully predicted at a residue-level resolution, and were used to compute the proportion of disordered and coiled residues in the different classes of exons.

To identify micro-exons overlapping protein domains, HMMSCAN (Eddy, 2010) was used with pfam-A hmm database. The protein structure of APBB2 was retrieved from PDB with ID 2ROZ. To identify a structure homologous to TNS4, HHPred (Soding et al., 2005) was used with TNS4 sequence as input. The structure of LDLH (3SO6) was then found to be appropriate as a proxy for TNS4 structure. Micro-exons within each protein were then mapped onto their corresponding structure using pyMol.

6 Chapter 6: Discussion

There are two common computational angles from which geneticists approach the question of how genetic variation influences phenotypic traits. One approach (comparative genomics) is to identify functional changes in the genomes of different species, while the other (population genomics) uses standing genetic variation within species to study phenotypic variation. Although both approaches share a common aim, population genomics has recently gained widespread interest owing to its success in detecting genome-wide functional sequence variation (mostly in yeast, flies, mouse and human). Indeed, population genomics studies have now identified thousands of variable loci that affect molecular phenotypes such as RNA expression levels (Pickrell et al., 2010a; Montgomery et al., 2010; Lappalainen et al., 2013), protein expression levels (Wu et al., 2013b) and transcription factor binding affinity (McVicker et al., 2013). Analyses in human populations also revealed hundreds of loci which underwent positive selection over the course of evolution, and these likely correspond to loci that affect phenotype (Grossman et al., 2013).

Despite these successes, population analyses are limited in scope to only the most recent evolutionary history of a species. Moreover, it is often difficult to distinguish positive from negative selection because they leave similar signatures on linked genetic loci (a reduction of genetic diversity). Consequently, comparative genomics can complement population genomics and researchers have begun to investigate possibilities of combining the two for greater power (Lawrie and Petrov, 2014). However, comparative genomics currently remains the only practical computational approach to study phenotypic innovations that occurred in the distant past. Comparative genomics allows us to infer properties of ancient genomes and to study the evolutionary forces that transformed them

into the contemporary genomes that we now observe.

The work presented in this thesis focuses on using and developing comparative genomics approaches to study genome evolution using high-throughput data. In particular, my aim has been to identify and to understand how evolutionary forces shaped ancient genomes by comparing large genomics datasets from various extant species. In this thesis, I described analyses of diverse evolutionary mechanisms including gene duplications, positive selection, incomplete lineage sorting, and gene architecture evolution. The ensemble of my work therefore provides us with a broad understanding of genome evolution.

In this last chapter, I discuss briefly my work and possible future directions.

6.1 Next-generation sequencing technologies are changing the way we study biology

The advent of high-throughput sequencing and constant technological advances encourage us to study biology in a new light. For instance, high-throughput RNA sequencing has become a great complement to genome sequencing projects in two important ways. Firstly, the use of RNA sequencing significantly increases the quality of gene annotations and secondly, it allows genomicists to study the transcriptomes of non-model organisms in a wide variety of context. Other high-throughput assays can also now be used to explore specific aspects of non-model organisms. To give but a couple of examples, RAD-seq is now commonly used to estimate genetic diversity in non-model populations (Davey et al., 2013; Wagner et al., 2013) and ChIP-seq can be used to identify genomic regions bound by selected transcription factors (Ricardi et al., 2014). Understanding the strengths and limitations of each method therefore al-

lows novel biological questions to be approached in the most relevant non-model organisms (e.g. speciation in cichlids).

Single-cell genomics

Single-cell genomics is an exciting area of research that allows us to study cellular biology at an unprecedented resolution. I therefore wished to better understand its limitations. Currently, detection of lowly-expressed transcripts is difficult at the single-cell level and reconstructing whole genome sequences is intractable. Sequencing small quantities of DNA or RNA, such as that present in single cells, is known to be error prone as biases and errors that occur in low levels are propagated through the heavy amplification steps needed for high-throughput sequencing. This often leads to biases in coverage and low technical reproducibility (Nawy, 2014).

Consistent with expectations, I observed that compared to bulk sequencing data, single-cell sequencing of human breast cancer cell lines (HCC38) detected a much smaller number of splice junctions, a proxy for transcriptional complexity. Even when pooling all single-cell sequencing, many splice junctions remained undetected. This means that either (1) some transcripts are not expressed in all cells (a proposition that is known to be the case; e.g. Kaufmann and van Oudenaarden (2007)), or that (2) rare transcripts have a small chance of being amplified and detected. Indeed, if a rare transcript has a one in a hundred (10^{-2}) chance of being amplified, a back of the envelope calculation gives us a probability of only 0.62 or $1 - (1 - 10^{-2})^{96}$ that it is detected in one sample out of 96 sequencing runs.

Unfortunately, distinguishing the two possibilities is difficult with our current

experimental design (though it is likely that both contribute to the differences observed between bulk and single-cell sequencing). In an elegant study from Marinov et al. (2014), the mRNAs from single cells were pooled together and then split again into the same number of reactions that were used to build libraries. Because the mRNAs were mixed, the proportion of each mRNA should be the same (up to sampling error) in each library. Sequencing the libraries therefore allowed them to measure exactly how much variation is associated to technical noise. Although they found that technical limitations prevent detection of lowly expressed genes (genes expressed at 1 FKPM[2] were detected in only 10% of all libraries, while genes expressed at $\geq$100FKPM were detected in almost all libraries), overall gene expression variability was found to be significantly higher than technical variability. This indicates that biological variation must exist between single cells (from a lymphoblastoid cell line). Interestingly, they found that 30 cells is the lower limit at which the transcriptome complexity begins to recapitulate that of bulk sequencing, an observation that I have also made from HCC38 transcriptomes (using transcriptomes produced using the C1 protocol). Clearly, single-cell RNA sequencing technology is still in its infancy. Carefully designed studies, such as the one from Marinov et al. (2014), must therefore be employed until single-cell technologies reach their full maturity.

A second use of single-cell RNA data is to identify single nucleotide polymorphisms (SNPs). I estimated that the false positive rates of the variant calling approach I used was around 50% using RNA-seq data generated using the C1 protocol. However, when SNPs are called in two or more samples, the false positive rate drops down to less than 5%. An interesting application is therefore to study somatic mutations that occur within transcript-coding regions of the genome. For example, by sequencing dividing cells after different numbers of

[2]Fragment Per Kilobase of transcript per Million mapped reads

replication rounds, it should be possible to study somatic mutation rates during cell division. Of course, this is restricted to regions corresponding to highly-expressed genes as variants in these regions should be detected in every cell. Another possibility is to sequence both the RNA and DNA of a single cell with protocols similar to GT-seq. Although current biases in single cell whole genome sequencing protocols do not permit reliable variant detection (due to uneven coverage and biases), exome sequencing may allow replication of variants detected at the RNA level.

SWiPS

Genome assembly remains to date a challenging task. Scaffolding programs such as SWiPS can help us leverage the many low-quality assembly genomes that were produced in the past 10 years (and are still being produced). However, they suffer from several weaknesses. First, despite the scaffolding error being low (less than 10% for most tested assemblies), errors in scaffolding propagate to downstream analyses. For example, gene fragments that are linked incorrectly can lead to errors in orthology prediction. In particular, the reconstructed gene may lead to false positives in positive selection analyses.

With long read sequencing becoming increasingly cheaper and accurate, short-read assemblers and scaffolding approaches are less necessary for genome assemblies. Indeed, PacBio sequencing is already commonly used for sequencing bacterial genomes as error rates approach 1%. Moleculo, an innovative approach that produces 10kb reads can also be used to further reduce the complexity of genome assembly (Kuleshov et al., 2014; Voskoboynik et al., 2013). Nevertheless, a major shortcoming of current long-read sequencing is their low throughput compared to short-read sequencing. This limits the cost-effectiveness of long-

read sequencing for building transcriptomes. A modified version of SWiPS could therefore be used to guide transcript assembly. *De novo* transcript assembly remains difficult, and relies heavily on sequencing depth. Even with high sequencing depths, lowly expressed transcripts remain difficult to reconstruct. For example, while the *de novo* transcript assembler Trinity is able to fully assemble most of the highly expressed transcripts (80-percentile), transcripts expressed at the 20-percentile were generally reconstructed only partially (Grabherr et al., 2011). Scaffolding transcript contigs using orthologous protein sequence should be more effective compared to scaffolding genomic contigs. This is because the proportion of transcript sequence represented by proteins is much higher than that of the genome. Less than 2% of the human genome is coding, while it is expected that polyadenylated[3] RNA consist of as much as 50% bases that are coding. Therefore, a scaffolding approach using protein sequences like SWiPS is promising for linking transcript fragments. Nevertheless, reconstructing multiple isoforms from a single protein remains difficult, in particular considering the large amount of turnover in alternative splicing between even closely-related species.

6.2 Gene duplication analyses

In general, ascribing a phenotypic change to a gene duplication event is extremely difficult. This is because a large fraction of genes have unknown functions. Gene duplication may also lead to phenotypic changes at the cellular level alone, which are difficult to assess. However, when the function of a duplicated gene is known and when there is a conspicuous change in an associated trait, a link between gene duplication and phenotypic change can be made.

[3]The polyadenylated transcriptome is generally the transcriptome fraction that is being sequenced

Beta-keratin gene family expansion in turtles

When I first discovered that a large expansion of the beta-keratin gene family in the genome of the painted turtle, I immediately hypothesised that they contributed to the emergence of the modern turtle shell. This is because the turtle shell is known to consist mostly of layers of beta-keratins (Dalla Valle et al., 2009b). Of course, other key innovations, such as the development of the turtle carapacial ridge, were needed before turtles acquired their shells as we know it. However, the beta-keratin expansion in turtles intrigued me the most considering that the emergence of feathers in birds was also associated with a large lineage-specific beta-keratin expansion at the syntenic genomic locus.

Interestingly, turtles retained two major groups of beta-keratin genes that originated over 160My years ago (since this period corresponds to the most recent common ancestors of the three turtles). Estimates from BEAST however, suggests that this duplication likely occurred even earlier, probably soon after the first turtle-like species emerged. As discussed previously, these two beta-keratin clusters may provide a good example of subfunctionalization. Indeed, evidence suggest that turtle shells evolved in two steps: first the plastron evolved (belly shield), followed by the carapace (Li et al., 2008). The ancestral turtle beta-keratin clusters may therefore have been retained because they first allowed a hard keratineous plastron, and then, when the morphological structure arose (i.e. the development of the carapacial ridge), a novel beta-keratin cluster emerged, that encode beta-keratins that would later form the carapace of the shell.

This hypothesis can be tested by designing cluster-specific primers to quantify the expression of beta-keratin genes from each cluster in plastron and carapace

precursor cells. If the plastron and carapace precursor cells express different sets of beta-keratin genes, then this will be the first molecular evidence that the turtle shell evolved in two steps. An additional prediction is that each set consists of beta-keratin genes that are near each other in the genome as they arose through tandem duplications. Subsequent to their duplications, their promoters evolved independently which allowed fine tuning of the expression levels and cell type specificity of beta-keratin gene.

6.3 Genome evolution in a rapidly diversifying clade

Incomplete lineage sorting

Recently, the study of primate genomes revealed that as much as 30% of the human and chimpanzee genomes was more closely related to the gorilla genome than to the chimpanzee and human genomes, respectively, despite chimpanzees being more closely related to humans than to gorillas. In other words, 30% of the human and chimp genomes are incompletely sorted relative to the gorilla genome. I found that 43% of the haplochromine genomes were incompletely sorted.

To arrive at this number, I used CoalHMM to estimate the coalescence time of DNA segments of the haplochromine genomes, using the tilapia genome as outgroup. CoalHMM uses hidden Markov models to assign genomic segments to a tree and then learns population parameters, such as mutation rates, population size and coalescence time to maximise the likelihood of the model given the observed sequences. The advantage of such approach is that segments are not fixed in length (e.g. windows of 100bp), and parameters are estimated for each segment, yielding distributions of various parameters that can be

interpreted in a biological context. A disadvantage of CoalHMM however is that it is time consuming and relies on well aligned sequences. The ILS estimates are therefore based on aligned regions of the cichlid genomes (i.e. present in four cichlids: *N. brichardi* sequences were not used), which amounts to only one fifth of the whole genome size. However, the amount of ILS is not expected to significantly differ in unaligned regions because overall ILS levels are dominated by neutrally evolving regions and unaligned regions are likely to evolve neutrally.

The high levels of ILS in the haplochromine cichlid genomes made me consider its implications on cichlid biology. Notably, I hypothesised that, as with DNA sequences, ILS could cause levels of gene expression to be more different between two closely-related species than two more distantly-related species. This is because ILS can occur in cis-regulatory regions or even in trans-acting genes. Brawand et al. (2011) observed that phylogenetic trees constructed from pairwise correlation based on gene expression levels in the several primate tissues were different from the species tree. More precisely, they found that for testis, humans and gorillas group together, while chimpanzees fall outside. Although they conjecture that this observation is explained by higher selective pressures on sperm production in chimpanzees relative to humans and bonobos, it is also possible that incomplete lineage sorting of gene regulatory elements (e.g. gene promoters or transcription factor) could also contribute to the discordance between species and expression trees. Unfortunately, possibly due to the lack of biological replicates, I was unable to find any correlation between the genealogy of gene promoter regions and trees built based on their expression levels (data not shown).

The role of *Ednrb1* evolution in cichlid colouration

Perhaps due to the short divergence time between cichlids, genome-wide scans of positive selection and of changes in gene architecture were largely fruitless. Statistical tests lack power to detect signatures of positive selection when a small proportion (<5%) of sites are under positive selection (Yang and dos Reis, 2011). Therefore, positive selection could not be tested in the genes I found to be undergoing rapid evolution as fewer than 2% of their amino acids were substituted. Nevertheless, I identified one gene, endothelin receptor b1 (*Ednrb1*), that shows both evidence of rapid evolution and change in gene architecture in Haplochromine ancestors. In mammals, Ednrb genes have long been studied for their roles during neural crest development, and in fishes, *Ednrb1* is known to be involved in pigment-cell development (Parichy et al., 2000).

A conjecture is therefore that *Edrnb1* polymorphisms were generated in ancestral haplochromines which played a role in the emergence of a large number of different haplochromine cichlid colour morphs. This can be tested by sequencing full length *Ednrb1* in a wide variety of cichlids possessing different colouration. Although Diepeveen and Salzburger (2011) sequenced *Ednrb1* in 26 cichlids to this end, we found that the most variable positions in the *Ednrb1* gene are located at the 3' end (C-terminal region of the protein), which they failed to sequence.

Haplochromine males are also characterised by spots on their anal fins which mimic real eggs and evidence suggest that these egg-spots play an important role in breeding behaviour either through sexual selection (Salzburger et al., 2007), or as marker of male dominance (Lehtonen and Meyer, 2011; Theis et al., 2012). Since *Ednrb1* likely plays a role in the formation of these egg-spots as they are

derived from neural crest cells (Diepeveen and Salzburger, 2011; Braasch et al., 2009), it is possible that the novel 5' UTR exon in *Ednrb1* contributed to the emergence of egg-spots in Haplochromines. Interestingly, a distantly related cichlid species, *O. ventralis*, possesses egg-spots with similar functions that instead evolved on their pelvic fins, likely through convergent evolution (Salzburger et al., 2007). I would therefore be interested to verify whether a similar gain-of-function mutation occurred in the 5' UTR of *Ednrb1* in *O. ventralis*.

6.4 Functions of micro-exons and alternative splicing

RNA splicing occurs through the formation of the spliceosome at exons on the pre-messenger RNA, a process that is guided by regulatory signals encoded within and near exonic splice sites. Therefore, the discovery that exons as small as 3nt were spliced into human transcripts astonished me as they do not have "room" to harbour even a single splicing enhancer. I discovered that AS micro-exons possess highly conserved intronic flanks that likely contain splicing enhancers. It is therefore possible that they help exon recognition. In particular, I found that Rbfox proteins show higher levels of RNA binding downstream of micro-exons, which is predicted to enhance exon inclusion levels. However, these conserved regions alone cannot explain how all micro-exons are recognised by the spliceosome. For instance, constitutively spliced micro-exons, which are spliced more efficiently, do not possess similar conserved intronic flanks.

This brings us to the first (of two) questions that I wish to investigate next. The first question is about exon recognition. As shown in Chapter 5, there are several lines of evidence that suggest a difficulty for the splicing machinery to recognise micro-exons: (1) there is a sharp drop in the number of exons under 51 nt, (2) reducing the size of an internal exon below 51nt induced exon

skipping in gene constructs (Dominski and Kole, 1991), (3) the ratio of AS to CS micro-exons is over twice higher than the ratio of AS to CS exons for longer exons, and (4) there is little evidence of noisy splicing of micro-exons as most are highly conserved across vertebrates. There should therefore be compensatory mechanisms or stronger splicing signals that allow micro-exons to be properly recognised constitutively. In my analyses, I was unable to find large differences in genomic attributes between CS micro-exons and longer CS exons. However, it is possible that the combined differences in several attributes can predict whether a micro-exon can be spliced properly. For example, I found that CS micro-exons had shorter flanking introns and slightly, but significantly, stronger splice sites than longer CS exons. Preliminary results from counting known enhancer hexamers (not shown) suggest that CS micro-exons also possess a higher density of exonic enhancers. A more rigorous approach verifying that the third sites of codons are more conserved in CS micro-exons than in CS exons (matching distance to splice site), a hallmark of purifying selection on exonic enhancers, would be of interest.

The second question pertains to the role of micro-exons in brain-related diseases. I have shown that micro-exons are likely to be involved in brain development. Furthermore, Rbfox proteins appear to regulate brain-specific patterns of micro-exon splicing and their dysregulation has been associated to several brain disorders including autism, mental retardation and epilepsy. It is therefore possible that aberrant splicing of micro-exons contributes to disease. In particular, Rbfox1 was found to regulate splicing of significantly more exonic targets that are differentially spliced in autism patients than would be expected by chance (Weyn-Vanhentenryck et al., 2014). Furthermore, Weyn-Vanhentenryck et al. (2014) also observed that a significantly larger

number of candidate autism-susceptibility genes (Basu et al., 2009) harboured Rbfox targets than would be expected by chance. It may therefore be possible to identify mutations associated with autism that lie near or within micro-exons. Mutations at these locations may alter splicing efficiency or amino acids encoded by micro-exons. Ideally, a set of *de novo* mutations associated with autism can be obtained from the study of trios (healthy parents, child with autism). We can then ask the question of whether there is a significantly larger number of *de novo* mutations near or within micro-exons than compared to normal exons.

Another interesting disease-related finding was the discovery of a 6nt micro-exon encoding 2 residues alternatively included in a beta-turn loop of the Amyloid-Beta Precursor Protein-Binding, Family B protein (*APBB*). By analysing the tertiary structure of APBB2 in complex with amyloid-beta protein, we found that this beta-turn loop likely interacts with the cytoplasmic tail of amyloid-beta protein. Aberrant splicing of the conserved AS micro-exons in Apbb genes may therefore affect APBB's interaction with amyloid-beta protein and contribute to the formation of amyloid plaques in Alzheimer's disease (Hu et al., 1999). Indeed, several studies identified intronic variants in Apbb2 that were associated to Alzheimer's disease (Golanska et al., 2008; Li et al., 2005). Inclusion of the APBB1 micro-exon is restricted to neuronal cells (Hu et al., 1999), which suggests it possesses a brain-specific role. Interestingly, the inclusion ratio of the 6nt micro-exon in Apbb2, and to a lesser extent in Apbb1, decreases in human ageing brains (data not shown). A hypothesis is that an appropriate ratio of Apbb2 isoforms (inclusion and exclusion of the micro-exon) must be maintained to prevent amyloid plaque build-ups, a balance that might be disrupted by the ageing process.

Bibliography

Aird, D., Ross, M. G., Chen, W. S., Danielsson, M., Fennell, T., Russ, C., Jaffe, D. B., Nusbaum, C., and Gnirke, A., 2011. Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biol.*, **12**(2):R18.

Alföldi, J., Di Palma, F., Grabherr, M., Williams, C., Kong, L., Mauceli, E., Russell, P., Lowe, C. B., Glor, R. E., Jaffe, J. D., *et al.*, 2011. The genome of the green anole lizard and a comparative analysis with birds and mammals. *Nature*, **477**(7366):587–591.

Alibardi, L., 2002. Immunocytochemical observations on the cornification of soft and hard epidermis in the turtle chrysemys picta. *Zoology (Jena)*, **105**(1):31–44.

Alibardi, L., Dalla Valle, L., Nardi, A., and Toni, M., 2009. Evolution of hard proteins in the sauropsid integument in relation to the cornification of skin derivatives in amniotes. *J Anat*, **214**(4):560–586.

Alibardi, L., Toni, M., and Dalla Valle, L., 2007. Hard cornification in reptilian epidermis in comparison to cornification in mammalian epidermis. *Experiment Dermatology*, **16**:961–976.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J., 1990. Basic local alignment search tool. *J. Mol. Biol.*, **215**(3):403–410.

Amemiya, C. T., Alfoldi, J., Lee, A. P., Fan, S., Philippe, H., Maccallum, I., Braasch, I., Manousaki, T., Schneider, I., Rohner, N., *et al.*, 2013. The African coelacanth genome provides insights into tetrapod evolution. *Nature*, **496**(7445):311–316.

Amit, M., Donyo, M., Hollander, D., Goren, A., Kim, E., Gelfman, S., Lev-Maor, G., Burstein, D., Schwartz, S., Postolsky, B., *et al.*, 2012. Differential GC content between exons and introns establishes distinct strategies of splice-site recognition. *Cell Rep*, **1**(5):543–556.

Aparicio, S., Chapman, J., Stupka, E., Putnam, N., Chia, J. M., Dehal, P., Christoffels, A., Rash, S., Hoon, S., Smit, A., *et al.*, 2002. Whole-genome shotgun assembly and analysis of the genome of Fugu rubripes. *Science*, **297**(5585):1301–1310.

Assis, R. and Bachtrog, D., 2014. Gradual divergence and diversification of mammalian duplicate gene functions. *biorxiv*, **1**(1):1.

Avery, O. T., Macleod, C. M., and McCarty, M., 1944. STUDIES ON THE CHEMICAL NATURE OF THE SUBSTANCE INDUCING TRANSFORMATION OF PNEUMOCOCCAL TYPES : INDUCTION OF TRANSFORMATION BY A DESOXYRIBONUCLEIC ACID FRACTION ISOLATED FROM PNEUMOCOCCUS TYPE III. *J. Exp. Med.*, **79**(2):137–158.

Babu, M. M., van der Lee, R., de Groot, N. S., and Gsponer, J., 2011. Intrinsically disordered proteins: regulation and disease. *Curr. Opin. Struct. Biol.*, **21**(3):432–440.

Barbosa-Morais, N. L., Irimia, M., Pan, Q., Xiong, H. Y., Gueroussov, S., Lee, L. J., Slobodeniuc, V., Kutter, C., Watt, S., Colak, R., *et al.*, 2012. The

evolutionary landscape of alternative splicing in vertebrate species. *Science*, **338**(6114):1587–1593.

Barten, R., Torkar, M., Haude, A., Trowsdale, J., and Wilson, M. J., 2001. Divergent and convergent evolution of nk-cell receptors. *Trends Immunol*, **22**(1):52–57.

Basu, S. N., Kollu, R., and Banerjee-Basu, S., 2009. AutDB: a gene reference resource for autism research. *Nucleic Acids Res.*, **37**(Database issue):D832–836.

Bergsten, J., 2005. A review of long-branch attraction. *Cladistics*, **21**(3):163–193.

Bernardi, G., 2000. Isochores and the evolutionary genomics of vertebrates. *Gene*, **241**(1):3–17.

Birney, E., Clamp, M., and Durbin, R., 2004. GeneWise and Genomewise. *Genome Res.*, **14**(5):988–995.

Birol, I., Jackman, S. D., Nielsen, C. B., Qian, J. Q., Varhol, R., Stazyk, G., Morin, R. D., Zhao, Y., Hirst, M., Schein, J. E., *et al.*, 2009. De novo transcriptome assembly with ABySS. *Bioinformatics*, **25**(21):2872–2877.

Black, D. L., 1991. Does steric interference between splice sites block the splicing of a short c-src neuron-specific exon in non-neuronal cells? *Genes Dev.*, **5**(3):389–402.

Black, D. L., 2000. Protein diversity from alternative splicing: a challenge for bioinformatics and post-genome biology. *Cell*, **103**(3):367–370.

Blanchette, M., Kent, W. J., Riemer, C., Elnitski, L., Smit, A. F., Roskin, K. M., Baertsch, R., Rosenbloom, K., Clawson, H., Green, E. D., *et al.*, 2004. Aligning

multiple genomic sequences with the threaded blockset aligner. *Genome Res.*, **14**(4):708–715.

Blekhman, R., Oshlack, A., Chabot, A. E., Smyth, G. K., and Gilad, Y., 2008. Gene regulation in primates evolves under tissue-specific selection pressures. *PLoS Genet.*, **4**(11):e1000271.

Blencowe, B. J., 2000. Exonic splicing enhancers: mechanism of action, diversity and role in human genetic diseases. *Trends Biochem. Sci.*, **25**(3):106–110.

Boyko, A. R., Williamson, S. H., Indap, A. R., Degenhardt, J. D., Hernandez, R. D., Lohmueller, K. E., Adams, M. D., Schmidt, S., Sninsky, J. J., Sunyaev, S. R., *et al.*, 2008. Assessing the evolutionary impact of amino acid mutations in the human genome. *PLoS Genet.*, **4**(5):e1000083.

Braasch, I., Volff, J. N., and Schartl, M., 2009. The endothelin system: evolution of vertebrate-specific ligand-receptor interactions by three rounds of genome duplication. *Mol. Biol. Evol.*, **26**(4):783–799.

Brawand, D., Soumillon, M., Necsulea, A., Julien, P., Csardi, G., Harrigan, P., Weier, M., Liechti, A., Aximu-Petri, A., Kircher, M., *et al.*, 2011. The evolution of gene expression levels in mammalian organs. *Nature*, **478**(7369):343–348.

Brett, D., Pospisil, H., Valcarcel, J., Reich, J., and Bork, P., 2002. Alternative splicing and genome complexity. *Nat. Genet.*, **30**(1):29–30.

Brooke, N. M., Garcia-Fernandez, J., and Holland, P. W., 1998. The ParaHox gene cluster is an evolutionary sister of the Hox gene cluster. *Nature*, **392**(6679):920–922.

Buljan, M., Chalancon, G., Eustermann, S., Wagner, G. P., Fuxreiter, M., Bateman, A., and Babu, M. M., 2012. Tissue-specific splicing of disordered seg-

ments that embed binding motifs rewires protein interaction networks. *Mol. Cell*, **46**(6):871–883.

Burge, C. and Karlin, S., 1997. Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.*, **268**(1):78–94.

Bustamante, C. D., Fledel-Alon, A., Williamson, S., Nielsen, R., Hubisz, M. T., Glanowski, S., Tanenbaum, D. M., White, T. J., Sninsky, J. J., Hernandez, R. D., *et al.*, 2005. Natural selection on protein-coding genes in the human genome. *Nature*, **437**(7062):1153–1157.

Caceres, E. F. and Hurst, L. D., 2013. The evolution, impact and properties of exonic splice enhancers. *Genome Biol.*, **14**(12):R143.

Carlo, T., Sierra, R., and Berget, S. M., 2000. A 5' splice site-proximal enhancer binds SF1 and activates exon bridging of a microexon. *Mol. Cell. Biol.*, **20**(11):3988–3995.

Carroll, S. B., 1995. Homeotic genes and the evolution of arthropods and chordates. *Nature*, **376**(6540):479–485.

Cebra-Thomas, J., Tan, F., Sistla, S., Estes, E., Bender, G., Kim, C., Riccio, P., and Gilbert, S. F., 2005. How the turtle forms its shell: a paracrine hypothesis of carapace formation. *J Exp Zool B Mol Dev Evol*, **304**(6):558–569.

Chang, D. and Duda, Jr, T. F., 2012. Extensive and continuous duplication facilitates rapid evolution and diversification of gene families. *Mol Biol Evol*, **29**(8):2019–2029.

Chen, Y. C., Liu, T., Yu, C. H., Chiang, T. Y., and Hwang, C. C., 2013. Effects of GC bias in next-generation-sequencing data on de novo genome assembly. *PLoS ONE*, **8**(4):e62856.

Chiaromonte, F., Weber, R. J., Roskin, K. M., Diekhans, M., Kent, W. J., and Haussler, D., 2003. The share of human genomic DNA under selection estimated from human-mouse genomic alignments. *Cold Spring Harb. Symp. Quant. Biol.*, **68**:245–254.

Clark, A. G., Glanowski, S., Nielsen, R., Thomas, P. D., Kejariwal, A., Todd, M. A., Tanenbaum, D. M., Civello, D., Lu, F., Murphy, B., *et al.*, 2003. Inferring nonneutral evolution from human-chimp-mouse orthologous gene trios. *Science*, **302**(5652):1960–1963.

Costantini, M., Cammarano, R., and Bernardi, G., 2009. The evolution of isochore patterns in vertebrate genomes. *BMC Genomics*, **10**:146.

Cuccurese, M., Russo, G., Russo, A., and Pietropaolo, C., 2005. Alternative splicing and nonsense-mediated mRNA decay regulate mammalian ribosomal gene expression. *Nucleic Acids Res.*, **33**(18):5965–5977.

Dalerba, P., Kalisky, T., Sahoo, D., Rajendran, P. S., Rothenberg, M. E., Leyrat, A. A., Sim, S., Okamoto, J., Johnston, D. M., Qian, D., *et al.*, 2011. Single-cell dissection of transcriptional heterogeneity in human colon tumors. *Nat. Biotechnol.*, **29**(12):1120–1127.

Dalla Valle, L., Nardi, A., Gelmi, C., Toni, M., Emera, D., and Alibardi, L., 2009a. Beta-keratins of the crocodilian epidermis: composition, structure, and phylogenetic relationships. *J Exp Zool B Mol Dev Evol*, **312**(1):42–57.

Dalla Valle, L., Nardi, A., Toni, M., Emera, D., and Alibardi, L., 2009b. Beta-keratins of turtle shell are glycine-proline-tyrosine rich proteins similar to those of crocodilians and birds. *J Anat*, **214**(2):284–300.

Davey, J. W., Cezard, T., Fuentes-Utrilla, P., Eland, C., Gharbi, K., and Blax-

ter, M. L., 2013. Special features of RAD Sequencing data: implications for genotyping. *Mol. Ecol.*, **22**(11):3151–3164.

De Grassi, A. and Ciccarelli, F. D., 2009. Tandem repeats modify the structure of human genes hosted in segmental duplications. *Genome Biol.*, **10**(12):R137.

Dehal, P. and Boore, J. L., 2005. Two rounds of whole genome duplication in the ancestral vertebrate. *PLoS Biol.*, **3**(10):e314.

Dehal, P., Satou, Y., Campbell, R. K., Chapman, J., Degnan, B., De Tomaso, A., Davidson, B., Di Gregorio, A., Gelpke, M., Goodstein, D. M., *et al.*, 2002. The draft genome of Ciona intestinalis: insights into chordate and vertebrate origins. *Science*, **298**(5601):2157–2167.

Dessimoz, C., Zoller, S., Manousaki, T., Qiu, H., Meyer, A., and Kuraku, S., 2011. Comparative genomics approach to detecting split-coding regions in a low-coverage genome: lessons from the chimaera Callorhinchus milii (Holocephali, Chondrichthyes). *Brief. Bioinformatics*, **12**(5):474–484.

Diepeveen, E. T. and Salzburger, W., 2011. Molecular characterization of two endothelin pathways in East African cichlid fishes. *J. Mol. Evol.*, **73**(5-6):355–368.

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T. R., 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, **29**(1):15–21.

Dohm, J. C., Lottaz, C., Borodina, T., and Himmelbauer, H., 2008. Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Res.*, **36**(16):e105.

Doi, T., Hiroaki, Y., Arimoto, I., Fujiyoshi, Y., Okamoto, T., Satoh, M., and

Furuichi, Y., 1997. Characterization of human endothelin B receptor and mutant receptors expressed in insect cells. *Eur. J. Biochem.*, **248**(1):139–148.

Dominski, Z. and Kole, R., 1991. Selection of splice sites in pre-mRNAs with short internal exons. *Mol. Cell. Biol.*, **11**(12):6075–6083.

Dorus, S., Vallender, E. J., Evans, P. D., Anderson, J. R., Gilbert, S. L., Mahowald, M., Wyckoff, G. J., Malcom, C. M., and Lahn, B. T., 2004. Accelerated evolution of nervous system genes in the origin of Homo sapiens. *Cell*, **119**(7):1027–1040.

Doyon, J. P., Chauve, C., and Hamel, S., 2009. Space of gene/species trees reconciliations and parsimonious models. *J. Comput. Biol.*, **16**(10):1399–1418.

Drummond, A. J. and Rambaut, A., 2007. Beast: Bayesian evolutionary analysis by sampling trees. *BMC Evol Biol*, **7**:214.

Duret, L. and Galtier, N., 2009. Biased gene conversion and the evolution of mammalian genomic landscapes. *Annu Rev Genomics Hum Genet*, **10**:285–311.

Dutheil, J. Y., Ganapathy, G., Hobolth, A., Mailund, T., Uyenoyama, M. K., and Schierup, M. H., 2009. Ancestral population genomics: the coalescent hidden Markov model approach. *Genetics*, **183**(1):259–274.

Dvir, S., Velten, L., Sharon, E., Zeevi, D., Carey, L. B., Weinberger, A., and Segal, E., 2013. Deciphering the rules by which 5'-UTR sequences affect protein expression in yeast. *Proc. Natl. Acad. Sci. U.S.A.*, **110**(30):E2792–2801.

Eckhart, L., Valle, L. D., Jaeger, K., Ballaun, C., Szabo, S., Nardi, A., Buchberger, M., Hermann, M., Alibardi, L., and Tschachler, E., *et al.*, 2008. Identification of reptilian genes encoding hair keratin-like proteins suggests

a new scenario for the evolutionary origin of hair. *Proc Natl Acad Sci U S A*, **105**(47):18419–18423.

Eddy, S., 2010. *HMMER user's guide version 3.0*. Department of Mathematics, Washington University in St. Louis.

Edgar, R. C., 2004. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*, **5**:113.

Eichler, E. E. and Sankoff, D., 2003. Structural dynamics of eukaryotic chromosome evolution. *Science*, **301**(5634):793–797.

Enard, W., Przeworski, M., Fisher, S. E., Lai, C. S., Wiebe, V., Kitano, T., Monaco, A. P., and Paabo, S., 2002. Molecular evolution of FOXP2, a gene involved in speech and language. *Nature*, **418**(6900):869–872.

Engstrom, P. G., Steijger, T., Sipos, B., Grant, G. R., Kahles, A., Ratsch, G., Goldman, N., Hubbard, T. J., Harrow, J., Guigo, R., *et al.*, 2013. Systematic evaluation of spliced alignment programs for RNA-seq data. *Nat. Methods*, **10**(12):1185–1191.

Ermakova, E. O., Nurtdinov, R. N., and Gelfand, M. S., 2006. Fast rate of evolution in alternatively spliced coding regions of mammalian genes. *BMC Genomics*, **7**:84.

Evans, P. D., Anderson, J. R., Vallender, E. J., Gilbert, S. L., Malcom, C. M., Dorus, S., and Lahn, B. T., 2004. Adaptive evolution of ASPM, a major determinant of cerebral cortical size in humans. *Hum. Mol. Genet.*, **13**(5):489–494.

Ewing, A. D., Ballinger, T. J., Earl, D., Harris, C. C., Ding, L., Wilson, R. K., and Haussler, D., 2013. Retrotransposition of gene transcripts leads to structural variation in mammalian genomes. *Genome Biol.*, **14**(3):R22.

Eyre-Walker, A., 2006. The genomic rate of adaptive evolution. *Trends Ecol. Evol. (Amst.)*, **21**(10):569–575.

Fairbrother, W. G., Yeh, R. F., Sharp, P. A., and Burge, C. B., 2002. Predictive identification of exonic splicing enhancers in human genes. *Science*, **297**(5583):1007–1013.

Fan, C., Chen, Y., and Long, M., 2008. Recurrent tandem gene duplication gave rise to functionally divergent genes in Drosophila. *Mol. Biol. Evol.*, **25**(7):1451–1458.

Felsenstein, J., 1989. Mathematics vs. Evolution: Mathematical Evolutionary Theory. *Science*, **246**(4932):941–942.

Fonseca, N. A., Rung, J., Brazma, A., and Marioni, J. C., 2012. Tools for mapping high-throughput sequencing data. *Bioinformatics*, **28**(24):3169–3177.

Fraser, H. B., 2013. Gene expression drives local adaptation in humans. *Genome Res.*, **23**(7):1089–1096.

Gan, X., Stegle, O., Behr, J., Steffen, J. G., Drewe, P., Hildebrand, K. L., Lyngsoe, R., Schultheiss, S. J., Osborne, E. J., Sreedharan, V. T., *et al.*, 2011. Multiple reference genomes and transcriptomes for Arabidopsis thaliana. *Nature*, **477**(7365):419–423.

Garcia-Fernandez, J., 2005. Hox, ParaHox, ProtoHox: facts and guesses. *Heredity (Edinb)*, **94**(2):145–152.

Gehman, L. T., Meera, P., Stoilov, P., Shiue, L., O'Brien, J. E., Meisler, M. H., Ares, M., Otis, T. S., and Black, D. L., 2012. The splicing regulator Rbfox2 is required for both cerebellar development and mature motor function. *Genes Dev.*, **26**(5):445–460.

Gehman, L. T., Stoilov, P., Maguire, J., Damianov, A., Lin, C. H., Shiue, L., Ares, M., Mody, I., and Black, D. L., 2011. The splicing regulator Rbfox1 (A2BP1) controls neuronal excitation in the mammalian brain. *Nat. Genet.*, **43**(7):706–711.

Gelfman, S., Burstein, D., Penn, O., Savchenko, A., Amit, M., Schwartz, S., Pupko, T., and Ast, G., 2012. Changes in exon-intron structure during vertebrate evolution affect the splicing pattern of exons. *Genome Res.*, **22**(1):35–50.

Genner, M. J., Seehausen, O., Lunt, D. H., Joyce, D. A., Shaw, P. W., Carvalho, G. R., and Turner, G. F., 2007. Age of cichlids: new dates for ancient lake fish radiations. *Mol. Biol. Evol.*, **24**(5):1269–1282.

Gertz, E. M., Yu, Y. K., Agarwala, R., Schaffer, A. A., and Altschul, S. F., 2006. Composition-based statistics and translated nucleotide searches: improving the TBLASTN module of BLAST. *BMC Biol.*, **4**:41.

Gilad, Y., Oshlack, A., and Rifkin, S. A., 2006. Natural selection on gene expression. *Trends Genet.*, **22**(8):456–461.

Gilbert, S. F., Loredo, G. A., Brukman, A., and Burke, A. C., 2001. Morphogenesis of the turtle shell: the development of a novel structure in tetrapod evolution. *Evol Dev*, **3**(2):47–58.

Gnerre, S., Maccallum, I., Przybylski, D., Ribeiro, F. J., Burton, J. N., Walker, B. J., Sharpe, T., Hall, G., Shea, T. P., Sykes, S., *et al.*, 2011. High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proc. Natl. Acad. Sci. U.S.A.*, **108**(4):1513–1518.

Golanska, E., Sieruta, M., Gresner, S. M., Hulas-Bigoszewska, K., Corder, E. H., Styczynska, M., Peplonska, B., Barcikowska, M., and Liberski, P. P., 2008.

Analysis of APBB2 gene polymorphisms in sporadic Alzheimer's disease. *Neurosci. Lett.*, **447**(2-3):164–166.

Goodstadt, L. and Ponting, C. P., 2006. Phylogenetic reconstruction of orthology, paralogy, and conserved synteny for dog and human. *PLoS Comput. Biol.*, **2**(9):e133.

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., *et al.*, 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.*, **29**(7):644–652.

Graveley, B. R., 2001. Alternative splicing: increasing diversity in the proteomic world. *Trends Genet.*, **17**(2):100–107.

Greenwold, M. J. and Sawyer, R. H., 2010. Genomic organization and molecular phylogenies of the beta (beta) keratin multigene family in the chicken (gallus gallus) and zebra finch (taeniopygia guttata): implications for feather evolution. *BMC Evol Biol*, **10**:148.

Greenwold, M. J. and Sawyer, R. H., 2011. Linking the molecular evolution of avian beta (β) keratins to the evolution of feathers. *J Exp Zool B Mol Dev Evol*, **316**(8):609–616.

Grossman, S. R., Andersen, K. G., Shlyakhter, I., Tabrizi, S., Winnicki, S., Yen, A., Park, D. J., Griesemer, D., Karlsson, E. K., Wong, S. H., *et al.*, 2013. Identifying recent adaptations in large-scale genomic data. *Cell*, **152**(4):703–713.

Guigo, R., Knudsen, S., Drake, N., and Smith, T., 1992. Prediction of gene structure. *J. Mol. Biol.*, **226**(1):141–157.

Guindon, S., Dufayard, J.-F., Lefort, V., Anisimova, M., Hordijk, W., and Gascuel, O., 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of phyml 3.0. *Syst Biol*, **59**(3):307–321.

Guindon, S. and Gascuel, O., 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst. Biol.*, **52**(5):696–704.

Haas, B. J., Salzberg, S. L., Zhu, W., Pertea, M., Allen, J. E., Orvis, J., White, O., Buell, C. R., and Wortman, J. R., 2008. Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biol.*, **9**(1):R7.

Hahn, M. W., Han, M. V., and Han, S. G., 2007. Gene family evolution across 12 Drosophila genomes. *PLoS Genet.*, **3**(11):e197.

Han, M. V., Demuth, J. P., McGrath, C. L., Casola, C., and Hahn, M. W., 2009. Adaptive evolution of young gene duplicates in mammals. *Genome Res.*, **19**(5):859–867.

Hardison, R. C., 2003. Comparative genomics. *PLoS Biol.*, **1**(2):E58.

Heger, A. and Ponting, C. P., 2007. Evolutionary rate analyses of orthologs and paralogs from 12 drosophila genomes. *Genome Res*, **17**(12):1837–1849.

Heger, A. and Ponting, C. P., 2008. Optic: orthologous and paralogous transcripts in clades. *Nucleic Acids Res*, **36**(Database issue):D267–D270.

Heger, A., Webber, C., Goodson, M., Ponting, C. P., and Lunter, G., 2013. GAT: a simulation framework for testing the association of genomic intervals. *Bioinformatics*, **29**(16):2046–2048.

Hiller, M., Huse, K., Szafranski, K., Jahn, N., Hampe, J., Schreiber, S., Backofen, R., and Platzer, M., 2004. Widespread occurrence of alternative splic-

ing at NAGNAG acceptors contributes to proteome plasticity. *Nat. Genet.*, **36**(12):1255–1257.

Hobolth, A., Christensen, O. F., Mailund, T., and Schierup, M. H., 2007. Genomic relationships and speciation times of human, chimpanzee, and gorilla inferred from a coalescent hidden Markov model. *PLoS Genet.*, **3**(2):e7.

Hoege, C., Pfander, B., Moldovan, G. L., Pyrowolakis, G., and Jentsch, S., 2002. RAD6-dependent DNA repair is linked to modification of PCNA by ubiquitin and SUMO. *Nature*, **419**(6903):135–141.

Holland, P. W., Garcia-Fernandez, J., Williams, N. A., and Sidow, A., 1994. Gene duplications and the origins of vertebrate development. *Dev. Suppl.*, :125–133.

Horstmeyer, A., Cramer, H., Sauer, T., Muller-Esterl, W., and Schroeder, C., 1996. Palmitoylation of endothelin receptor A. Differential modulation of signal transduction activity by post-translational modification. *J. Biol. Chem.*, **271**(34):20811–20819.

Hu, Q., Hearn, M. G., Jin, L. W., Bressler, S. L., and Martin, G. M., 1999. Alternatively spliced isoforms of FE65 serve as neuron-specific and non-neuronal markers. *J. Neurosci. Res.*, **58**(5):632–640.

Huang, S., Chen, Z., Huang, G., Yu, T., Yang, P., Li, J., Fu, Y., Yuan, S., Chen, S., and Xu, A., *et al.*, 2012. HaploMerger: reconstructing allelic relationships for polymorphic diploid genome assemblies. *Genome Res.*, **22**(8):1581–1588.

Huang, Y., Li, Y., Burt, D. W., Chen, H., Zhang, Y., Qian, W., Kim, H., Gan, S., Zhao, Y., Li, J., *et al.*, 2013. The duck genome and transcriptome provide insight into an avian influenza virus reservoir species. *Nat. Genet.*, **45**(7):776–783.

Hunt, M., Kikuchi, T., Sanders, M., Newbold, C., Berriman, M., and Otto, T. D., 2013. REAPR: a universal tool for genome assembly evaluation. *Genome Biol.*, **14**(5):R47.

Illumina Human Body Map 2.0, ., 2011. Accesesed from `http://www.illumina.com/science/data_library.ilmn`.

Innan, H. and Kondrashov, F., 2010. The evolution of gene duplications: classifying and distinguishing between models. *Nat Rev Genet*, **11**(2):97–108.

International Chicken Genome Sequencing Consortium, 2004. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature*, **432**(7018):695–716.

Iqbal, Z., Caccamo, M., Turner, I., Flicek, P., and McVean, G., 2012. De novo assembly and genotyping of variants using colored de Bruijn graphs. *Nat. Genet.*, **44**(2):226–232.

Jiang, L., Schlesinger, F., Davis, C. A., Zhang, Y., Li, R., Salit, M., Gingeras, T. R., and Oliver, B., 2011. Synthetic spike-in standards for RNA-seq experiments. *Genome Res.*, **21**(9):1543–1551.

Johnson, S. C., Rabinovitch, P. S., and Kaeberlein, M., 2013. mTOR is a key modulator of ageing and age-related disease. *Nature*, **493**(7432):338–345.

Jones, D. T., 1999. Protein secondary structure prediction based on position-specific scoring matrices. *J Mol Biol*, **292**(2):195–202.

Joyce, D. A., Lunt, D. H., Genner, M. J., Turner, G. F., Bills, R., and Seehausen, O., 2011. Repeated colonization and hybridization in Lake Malawi cichlids. *Curr. Biol.*, **21**(3):R108–109.

Joyce, W. G., Lucas, S. G., Scheyer, T. M., Heckert, A. B., and Hunt, A. P., 2009. A thin-shelled reptile from the late triassic of north america and the origin of the turtle shell. *Proc Biol Sci*, **276**(1656):507–513.

Kaessmann, H., 2010. Origins, evolution, and phenotypic impact of new genes. *Genome Res.*, **20**(10):1313–1326.

Katz, Y., Wang, E. T., Airoldi, E. M., and Burge, C. B., 2010. Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nat. Methods*, **7**(12):1009–1015.

Kaufmann, B. B. and van Oudenaarden, A., 2007. Stochastic gene expression: from single molecules to the proteome. *Curr. Opin. Genet. Dev.*, **17**(2):107–112.

Kelemen, O., Convertini, P., Zhang, Z., Wen, Y., Shen, M., Falaleeva, M., and Stamm, S., 2013. Function of alternative splicing. *Gene*, **514**(1):1–30.

Keren, H., Lev-Maor, G., and Ast, G., 2010. Alternative splicing and evolution: diversification, exon definition and function. *Nat. Rev. Genet.*, **11**(5):345–355.

Khan, Z., Ford, M. J., Cusanovich, D. A., Mitrano, A., Pritchard, J. K., and Gilad, Y., 2013. Primate transcript and protein expression levels evolve under compensatory selection pressures. *Science*, **342**(6162):1100–1104.

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., and Salzberg, S. L., 2013. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.*, **14**(4):R36.

Kim, E. B., Fang, X., Fushan, A. A., Huang, Z., Lobanov, A. V., Han, L., Marino, S. M., Sun, X., Turanov, A. A., Yang, P., *et al.*, 2011. Genome sequencing reveals insights into physiology and longevity of the naked mole rat. *Nature*, **479**(7372):223–227.

Kim, Y. and Nirenberg, M., 1989. Drosophila NK-homeobox genes. *Proc. Natl. Acad. Sci. U.S.A.*, **86**(20):7716–7720.

Kimura, M., 1983. *The Neutral Theory of Molecular Evolution*. Cambridge University Press, Sunderland, MA.

Kocher, T. D., 2004. Adaptive evolution and explosive speciation: the cichlid fish model. *Nat. Rev. Genet.*, **5**(4):288–298.

Konopka, G., Bomar, J. M., Winden, K., Coppola, G., Jonsson, Z. O., Gao, F., Peng, S., Preuss, T. M., Wohlschlegel, J. A., and Geschwind, D. H., *et al.*, 2009. Human-specific transcriptional regulation of CNS development genes by FOXP2. *Nature*, **462**(7270):213–217.

Kosiol, C., Vinar, T., da Fonseca, R. R., Hubisz, M. J., Bustamante, C. D., Nielsen, R., and Siepel, A., 2008. Patterns of positive selection in six Mammalian genomes. *PLoS Genet.*, **4**(8):e1000144.

Kudaravalli, S., Veyrieras, J. B., Stranger, B. E., Dermitzakis, E. T., and Pritchard, J. K., 2009. Gene expression levels are a target of recent natural selection in the human genome. *Mol. Biol. Evol.*, **26**(3):649–658.

Kuleshov, V., Xie, D., Chen, R., Pushkarev, D., Ma, Z., Blauwkamp, T., Kertesz, M., and Snyder, M., 2014. Whole-genome haplotyping using long reads and statistical methods. *Nat. Biotechnol.*, **32**(3):261–266.

Kuraku, S., Usuda, R., and Kuratani, S., 2005. Comprehensive survey of carapacial ridge-specific genes in turtle implies co-option of some regulatory genes in carapace evolution. *Evol Dev*, **7**(1):3–17.

Lalueza-Fox, C., Rompler, H., Caramelli, D., Staubert, C., Catalano, G., Hughes, D., Rohland, N., Pilli, E., Longo, L., Condemi, S., *et al.*, 2007. A melanocortin

1 receptor allele suggests varying pigmentation among Neanderthals. *Science*, **318**(5855):1453–1455.

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., *et al.*, 2001. Initial sequencing and analysis of the human genome. *Nature*, **409**(6822):860–921.

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S. L., 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, **10**(3):R25.

Lappalainen, T., Sammeth, M., Friedlander, M. R., 't Hoen, P. A., Monlong, J., Rivas, M. A., Gonzalez-Porta, M., Kurbatova, N., Griebel, T., Ferreira, P. G., *et al.*, 2013. Transcriptome and genome sequencing uncovers functional variation in humans. *Nature*, **501**(7468):506–511.

Laun, K., Coggill, P., Palmer, S., Sims, S., Ning, Z., Ragoussis, J., Volpi, E., Wilson, N., Beck, S., Ziegler, A., *et al.*, 2006. The leukocyte receptor complex in chicken is characterized by massive expansion and diversification of immunoglobulin-like loci. *PLoS Genet*, **2**(5):e73.

Lawless, C., Pearson, R. D., Selley, J. N., Smirnova, J. B., Grant, C. M., Ashe, M. P., Pavitt, G. D., and Hubbard, S. J., 2009. Upstream sequence elements direct post-transcriptional regulation of gene expression under stress conditions in yeast. *BMC Genomics*, **10**:7.

Lawrie, D. S. and Petrov, D. A., 2014. Comparative population genomics: power and principles for the inference of functionality. *Trends Genet.*, **30**(4):133–139.

Lee, M. S. Y., 1996. Correlated progression and the origin of turtles. *Nature*, **379**:812–815.

Lehtonen, T. K. and Meyer, A., 2011. Heritability and adaptive significance of the number of egg-dummies in the cichlid fish Astatotilapia burtoni. *Proc. Biol. Sci.*, **278**(1716):2318–2324.

Li, C., Wu, X.-C., Rieppel, O., Wang, L.-T., and Zhao, L.-J., 2008. An ancestral turtle from the late triassic of southwestern china. *Nature*, **456**(7221):497–501.

Li, H., 2013. Aligning sequence reads, clone sequences and assembly contigs with bwa-mem. *arxiv*, **1**(1):1.

Li, H. and Durbin, R., 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, **25**(14):1754–1760.

Li, H. and Durbin, R., 2010. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics*, **26**(5):589–595.

Li, H. and Durbin, R., 2011. Inference of human population history from individual whole-genome sequences. *Nature*, **475**(7357):493–496.

Li, R., Zhu, H., Ruan, J., Qian, W., Fang, X., Shi, Z., Li, Y., Li, S., Shan, G., Kristiansen, K., *et al.*, 2010. De novo assembly of human genomes with massively parallel short read sequencing. *Genome Res.*, **20**(2):265–272.

Li, Y., Hollingworth, P., Moore, P., Foy, C., Archer, N., Powell, J., Nowotny, P., Holmans, P., O'Donovan, M., Tacey, K., *et al.*, 2005. Genetic association of the APP binding protein 2 gene (APBB2) with late onset Alzheimer disease. *Hum. Mutat.*, **25**(3):270–277.

Li, Y. I. and Copley, R. R., 2013. Scaffolding low quality genomes using orthologous protein sequences. *Bioinformatics*, **29**(2):160–165.

Li, Y. I., Kong, L., Ponting, C. P., and Haerty, W., 2013. Rapid evolution of

Beta-keratin genes contribute to phenotypic differences that distinguish turtles and birds from other reptiles. *Genome Biol Evol*, **5**(5):923–933.

Liu, B. and Wu, D., 2003. The first inner loop of endothelin receptor type B is necessary for specific coupling to Galpha 13. *J. Biol. Chem.*, **278**(4):2384–2387.

Locke, D. P., Hillier, L. W., Warren, W. C., Worley, K. C., Nazareth, L. V., Muzny, D. M., Yang, S. P., Wang, Z., Chinwalla, A. T., Minx, P., *et al.*, 2011. Comparative and demographic analysis of orang-utan genomes. *Nature*, **469**(7331):529–533.

Loh, Y. H., Bezault, E., Muenzel, F. M., Roberts, R. B., Swofford, R., Barluenga, M., Kidd, C. E., Howe, A. E., Di Palma, F., Lindblad-Toh, K., *et al.*, 2013. Origins of shared genetic variation in African cichlids. *Mol. Biol. Evol.*, **30**(4):906–917.

Lonsdale, J., Thomas, J., Salvatore, M., Phillips, R., Lo, E., Shad, S., Hasz, R., Walters, G., Garcia, F., Young, N., *et al.*, 2013. The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.*, **45**(6):580–585.

Lovci, M. T., Ghanem, D., Marr, H., Arnold, J., Gee, S., Parra, M., Liang, T. Y., Stark, T. J., Gehman, L. T., Hoon, S., *et al.*, 2013. Rbfox proteins regulate alternative mRNA splicing through evolutionarily conserved RNA bridges. *Nat. Struct. Mol. Biol.*, **20**(12):1434–1442.

Lunter, G. and Goodson, M., 2011. Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads. *Genome Res.*, **21**(6):936–939.

Lynch, M., 2007. *The origins of genome architecture*. Sinauer Associates, Sunderland, MA.

Lyson, T. R. and Joyce, W. G., 2012. Evolution of the turtle bauplan: the topological relationship of the scapula relative to the ribcage. *Biol Lett*, **8(6)**:1028–31.

Macaulay, I. C. and Voet, T., 2014. Single cell genomics: advances and future perspectives. *PLoS Genet.*, **10**(1):e1004126.

Mackay, T. F., Richards, S., Stone, E. A., Barbadilla, A., Ayroles, J. F., Zhu, D., Casillas, S., Han, Y., Magwire, M. M., Cridland, J. M., *et al.*, 2012. The Drosophila melanogaster Genetic Reference Panel. *Nature*, **482**(7384):173–178.

Marinov, G. K., Williams, B. A., McCue, K., Schroth, G. P., Gertz, J., Myers, R. M., and Wold, B. J., 2014. From single-cell to cell-pool transcriptomes: stochasticity in gene expression and RNA splicing. *Genome Res.*, **24**(3):496–510.

Mayr, E., 1942. *Systematics and the Origin of Species*. Harvard University Press, Sunderland, MA.

Mazin, P., Xiong, J., Liu, X., Yan, Z., Zhang, X., Li, M., He, L., Somel, M., Yuan, Y., Phoebe Chen, Y. P., *et al.*, 2013. Widespread splicing changes in human brain development and aging. *Mol. Syst. Biol.*, **9**:633.

McBride, C. S., 2007. Rapid evolution of smell and taste receptor genes during host specialization in drosophila sechellia. *Proc Natl Acad Sci U S A*, **104**(12):4996–5001.

McVicker, G., van de Geijn, B., Degner, J. F., Cain, C. E., Banovich, N. E., Raj, A., Lewellen, N., Myrthil, M., Gilad, Y., and Pritchard, J. K., *et al.*, 2013. Identification of genetic variants that affect histone modifications in human cells. *Science*, **342**(6159):747–749.

Meader, S., Ponting, C. P., and Lunter, G., 2010. Massive turnover of functional sequence in human and other mammalian genomes. *Genome Res.*, **20**(10):1335–1343.

Merkin, J., Russell, C., Chen, P., and Burge, C. B., 2012. Evolutionary dynamics of gene and isoform regulation in Mammalian tissues. *Science*, **338**(6114):1593–1599.

Messer, P. W. and Petrov, D. A., 2013. Frequent adaptation and the McDonald-Kreitman test. *Proc. Natl. Acad. Sci. U.S.A.*, **110**(21):8615–8620.

Meunier, J. and Duret, L., 2004. Recombination drives the evolution of gc-content in the human genome. *Mol Biol Evol*, **21**(6):984–990.

Meyer, A. and Schartl, M., 1999. Gene and genome duplications in vertebrates: the one-to-four (-to-eight in fish) rule and the evolution of novel gene functions. *Curr. Opin. Cell Biol.*, **11**(6):699–704.

Modrek, B. and Lee, C. J., 2003. Alternative splicing in the human, mouse and rat genomes is associated with an increased frequency of exon creation and/or loss. *Nat. Genet.*, **34**(2):177–180.

Moffett, S., Mouillac, B., Bonin, H., and Bouvier, M., 1993. Altered phosphorylation and desensitization patterns of a human beta 2-adrenergic receptor lacking the palmitoylated Cys341. *EMBO J.*, **12**(1):349–356.

Montavon, T., Soshnikova, N., Mascrez, B., Joye, E., Thevenet, L., Splinter, E., de Laat, W., Spitz, F., and Duboule, D., 2011. A regulatory archipelago controls Hox genes transcription in digits. *Cell*, **147**(5):1132–1145.

Montgomery, S. B., Sammeth, M., Gutierrez-Arcelus, M., Lach, R. P., Ingle, C., Nisbett, J., Guigo, R., and Dermitzakis, E. T., 2010. Transcriptome genet-

ics using second generation sequencing in a Caucasian population. *Nature*, **464**(7289):773–777.

Mortazavi, A., Schwarz, E. M., Williams, B., Schaeffer, L., Antoshechkin, I., Wold, B. J., and Sternberg, P. W., 2010. Scaffolding a Caenorhabditis nematode genome with RNA-seq. *Genome Res.*, **20**(12):1740–1747.

Nagashima, H., Kuraku, S., Uchida, K., Ohya, Y. K., Narita, Y., and Kuratani, S., 2007. On the carapacial ridge in turtle embryos: its developmental origin, function and the chelonian body plan. *Development*, **134**(12):2219–2226.

Nagashima, H., Sugahara, F., Takechi, M., Ericsson, R., Kawashima-Ohya, Y., Narita, Y., and Kuratani, S., 2009. Evolution of the turtle body plan by the folding and creation of new muscle connections. *Science*, **325**(5937):193–196.

Nawy, T., 2014. Single-cell sequencing. *Nat. Methods*, **11**(1):18.

Near, T. J., Meylan, P. A., and Shaffer, H. B., 2005. Assessing concordance of fossil calibration points in molecular clock studies: an example using turtles. *Am Nat*, **165**(2):137–146.

Nellaker, C., Keane, T. M., Yalcin, B., Wong, K., Agam, A., Belgard, T. G., Flint, J., Adams, D. J., Frankel, W. N., and Ponting, C. P., *et al.*, 2012. The genomic landscape shaped by selection on transposable elements across 18 mouse strains. *Genome Biol.*, **13**(6):R45.

Nielsen, R., Williamson, S., Kim, Y., Hubisz, M. J., Clark, A. G., and Bustamante, C., 2005. Genomic scans for selective sweeps using SNP data. *Genome Res.*, **15**(11):1566–1575.

Nikolaidis, N., Makalowska, I., Chalkia, D., Makalowski, W., Klein, J., and Nei, M., 2005. Origin and evolution of the chicken leukocyte receptor complex. *Proc Natl Acad Sci U S A*, **102**(11):4057–4062.

Ohno, S., 1970. *The evolution by gene duplication*. Springer-Verlag, Berlin.

Ohta, T., 1973. Slightly deleterious mutant substitutions in evolution. *Nature*, **246**(5428):96–98.

Okamoto, Y., Ninomiya, H., Tanioka, M., Sakamoto, A., Miwa, S., and Masaki, T., 1998. Cysteine residues in the carboxyl terminal domain of the endothelin-B receptor are required for coupling with G-proteins. *J. Cardiovasc. Pharmacol.*, **31 Suppl 1**:S230–232.

Pan, Q., Bakowski, M. A., Morris, Q., Zhang, W., Frey, B. J., Hughes, T. R., and Blencowe, B. J., 2005. Alternative splicing of conserved exons is frequently species-specific in human and mouse. *Trends Genet.*, **21**(2):73–77.

Parichy, D. M., Mellgren, E. M., Rawls, J. F., Lopes, S. S., Kelsh, R. N., and Johnson, S. L., 2000. Mutational analysis of endothelin receptor b1 (rose) during neural crest and pigment pattern development in the zebrafish Danio rerio. *Dev. Biol.*, **227**(2):294–306.

Parker, J., Tsagkogeorga, G., Cotton, J. A., Liu, Y., Provero, P., Stupka, E., and Rossiter, S. J., 2013. Genome-wide signatures of convergent evolution in echolocating mammals. *Nature*, **502**(7470):228–231.

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

Parra, G., Bradnam, K., Ning, Z., Keane, T., and Korf, I., 2009. Assessing the gene space in draft genomes. *Nucleic Acids Res.*, **37**(1):289–297.

Payandeh, J., Gamal El-Din, T. M., Scheuer, T., Zheng, N., and Catterall, W. A., 2012. Crystal structure of a voltage-gated sodium channel in two potentially inactivated states. *Nature*, **486**(7401):135–139.

Pearson, J. C., Lemons, D., and McGinnis, W., 2005a. Modulating hox gene functions during animal body patterning. *Nat Rev Genet*, **6**(12):893–904.

Pearson, J. C., Lemons, D., and McGinnis, W., 2005b. Modulating Hox gene functions during animal body patterning. *Nat. Rev. Genet.*, **6**(12):893–904.

Pereira, S. L. and Baker, A. J., 2006a. A mitogenomic timescale for birds detects variable phylogenetic rates of molecular evolution and refutes the standard molecular clock. *Mol. Biol. Evol.*, **23**(9):1731–1740.

Pereira, S. L. and Baker, A. J., 2006b. A mitogenomic timescale for birds detects variable phylogenetic rates of molecular evolution and refutes the standard molecular clock. *Mol Biol Evol*, **23**(9):1731–1740.

Perry, G. H., Dominy, N. J., Claw, K. G., Lee, A. S., Fiegler, H., Redon, R., Werner, J., Villanea, F. A., Mountain, J. L., Misra, R., *et al.*, 2007. Diet and the evolution of human amylase gene copy number variation. *Nat. Genet.*, **39**(10):1256–1260.

Perry, G. H., Melsted, P., Marioni, J. C., Wang, Y., Bainer, R., Pickrell, J. K., Michelini, K., Zehr, S., Yoder, A. D., Stephens, M., *et al.*, 2012. Comparative RNA sequencing reveals substantial genetic variation in endangered primates. *Genome Res.*, **22**(4):602–610.

Peto, R., Roe, F. J., Lee, P. N., Levy, L., and Clack, J., 1975. Cancer and ageing in mice and men. *Br. J. Cancer*, **32**(4):411–426.

Pickrell, J. K., Marioni, J. C., Pai, A. A., Degner, J. F., Engelhardt, B. E., Nkadori, E., Veyrieras, J. B., Stephens, M., Gilad, Y., and Pritchard, J. K., *et al.*, 2010a. Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature*, **464**(7289):768–772.

Pickrell, J. K., Pai, A. A., Gilad, Y., and Pritchard, J. K., 2010b. Noisy splicing drives mRNA isoform diversity in human cells. *PLoS Genet.*, **6**(12):e1001236.

Piskol, R., Ramaswami, G., and Li, J. B., 2013. Reliable identification of genomic variants from RNA-seq data. *Am. J. Hum. Genet.*, **93**(4):641–651.

Pop, M. and Salzberg, S. L., 2008. Bioinformatics challenges of new sequencing technology. *Trends Genet.*, **24**(3):142–149.

Posada, D., 2008. jmodeltest: phylogenetic model averaging. *Mol Biol Evol*, **25**(7):1253–1256.

Prud'homme, B., Gompel, N., and Carroll, S. B., 2007. Emerging principles of regulatory evolution. *Proc. Natl. Acad. Sci. U.S.A.*, **104 Suppl 1**:8605–8612.

Prud'homme, B., Gompel, N., Rokas, A., Kassner, V. A., Williams, T. M., Yeh, S. D., True, J. R., and Carroll, S. B., 2006. Repeated morphological evolution through cis-regulatory changes in a pleiotropic gene. *Nature*, **440**(7087):1050–1053.

Prufer, K., Munch, K., Hellmann, I., Akagi, K., Miller, J. R., Walenz, B., Koren, S., Sutton, G., Kodira, C., Winer, R., *et al.*, 2012. The bonobo genome compared with the chimpanzee and human genomes. *Nature*, **486**(7404):527–531.

Ramskold, D., Luo, S., Wang, Y. C., Li, R., Deng, Q., Faridani, O. R., Daniels, G. A., Khrebtukova, I., Loring, J. F., Laurent, L. C., *et al.*, 2012. Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. *Nat. Biotechnol.*, **30**(8):777–782.

Rasmussen, M. D. and Kellis, M., 2007. Accurate gene-tree reconstruction by learning gene- and species-specific substitution rates across multiple complete genomes. *Genome Res.*, **17**(12):1932–1942.

Reisz, R. R. and Head, J. J., 2008. Palaeontology: Turtle origins out to sea. *Nature*, **456**(7221):450–451.

Rennison, D. J., Owens, G. L., and Taylor, J. S., 2012. Opsin gene duplication and divergence in ray-finned fish. *Mol Phylogenet Evol*, **62**(3):986–1008.

Reyes, A., Anders, S., Weatheritt, R. J., Gibson, T. J., Steinmetz, L. M., and Huber, W., 2013. Drift and conservation of differential exon usage across tissues in primate species. *Proc. Natl. Acad. Sci. U.S.A.*, **110**(38):15377–15382.

Ricardi, M. M., Gonzalez, R. M., Zhong, S., Dominguez, P. G., Duffy, T., Turjanski, P. G., Salgado Salter, J. D., Alleva, K., Carrari, F., Giovannoni, J. J., *et al.*, 2014. Genome-wide data (ChIP-seq) enabled identification of cell wall-related and aquaporin genes as targets of tomato ASR1, a drought stress-responsive transcription factor. *BMC Plant Biol.*, **14**:29.

Rice, P., Longden, I., and Bleasby, A., 2000. EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet*, **16**(6):276–7.

Robertson, H. M. and Wanner, K. W., 2006. The chemoreceptor superfamily in the honey bee, apis mellifera: expansion of the odorant, but not gustatory, receptor family. *Genome Res*, **16**(11):1395–1403.

Rodriguez-Lopez, A. M., Jackson, D. A., Nehlin, J. O., Iborra, F., Warren, A. V., and Cox, L. S., 2003. Characterisation of the interaction between WRN, the helicase/exonuclease defective in progeroid Werner's syndrome, and an essential replication factor, PCNA. *Mech. Ageing Dev.*, **124**(2):167–174.

Romero, P. R., Zaidi, S., Fang, Y. Y., Uversky, V. N., Radivojac, P., Oldfield, C. J., Cortese, M. S., Sickmeier, M., LeGall, T., Obradovic, Z., *et al.*, 2006. Alternative splicing in concert with protein intrinsic disorder enables increased

functional diversity in multicellular organisms. *Proc. Natl. Acad. Sci. U.S.A.*, **103**(22):8390–8395.

Ronquist, F., Teslenko, M., van der Mark, P., Ayres, D. L., Darling, A., Höhna, S., Larget, B., Liu, L., Suchard, M. A., and Huelsenbeck, J. P., *et al.*, 2012. Mrbayes 3.2: efficient bayesian phylogenetic inference and model choice across a large model space. *Syst Biol*, **61**(3):539–542.

Salzberg, S. L., Delcher, A. L., Kasif, S., and White, O., 1998. Microbial gene identification using interpolated Markov models. *Nucleic Acids Res.*, **26**(2):544–548.

Salzberg, S. L., Sommer, D. D., Puiu, D., and Lee, V. T., 2008. Gene-boosted assembly of a novel bacterial genome from very short reads. *PLoS Comput. Biol.*, **4**(9):e1000186.

Salzburger, W., Braasch, I., and Meyer, A., 2007. Adaptive sequence evolution in a color gene involved in the formation of the characteristic egg-dummies of male haplochromine cichlid fishes. *BMC Biol.*, **5**:51.

Sankararaman, S., Mallick, S., Dannemann, M., Prufer, K., Kelso, J., Paabo, S., Patterson, N., and Reich, D., 2014. The genomic landscape of Neanderthal ancestry in present-day humans. *Nature*, **507**(7492):354–357.

Sawyer, R. H., Rogers, L., Washington, L., Glenn, T. C., and Knapp, L. W., 2005. Evolutionary origin of the feather epidermis. *Dev Dyn*, **232**(2):256–267.

Sawyer, S., 1999. *GENECONV: A computer package for the statistical detection of gene conversion.* Department of Mathematics, Washington University in St. Louis.

Scally, A., Dutheil, J. Y., Hillier, L. W., Jordan, G. E., Goodhead, I., Herrero, J., Hobolth, A., Lappalainen, T., Mailund, T., Marques-Bonet, T., *et al.*, 2012.

Insights into hominid evolution from the gorilla genome sequence. *Nature*, **483**(7388):169–175.

Schluter, D., 2000. *The Ecology of Adaptive Radiation*. Oxford University Press, Berlin.

Schwartz, S., Meshorer, E., and Ast, G., 2009. Chromatin organization marks exon-intron structure. *Nat. Struct. Mol. Biol.*, **16**(9):990–995.

Shaffer, H. B., Minx, P., Warren, D. E., Shedlock, A. M., Thomson, R. C., Valenzuela, N., Abramyan, J., Amemiya, C. T., Badenhorst, D., Biggar, K. K., *et al.*, 2013. The western painted turtle genome, a model for the evolution of extreme physiological adaptations in a slowly evolving lineage. *Genome Biol.*, **14**(3):R28.

Shalek, A. K., Satija, R., Adiconis, X., Gertner, R. S., Gaublomme, J. T., Raychowdhury, R., Schwartz, S., Yosef, N., Malboeuf, C., Lu, D., *et al.*, 2013. Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. *Nature*, **498**(7453):236–240.

Shapiro, M. D., Marks, M. E., Peichel, C. L., Blackman, B. K., Nereng, K. S., Jonsson, B., Schluter, D., and Kingsley, D. M., 2004. Genetic and developmental basis of evolutionary pelvic reduction in threespine sticklebacks. *Nature*, **428**(6984):717–723.

Sharon, D., Tilgner, H., Grubert, F., and Snyder, M., 2013. A single-molecule long-read survey of the human transcriptome. *Nat. Biotechnol.*, **31**(11):1009–1014.

Shin, M. K., Russell, L. B., and Tilghman, S. M., 1997. Molecular characterization of four induced alleles at the Ednrb locus. *Proc. Natl. Acad. Sci. U.S.A.*, **94**(24):13105–13110.

Siepel, A., Bejerano, G., Pedersen, J. S., Hinrichs, A. S., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L. W., Richards, S., *et al.*, 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.*, **15**(8):1034–1050.

Simakov, O., Marletaz, F., Cho, S. J., Edsinger-Gonzales, E., Havlak, P., Hellsten, U., Kuo, D. H., Larsson, T., Lv, J., Arendt, D., *et al.*, 2013. Insights into bilaterian evolution from three spiralian genomes. *Nature*, **493**(7433):526–531.

Simpson, C. G., Hedley, P. E., Watters, J. A., Clark, G. P., McQuade, C., Machray, G. C., and Brown, J. W., 2000. Requirements for mini-exon inclusion in potato invertase mRNAs provides evidence for exon-scanning interactions in plants. *RNA*, **6**(3):422–433.

Sims, D., Sudbery, I., Ilott, N. E., Heger, A., and Ponting, C. P., 2014. Sequencing depth and coverage: key considerations in genomic analyses. *Nat. Rev. Genet.*, **15**(2):121–132.

Smith, J. J., Kuraku, S., Holt, C., Sauka-Spengler, T., Jiang, N., Campbell, M. S., Yandell, M. D., Manousaki, T., Meyer, A., Bloom, O. E., *et al.*, 2013. Sequencing of the sea lamprey (Petromyzon marinus) genome provides insights into vertebrate evolution. *Nat. Genet.*, **45**(4):415–421.

Smith, T. F. and Waterman, M. S., 1981. Identification of common molecular subsequences. *J. Mol. Biol.*, **147**(1):195–197.

Soding, J., Biegert, A., and Lupas, A. N., 2005. The HHpred interactive server for protein homology detection and structure prediction. *Nucleic Acids Res.*, **33**(Web Server issue):W244–248.

Solomon, S. E., Hendrickson, J. R., and Hendrickson, L. P., 1986. The structure

of the carapace and plastron of juvenile turtles, chelonia mydas (the green turtle) and caretta caretta (the loggerhead turtle). *J Anat*, **145**:123–131.

Sorek, R., Shamir, R., and Ast, G., 2004. How prevalent is functional alternative splicing in the human genome? *Trends Genet.*, **20**(2):68–71.

Stamatakis, A., 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, **30**(9):1312–1313.

Stanke, M., Keller, O., Gunduz, I., Hayes, A., Waack, S., and Morgenstern, B., 2006. AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Res.*, **34**(Web Server issue):W435–439.

Stoltzfus, A., 1999. On the possibility of constructive neutral evolution. *J. Mol. Evol.*, **49**(2):169–181.

Surget-Groba, Y. and Montoya-Burgos, J. I., 2010. Optimization of de novo transcriptome assembly from next-generation sequencing data. *Genome Res.*, **20**(10):1432–1440.

Takeda, J., Suzuki, Y., Sakate, R., Sato, Y., Seki, M., Irie, T., Takeuchi, N., Ueda, T., Nakao, M., Sugano, S., *et al.*, 2008. Low conservation and species-specific evolution of alternative splicing in humans and mice: comparative genomics analysis using well-annotated full-length cDNAs. *Nucleic Acids Res.*, **36**(20):6386–6395.

Teshima, K. M. and Innan, H., 2004. The effect of gene conversion on the divergence between duplicated genes. *Genetics*, **166**(3):1553–1560.

Theis, A., Salzburger, W., and Egger, B., 2012. The function of anal fin egg-spots in the cichlid fish Astatotilapia burtoni. *PLoS ONE*, **7**(1):e29878.

Toni, M., Valle, L. D., and Alibardi, L., 2007. Hard (beta-)keratins in the epidermis of reptiles: composition, sequence, and molecular organization. *J Proteome Res*, **6**(9):3377–3392.

Tournamille, C., Colin, Y., Cartron, J. P., and Le Van Kim, C., 1995. Disruption of a GATA motif in the Duffy gene promoter abolishes erythroid gene expression in Duffy-negative individuals. *Nat. Genet.*, **10**(2):224–228.

Trapnell, C., Pachter, L., and Salzberg, S. L., 2009. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics*, **25**(9):1105–1111.

Tsai, I. J., Zarowiecki, M., Holroyd, N., Garciarrubio, A., Sanchez-Flores, A., Brooks, K. L., Tracey, A., Bobes, R. J., Fragoso, G., Sciutto, E., *et al.*, 2013. The genomes of four tapeworm species reveal adaptations to parasitism. *Nature*, **496**(7443):57–63.

Vandebergh, W. and Bossuyt, F., 2012. Radiation and functional diversification of alpha keratins during early vertebrate evolution. *Mol Biol Evol*, **29**(3):995–1004.

Vavouri, T., Semple, J. I., Garcia-Verdugo, R., and Lehner, B., 2009. Intrinsic protein disorder and interaction promiscuity are widely associated with dosage sensitivity. *Cell*, **138**(1):198–208.

Venkatesh, B., Kirkness, E. F., Loh, Y. H., Halpern, A. L., Lee, A. P., Johnson, J., Dandona, N., Viswanathan, L. D., Tay, A., Venter, J. C., *et al.*, 2007. Survey sequencing and comparative analysis of the elephant shark (Callorhinchus milii) genome. *PLoS Biol.*, **5**(4):e101.

Vilella, A. J., Severin, J., Ureta-Vidal, A., Heng, L., Durbin, R., and Birney, E., 2009. EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. *Genome Res.*, **19**(2):327–335.

Voskoboynik, A., Neff, N. F., Sahoo, D., Newman, A. M., Pushkarev, D., Koh, W., Passarelli, B., Fan, H. C., Mantalas, G. L., Palmeri, K. J., *et al.*, 2013. The genome sequence of the colonial chordate, Botryllus schlosseri. *Elife*, **2**:e00569.

Wagner, C. E., Harmon, L. J., and Seehausen, O., 2012. Ecological opportunity and sexual selection together predict adaptive radiation. *Nature*, **487**(7407):366–369.

Wagner, C. E., Keller, I., Wittwer, S., Selz, O. M., Mwaiko, S., Greuter, L., Sivasundar, A., and Seehausen, O., 2013. Genome-wide RAD sequence data provide unprecedented resolution of species boundaries and relationships in the Lake Victoria cichlid adaptive radiation. *Mol. Ecol.*, **22**(3):787–798.

Wang, Z., Gerstein, M., and Snyder, M., 2009. RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.*, **10**(1):57–63.

Wang, Z., Pascual-Anaya, J., Zadissa, A., Li, W., Niimura, Y., Huang, Z., Li, C., White, S., Xiong, Z., Fang, D., *et al.*, 2013. The draft genomes of soft-shell turtle and green sea turtle yield insights into the development and evolution of the turtle-specific body plan. *Nat. Genet.*, **45**(6):701–706.

Warren, W. C., Clayton, D. F., Ellegren, H., Arnold, A. P., Hillier, L. W., Künstner, A., Searle, S., White, S., Vilella, A. J., Fairley, S., *et al.*, 2010. The genome of a songbird. *Nature*, **464**(7289):757–762.

Weischenfeldt, J., Waage, J., Tian, G., Zhao, J., Damgaard, I., Jakobsen, J. S., Kristiansen, K., Krogh, A., Wang, J., and Porse, B. T., *et al.*, 2012. Mammalian tissues defective in nonsense-mediated mRNA decay display highly aberrant splicing patterns. *Genome Biol.*, **13**(5):R35.

Weyn-Vanhentenryck, S. M., Mele, A., Yan, Q., Sun, S., Farny, N., Zhang, Z., Xue, C., Herre, M., Silver, P. A., Zhang, M. Q., *et al.*, 2014. HITS-CLIP

and integrative modeling define the Rbfox splicing-regulatory network linked to brain development and autism. *Cell Rep*, **6**(6):1139–1152.

Williams, L. J., Tabbaa, D. G., Li, N., Berlin, A. M., Shea, T. P., Maccallum, I., Lawrence, M. S., Drier, Y., Getz, G., Young, S. K., *et al.*, 2012. Paired-end sequencing of Fosmid libraries by Illumina. *Genome Res.*, **22**(11):2241–2249.

Wittkopp, P. J., Haerum, B. K., and Clark, A. G., 2004. Evolutionary changes in cis and trans gene regulation. *Nature*, **430**(6995):85–88.

Wu, J., Anczukow, O., Krainer, A. R., Zhang, M. Q., and Zhang, C., 2013a. OLego: fast and sensitive mapping of spliced mRNA-Seq reads using small seeds. *Nucleic Acids Res.*, **41**(10):5149–5163.

Wu, L., Candille, S. I., Choi, Y., Xie, D., Jiang, L., Li-Pook-Than, J., Tang, H., and Snyder, M., 2013b. Variation and genetic control of protein abundance in humans. *Nature*, **499**(7456):79–82.

Wu, T. D. and Nacu, S., 2010. Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics*, **26**(7):873–881.

Xing, J., Wang, H., Belancio, V. P., Cordaux, R., Deininger, P. L., and Batzer, M. A., 2006. Emergence of primate genes by retrotransposon-mediated sequence transduction. *Proc. Natl. Acad. Sci. U.S.A.*, **103**(47):17608–17613.

Xing, Y. and Lee, C., 2006. Alternative splicing and RNA selection pressure–evolutionary consequences for eukaryotic genomes. *Nat. Rev. Genet.*, **7**(7):499–509.

Xue, Y., Ouyang, K., Huang, J., Zhou, Y., Ouyang, H., Li, H., Wang, G., Wu, Q., Wei, C., Bi, Y., *et al.*, 2013. Direct conversion of fibroblasts to neurons by reprogramming PTB-regulated microRNA circuits. *Cell*, **152**(1-2):82–96.

Yang, Z., 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol*, **24**(8):1586–91.

Yang, Z. and dos Reis, M., 2011. Statistical properties of the branch-site test of positive selection. *Mol. Biol. Evol.*, **28**(3):1217–1228.

Yeo, G. and Burge, C. B., 2004. Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *J. Comput. Biol.*, **11**(2-3):377–394.

Yim, H. S., Cho, Y. S., Guang, X., Kang, S. G., Jeong, J. Y., Cha, S. S., Oh, H. M., Lee, J. H., Yang, E. C., Kwon, K. K., *et al.*, 2014. Minke whale genome and aquatic adaptation in cetaceans. *Nat. Genet.*, **46**(1):88–92.

Yuan, F., Bernard, G. D., Le, J., and Briscoe, A. D., 2010. Contrasting modes of evolution of the visual pigments in heliconius butterflies. *Mol Biol Evol*, **27**(10):2392–2405.

Zerbino, D. R. and Birney, E., 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res.*, **18**(5):821–829.

Zhang, C., Zhang, Z., Castle, J., Sun, S., Johnson, J., Krainer, A. R., and Zhang, M. Q., 2008. Defining the regulatory network of the tissue-specific splicing factors Fox-1 and Fox-2. *Genes Dev.*, **22**(18):2550–2563.

Zhang, G., Cowled, C., Shi, Z., Huang, Z., Bishop-Lilly, K. A., Fang, X., Wynne, J. W., Xiong, Z., Baker, M. L., Zhao, W., *et al.*, 2013. Comparative analysis of bat genomes provides insight into the evolution of flight and immunity. *Science*, **339**(6118):456–460.

Zhang, J., Webb, D. M., and Podlaha, O., 2002. Accelerated protein evolution and origins of human-specific features: Foxp2 as an example. *Genetics*, **162**(4):1825–1835.

Zhu, L., Zhang, Y., Zhang, W., Yang, S., Chen, J. Q., and Tian, D., 2009. Patterns of exon-intron architecture variation of genes in eukaryotic genomes. *BMC Genomics*, **10**:47.

Zibetti, C., Adamo, A., Binda, C., Forneris, F., Toffolo, E., Verpelli, C., Ginelli, E., Mattevi, A., Sala, C., and Battaglioli, E., *et al.*, 2010. Alternative splicing of the histone demethylase LSD1/KDM1 contributes to the modulation of neurite morphogenesis in the mammalian nervous system. *J. Neurosci.*, **30**(7):2521–2532.

Zong, C., Lu, S., Chapman, A. R., and Xie, X. S., 2012. Genome-wide detection of single-nucleotide and copy-number variations of a single human cell. *Science*, **338**(6114):1622–1626.