Characterising selection in Conserved Noncoding Elements (CNEs)

Dilrini R. De Silva

Submitted in partial fulfillment of the requirements of the Degree of Doctor of Philosophy

QUEEN MARY UNIVERSITY OF LONDON

Declaration

This dissertation is submitted for the degree of Doctor of Philosophy at the University of London.

This dissertation is the result of my own work and includes nothing of which is the outcome of work done in collaboration except where specifically indicated.

No part of this dissertation is being submitted for any other qualification or at any other university.

Dilrini De Silva

October 2013

Acknowledgements

I would like to extend my gratitude to my co-supervisors Greg Elgar (NIMR) and Richard Nichols (QMUL) firstly, for condoning my galavants across the globe during my time as a PhD student in addition to entertaining my request to spend the summer of 2012 volunteering at the London 2012 Olympic Games – it certainly enhanced my experience in the UK. More importantly, thank you for your support and guidance right throughout - I will strive to emulate your work ethic and leadership in my career.

I am grateful to Queen Mary University of London for my Graduate Teaching Studentship award that funded this PhD research. I am thankful to members of both labs for creating an enjoyable, productive working environment. I thank Paul Piccinelli for helping me get up to speed with programming in Perl during my first year, Robert Verity for helpful discussions and guidance with mathematical models when I needed it the most, Boris Noyvert for lending me a laptop cooler to keep my laptop alive until submission (as well as access to snacks from his locker when I worked late), Stefan Pauls for useful formatting tips and Joseph Price for lending me a USB stick countless times. I also thank Alexandros Stamatakis at the Heidelberg Institute of Theoretical Studies (HITS) for giving me the opportunity to pursue an internship at their lab, and the members of his lab for their kind support.

I extend special thanks to my friends for their generosity. I thank Chesmal Siriwardhana for insightful discussions and sharing a homemade meal with a hungry, exhausted PhD student, Shaun Choquet for an endless supply of Starbucks cookies, coffee, entertaining conversations and encouragement, and Michela Veronese for looking out for my wellbeing and spoiling me with delicious Italian cooking during her brief stay in London.

Finally, I thank my resilient family - my parents, Alex and Rose, for the sacrifices they made to put education first and their words of encouragement in every telephone conversation over the past few years. This work is especially dedicated to my sister, Gayathri De Silva, who took two years out of her own education to care for our ailing parents so I could finish what I started. You're a star.

An ode to a PhD student

Dilly the Merciful
O blessed are thou eyes
And may they look at us, unworthy

Dilly the Untainted
Thou dwells in rarified heights
May thou bow and descend to us, oh fallen

Dilly the Mindful
Forceful is the clarity of thy purpose
May we partake in thy mastery, us uncalled

Dilly the Eternal
Few have seen, fewer shall know
May thou reveal to us thy presence, oh forgotten

- Shaun Choquet (2013)

Abstract

Comparative genomic studies have identified noncoding regions of the genome which are often more highly conserved between species than protein-coding sequences. One possible explanation for this conservation of non-coding sequences is some form of selective constraint since sequence conservation at great evolutionary depths is a preliminary indication of functional constraint. Here, I consider nearly 2500 putative regulatory elements, termed Conserved Noncoding Elements (CNEs), that are conserved across seven vertebrate species (human, macaque, mouse, chicken, frog, zebrafish and fugu). I distinguish between CNEs that show accelerated rates of evolution and those that have remained more constrained throughout evolution, and identify CNEs that show higher than expected substitution rates in the human lineage that may be potential candidates of adaptive evolution. However, it is not trivial to demonstrate the action of selection on such sequences. It is relatively easier in the case of protein-coding DNA, since selection would be predicted to result in different rates of substitution for synonymous and non-synonymous sites. Hence, I use the same seven species to define phylogenetically invariant positions in CNEs in contrast to those that have at least one substitution and analyse them independently to determine if there is a positive correlation between evolutionary conservation and the strength of purifying selection at individual sites. In the 1000 Genomes, but not the HapMap, data I find a significant excess of rare derived alleles in CNEs relative to coding sequences. This excess of rare alleles can be best explained if selection is relatively consistent across sites, with most mutations resulting in a similar reduction in fitness. Finally, I explore patterns of variation in the allele-frequencies within human populations, however do not detect any significant differences in the underlying distribution of negatively selected variants among human populations.

Table of Contents

Characterising selection in Conserved Noncoding Elements (CNEs)

Chapter 1

1. Literature Review

The discovery of the double helical structure of DNA (Watson and Crick, 1953) and advances in recombinant DNA technology (Maxam and Gilbert, 1977; Sanger et al., 1977) that allowed DNA to be sequenced in a scalable fashion heralded the dawn of the Human Genome Project in 1990 and with it, the genomic era. The main aim of this coordinated international effort was to produce a map of the entire human genome to further our understanding of the genetic factors responsible for diseases that affect the lives of millions. The advent of the Internet, which facilitated data access and sharing among international collaborating centres, was especially instrumental in the progress made within the stipulated time. The human genome was drafted (International Human Genome Sequencing Consortium, 2001), completed (International Human Genome Sequencing Consortium, 2003) and, a decade later, we are closer to realising those aims. Personalised genomics, where treatment and therapy can be tailored according one's genetic makeup, is revolutionising the healthcare industry with an immediate impact on public policy pertaining to data access/sharing and ethical issues.

The most radical innovations in this sector have resulted in the ascent of sequencing technologies. The throughput of first generation sequencing technology (automated

Sanger sequencing) was limited by the need to isolate and amplify individual clones, labelling with fluorescent dyes and separating cloned fragments by gel electrophoresis.

The advent of next generation sequencing (NGS) by Roche/454, Life Technologies SOLiD and Illumina, has since reduced the time taken to sequence a human genome dramatically because theyeliminate the tedious step of cloning DNA in bacteria. The main advantage of 454 sequencing over other NGS methods is its capacity to generate long reads (> 250bp initially, now approaching 1kbp with recent technical refinements) but is known to be prone to errors that introduce insertions and deletions (indels) into the sequence (Hutchison, 2007). SOLiD and Illumina sequencing platforms are capable of producing a higher number of sequence reads albeit generally of shorter length (up to 150bp initially, although Illumina MiSeq now generates reads of up to 300bp comparable to 454 technology) (Metzker, 2010). However, the higher throughput of these massively parallel sequencing technologies does not translate to greater accuracy - platform-specific biases are introduced owing to the chemistry behind NGS methods and should be taken into account during data processing. For example Illumina chemistry which currently dominates the NGS market, has a relatively higher error rate calling a base after 'G' and also tends toward a higher read-density in GC rich regions (Dohm et al., 2008). Hence some sequencing techniques may be better suited for re-sequencing rather than *de novo* genome assembly.

Significant improvements across multiple disciplines such as molecular biology, chemistry and engineering have amalgamated to advance the technology to where it is today. So much so that the limiting factor in genomic analysis is no longer generating sequence data, but bioinformatics, i.e. data analysis and interpretation (Mardis, 2011).

Emerging platforms by Oxford Nanopore and IBM/Roche that use nanopore technology where nucleotides of a DNA strand are read as they pass through a protein nanopore, and Pacific Biosciences that use single molecule real time (SMRT) sequencing where incorporation of individual nucleotides by the DNA polymerase enzyme is detected, are expected to generate an unprecedented amount of data and exacerbate this bioinformatics bottleneck even further.

Additionally, IT infrastructure capable of handling computer memory-intensive analytical tasks have been integrated from other disciplines to deal with the large quantities of data being generated, which have reached terabytes in volume (Yap et al., 1996). High-performance computing (HPC) clusters, traditionally used in physics and engineering, are increasingly commonplace in mainstream bioinformatics (Collins et al., 2003; Bader, 2004). For example, algorithms such as Cloudburst (Schatz, 2009) for mapping NGS reads and variant discovery, and CloudBLAST, a parallelised version of the NCBI BLAST2 algorithm (Matsunaga et al., 2008), utilise the Hadoop architecture for large scale parallelised data analysis and are available through cloud-based services (Taylor, 2010).

Furthermore, clusters of Graphics Processing Units (GPUs), originally used in the video game industry, also add to this multidisciplinary approach where their massively parallelised architecture and high performance relative to conventional CPUs (Central Processing Units), is exploited for example to analyse genome-wide epistatic interactions in multifactorial diseases (Sinnott-Armstrong et al., 2009). The bioinformatics community has also benefitted from a multitude of software modules for data retrieval and analysis from the open-source community including BioPerl (Stajich

et al., 2002), BioPython (Cock et al., 2009), BioRuby (Goto et al., 2010) and BioJava (Prlic et al., 2012) etc. The growing volume of data generated by third-generation sequencing technologies such as SMRT, especially if sequencing cancer genomes that require twice as much (if not more) capacity due to the need for tumour and normal genomes from the same individual, will only push the envelope for even more innovative multidisciplinary solutions for data storage, transfer, access and analysis (Schadt et al. 2010).

1.1. Comparative genomics and Conserved Noncoding Elements (CNEs)

Since the Human Genome Project, the genomes of a number of different species have been sequenced. These datasets have opened up the exciting field of comparative genomics for studying models of human disease in other organisms as well as understanding our evolution. The current Ensembl repository (Hubbard et al., 2002; Flicek et al., 2013) hosts over 50 vertebrate genomes and 17 more are available in draft format on Pre!Ensembl. The pufferfish (*Takifugu rubripes*) with a genome 7.5 times smaller than the human genome presented the ideal model vertebrate organism to discover human genes since it consisted of a similar gene repertoire to humans (Brenner et al., 1993, Elgar et al., 1996). Additionally, the fugu genome proved useful in identifying putative regulatory regions in mammalian genomes by virtue of sequence conservation. For example, Aparicio et al. (1995) compared the Hoxb-4 region between fugu and mouse and identified regulatory elements conserved in both species. Importantly, the homologous fugu regulatory sequences could drive enhancer activity in a mammal experimental system, transgenic mice; demonstrating that the sequences had retained their regulatory function.

There may be many uncharacterised regulatory regions distributed around the genome. Once the noncoding regions (regions that are not annotated structural loci) of the fugu draft genome assembly was compared with the human genome sequence, over a thousand conserved noncoding elements (CNEs) with an average percent identity of 85% over at least 100bp in length were identified (Woolfe et al., 2005). The level of sequence conservation in some instances is exceptional: some CNEs are more highly conserved between the two species (>90% identity over ~500 bp) than most coding regions. Their high degree of sequence conservation suggests purifying selection to preserve some function, but further evidence is needed to verify such interpretations. In many instances the human element mapped to two locations on the fugu genome. These double-matches are thought to be explained by an event in the evolutionary history of teleosts, when their ancestral lineage underwent whole-genome duplication. In these cases the longest match in the fugu genome was used as the comparison for subsequent analysis (Woolfe et al., 2005).

Further fish-mammal comparisons including human, mouse, rat, dog and fugu (as the baseline organism) identified a complement of approximately 7000 CNEs with 65% identity over 40bp, which are catalogued in the COnserved NoncoDing ORthologous Regions (CONDOR) database along with functional data from in-vivo enhancer assays in zebrafish embryos for a subset of the elements (Woolfe et al., 2007). The authors used a synteny map created using whole genome comparisons of noncoding human and fugu genomes to define boundaries in which to look for clusters of CNEs. The boundaries of syntenic CNE clusters were then extended out in each direction to the nearest non-syntenic gene. Orthologous CNEs in divergent vertebrates (mouse, rat and dog) were

identified using these extended syntenic regions, thus avoiding spurious matches. CNEs overlapping known exons and noncoding RNAs were excluded although CNEs in UTRs (untranslated regions) were retained.

An interesting feature of CNEs is that they are not uniformly distributed across the genome. Most notably, CNEs are clustered around transcriptional-regulation and developmental genes (*trans-dev* genes). Moreover, although genome sizes vary across the vertebrate species, the order of CNEs relative to their associated genes remains conserved, suggesting such an association is important for function. For instance, the orientation of several highly conserved noncoding elements is retained in addition to perfect synteny of 16 genes around the SHH locus in human and fugu. Goode et al. (2005) demonstrated that CNEs in this region are able to drive reporter-gene expression in tissues where endogenous expression would normally occur.

A clearer picture about the need to maintain synteny began to emerge from studies involving chromosomal aberrations that decouple a regulatory element from its target gene creating a 'position effect' (Kleinjan and van Heyningen, 1998). For example, Lettice et al., (2011) described the effect of a translocation that places the sonic hedgehog (SHH) gene under the influence of a highly conserved enhancer different to its own, which results in severe limb abnormalities. As regulators of transcription, such highly conserved *cis*-regulatory elements are potentially at the heart of complex interactions across several genes and form essential *cis*-regulatory modules (CRMs) within a larger framework of gene regulatory networks (GRNs) (Davidson et al., 2002).

The most likely mechanism of *cis*-regulation, by which these sequences mediate gene expression, is protein-DNA interactions involving transcription factors facilitated by

motifs harboured in the regulatory sequences (Harbison et al., 2004). However, transcription factor binding sites (TFBS) are short stretches of DNA (6-10bp) that are abundant across the genome and have a rapid rate of turnover (Dermitzakis et al., 2002; Moses et al., 2003; Moses et al., 2006). Therefore, there is no apparent reason for such deep phylogenetic conservation over distances of hundreds of bases unless theyhve a similar function to enhancers which are typically . A putative explanation is that of an overlap of binding sites of more than one transcription factor (discussed in Elgar and Vavouri, 2008) and proximal binding sites for to facilitate co-operative binding of different transcription factors (e.g. Négre et al., 2011), similar to functional enhancer elements, which are typically about 500bp in length (Loots, 2008).

A number of studies demonstrate that the CNE-associated *trans-dev* genes have conserved elements with *cis*-regulatory capacity (in their vicinity and beyond). For example, Spitz *et al*. (2003) identified a regulatory element conserved between human and fugu that has a regulatory domain capable of regulating several genes at the HOXD (homeo-box D) locus, which is required for proper development of limbs and genitalia in vertebrates. PAX6 (paired-box 6), involved in the development of the eye, has enhancer elements that are conserved between human and fugu; additionally, other elements are conserved only in mammals, possibly because they are responsible for regulatory patterns specific to mammals (Kleinjan et al., 2004).

In addition to being able to drive gene expression proximally, CNEs are also capable of driving gene expression distally. For example, the human SOX9 (Sry-related HMG box 9) gene is flanked by large intergenic sequences that harbour conserved long-range regulatory elements ~300 kb upstream (Bagheri-Fam et al., 2001). There can be intervening coding sequence between the CNE and the loci where expression is

affected; including conserved long-range enhancer elements which drive expression whilst embedded in the introns of neighbouring genes (Lettice et al., 2003). In one case, the authors demonstrate the importance of sequence conservation in the CNE: single base-pair changes in a regulatory element associated with the SHH gene are associated with the congenital abnormality preaxial polydactyly in humans. Numerous other instances of human disorders attributed to disruptions in conserved noncoding regulatory elements through various mechanisms have been reported (Kleinjan and van Heyningen, 2005). For example, an array of heritable and *de novo* microdeletions, point mutations and translocations in highly conserved sequences flanking the SOX9 locus cause cranio-facial abnormalities in the Pierre-Robin syndrome (Benko et al., 2009). Furthermore, a *de novo* deletion in the FOXL2 locus that deletes several highly conserved elements affects development of the eyelids and ovaries (D'haene et al., 2009). Yet, no systematic screen for such evolutionarily conserved noncoding regions currently exists in diagnosing/characterising human developmental disorders.

Elements related to vertebrate CNEs are not detectable in the genomes of invertebrates: Woolfe et al., (2005) could not find similar sequences in surveys of the genomes of *Ciona intestinalis*, *Drosophila melanogaster*, and *Caenorhabitis elegans*, even though the homologous gene complement is present. However, conserved noncoding elements unique to invertebrate species have been defined in *Drosophila* (Bergman and Kreitman, 2001), *Saccharomyces* species (Kellis et al., 2003), *Caenorhabditis elegans* (Vavouri et al., 2007) and *Ciona intestinalis* (Doglio et al., in press). The conservation of CNEs over such a large phylogenetic distance and their absence in invertebrates suggests these sequences are critical for vertebrate development and therefore subject to strong selection. Studies on conserved noncoding elements in mouse (Keightley et al., 2005)

and mammals (Drake et al., 2005) have indeed demonstrated that the conservation stems from strong purifying selection rather than mutational coldspots in the genome.

The search for regulatory regions using conservation criteria has also revealed ultraconserved elements (UCEs) longer than 200 bp that are 100% conserved across human, mouse and rat genomes (Bejerano et al., 2004), ultra-conserved regions (UCRs) conserved between human and pufferfish with 95% sequence identity over at least 50bp between human and mouse (Sandelin et al., 2004) and highly conserved elements (HCEs) that are longer but less well conserved across human, mouse, rat, chicken and fugu genomes (Siepel et al., 2005). Notably, these sequences (with the exception of UCRs) often overlap exons of known genes and are sometimes associated with genes other than *trans-dev* genes. Therefore, CNEs are representative of a less well-conserved noncoding regulatory landscape still recognisable in teleosts, which has shaped the fundamental aspects of vertebrate development for nearly 450 million years. With information from genomes of a multitude of different species and increasingly, a multitude of individuals within species, we are now in a position to further interrogate patterns that could reveal the forces that affect their evolution.

1.2. *Cis*-regulation and vertebrate evolution

It has been proposed that there have been two rounds of whole genome duplication events in ancestral vertebrates (Ohno, 1970), referred to as 1R and 2R, which have shaped the current vertebrate genomic landscape, developmental and morphological complexity. However, other lines of evidence, based on phylogenetic distribution and patterns of amino acid sequences in duplicated genes of *Drosophila* and vertebrates, suggest vertebrate genomes could also have arisen from local duplication events and

subsequent translocations as opposed to whole genome duplication events (Hughes, 1999). The HOX gene family, a family of homeobox-containing transcription factors that govern body patterning, has served as a paradigm for exploring the dynamics of vertebrate gene evolution including developmental regulation and whole genome duplication (Sharman and Holland, 1998). Invertebrates have a single cluster of HOX genes whereas vertebrates have at least two HOX clusters (Garcia-Fernàndez, 1994), attributed to duplication events after the split between vertebrates and invertebrates. Studies on HOX gene clusters have also supported the idea of an additional teleost-specific genome duplication event referred to as 3R (Figure 1) (Aparicio et al., 1997; Amores et al., 1998; Meyer and Van de Peer, 2005). Nonetheless phylogenetic analysis of gene clusters including HOX genes suggest that duplication events across the different gene families are not necessarily simultaneous while some genes present evidence of duplication prior to the origin of vertebrates (Hughes et al., 2001). Asrar et al. (2013) also conclude that segmental duplications and chromosomal rearrangements at different time points could have given rise to the configuration of HOX clusters observed in vertebrates.

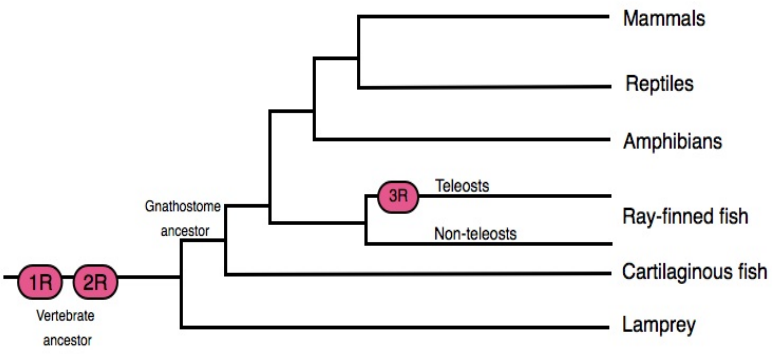

Figure 1: Vertebrate phylogeny depicting possible genome duplication events.

Whatever the exact cause of the duplication, the current-day genome must be interpreted in light of the evolution that has taken place after the duplication event. The "duplication–degeneration–complementation" (DDC) model (Force et al. 1999) proposes three possible scenarios for the fate of paralogous genes post-duplication. The new paralog can be lost (nonfunctionalisation), diversify to acquire new functions due to functional redundancy of the duplicates (neofunctionalisation) or paralogs can diversify in a manner that requires both copies to be retained to maintain ancestral function (subfunctionalisation). Differences in expression patterns of duplicates that are retained have been attributed to changes in their regulatory sequences (Hughes & Hughes, 1993). Blomme et al. (2006) observed a bias for the retention of regulatory genes post-duplication and noted the same trend in plants and yeast. Comparisons between two paralogous Hox gene clusters in the teleosts fugu and zebrafish revealed asymmetry in the rate of evolution between paralogous clusters post-duplication resulting in accelerated evolution in one lineage but not the other (Wagner, 2004). A similar trend was also observed with a larger set of genes in teleosts (Steinke et al 2006). Wagner (2004) also noted that the asymmetry in the rate of evolution of these genes correlates with the asymmetry in the rate of evolution of conserved noncoding DNA (putative *cis*-regulatory elements) in the different lineages. Such a pattern could be explained if some genes have dosage effects and need to be copied/retained along with their regulatory elements to maintain functional gene regulatory networks (Teichmann & Babu, 2004).

Detailed studies of other CNEs also show evidence of coevolution of CNE and coding locus post duplication. Comparisons between human and pufferfish genomes identified a large proportion of duplicated CNEs (dCNEs) in the human genome around transcriptional regulation and developmental genes that are retained in the same

orientation and relative position in the respective genomes contrary to surrounding genomic regions (McEwen et al., 2006). The authors observed that the human dCNEs are also duplicated in a number of other vertebrates and that their regulatory patterns partially overlap those of the ancestral sequence, such as seen with dCNEs associated with the FOX gene family. Further investigation of duplicated CNEs in the fugu genome resulting from the additional teleost-specific duplication event show evidence of putative regulatory subfunctionalisation and an asymmetric distribution of CNEs associated with duplicated genes (Woolfe and Elgar, 2007). Asymmetrical rates of evolution were observed in the duplicated copies when compared with the orthologous single copy in the human genome. In pairs of fugu dCNEs, one co-ortholog had degenerated at the edges while others had degenerated in the centre creating split elements, potentially altering the combination of binding sites available for transcription factor interaction.

Moreover, functional studies in duplicated fugu CNEs associated with PAX2 show that dCNEs can drive completely different patterns of expression both spatially and temporally despite a high level of sequence identity, emphasizing that in some instances, very few changes are required for major differences in expression (Goode et al., 2011).

Evidence is accumulating that similar regulatory interactions extend beyond the vertebrate lineage, and that comparable co-evolutionary dynamics take place. Over these longer evolutionary time periods it becomes difficult to identify homologous regulatory processes; this problem is apparent even with comparisons among distantly related vertebrate species: data from the recently sequenced genome of lamprey, an agnathan representing one of the earliest diverging lineages of jawless vertebrates, shows shared

identity across only ~53% of the length of gnathostome CNEs and a paucity of lamprey CNEs (Smith et al., 2013). These results might be interpreted as showing that CNEs became functionally constrained among vertebrates after the lamprey-gnathostome divergence, or that CNEs have evolved much more rapidly in the lamprey lineage.

In the case of invertebrates, there is evidence of *cis*-regulatory changes responsible for generating modified patterns of expression in orthologous genes in invertebrates, which have been well documented, yet they do not involve sequences with detectable homology to vertebrate CNEs. For example, differences in inter-specific gene expression in *Drosophila* are predominantly attributed to *cis*-regulatory changes (Wittkopp et al., 2004). Fewer than 10 point mutations in the same *cis*-regulatory element were identified to be the cause of parallel evolution of the same wing pattern in two species of *Drosophila* (Prud'Homme et al., 2006). Similarly, subtle differences in regulatory regions have been recognised as a major contributor to morphological innovations throughout vertebrate evolution, especially in limb and fin development (Abbassi et al., 2011). For instance, the long-range enhancer 'ShARE', which regulates the Shh gene responsible for limb development in gnathostomes, is notably absent in snakes and the lamprey, which lack paired-appendages (Smith et al., 2013).

Additionally, a variety of developmental changes are brought about by the evolution of *cis*-regulatory elements associated with protein-coding genes in vertebrates (reviewed by Wray, 2007), highlighting the importance of regulatory regions in the evolution of both vertebrates and invertebrates.

1.3. Phylogenetic comparisons to detect sequences constrained by selection

At first sight, it seems reasonable to assume that the parts of the genome that are highly conserved among species must be constrained by selection. There are however, other reasons why some regions of the genome might show higher or lower numbers of differences from other species: they include differences in the mutation rate, differences in the intensity of selection on linked loci or recombination rate (which changes the degree of linkage to selected loci), or variation in time back to a common ancestor – since different loci are descended from different ancestral individuals within the ancestral species (Nichols, 2001).

In the case of protein coding sequences it is possible to make use of our knowledge of the genetic code to allow for these problems. In the absence of selection, the effects of mutation rate, time to an ancestor and linkage to selected loci would be expected to affect synonymous and non-synonymous sites equally. Hence changes in relative rates of non-synonymous and synonymous substitutions can provide clues as to the action of selection. However, assessing divergence in noncoding DNA is complicated by the lack of a syntax analogous to the genetic code to illuminate changes in regulatory elements.

Two techniques widely used to overcome this difficulty are phylogenetic footprinting and phylogenetic shadowing. Footprinting makes comparisons of sequences across highly divergent species whereas shadowing makes comparisons across more closely related species. Sequences that have low substitution rates over millions of years of evolution, compared to the rest of the genome, are interpreted as having a large proportion of sites constrained by selection. Concerns over variation in the substitution

rate across the genome that generates mutational hotspots and coldspots can be discounted primarily because, over long evolutionary periods, changes are expected to have accumulated even at mutational coldspots in the absence of selective constraint. This approach has successfully identified sequences under functional constraint such as transcription factor binding sites and other regulatory sequences (Gumucio et al., 1993; Zhang & Gerstein, 2003; Ganley & Kobayashi, 2007). In such regions there can be a high variance in substitution rate, for example the substitution rate in the sequence encoding the stem-structure of primate rRNA genes (Rzhetsky, 1995) is very low relative to the rate in the loop.

Footprinting is less effective in identifying regions that have been constrained over shorter evolutionary periods, because of concerns over variation in substitution rate due to processes other than selective constraint outlined above, and simply because there may not have been enough time for differences to have accumulated in the unconstrained sites. Phylogenetic shadowing overcomes this problem by making comparisons across a large number of related species. The less constrained areas may not have accumulated substitutions in any one lineage, however if a whole clade of the phylogenetic tree is surveyed, there is more chance of observing a substitution in at least one of the branches. This approach has proved successful on primate sequences (Bofelli et al., 2003), since a large number of genomes have been sequenced from relatively closely related species.

The realization of the full potential of phylogenetic shadowing has been limited by the under-representation of whole genome sequences from non-mammalian vertebrate lineages. For example, *Xenopus tropicalis* and *Xenopus laevis*, two model organisms routinely used in developmental studies, are the only representative amphibians that

currently have fully sequenced genomes (Bowes et al., 2009).

Our understanding of key evolutionarily processes will be enhanced as genome sequences are obtained from across a range of divergence times (to distinguish signals of passive constraint resulting from short divergence time), but also as additional whole genome sequences are obtained from within clades. For example, *Lepisosteus oculatus* (spotted gar), the genome of which will soon be available, diverged before the teleost-specific whole genome duplication event and is an important outgroup species to the teleost lineage because it allows patterns of sequence evolution pre and post whole genome duplication to be compared (Amores et al., 2011). Further sequencing of key vertebrate species whose genomes represent a range of mutation rates, generation times and effective population sizes are warranted to inform evolutionary studies that can be exploited for medical, commercial and agricultural purposes.

1.4. Population genomics

The impact of advances in sequencing technologies has arguably been most salient in its application to population genomics. It has informed our understanding of host-parasite interactions, epidemiological dynamics with potential impacts on global health (Rambaut et al., 2008; Smith et al., 2009), explaining variable drug response and designing personalised therapy (Goldstein et al., 2004; Sadee et al, 2005; Roden et al., 2006, Potti et al., 2006; Garnett et al., 2012). It has already played a substantial role in the agricultural industry improving stock selection (Georges 2001; Rohrer, 2004; Rothschild and Plastow, 2008), phenotypic expression to enhance crop yield (Stuber et al., 1999; Tuberosa and Salvi, 2006) and in conservation genetics (Kohn et al., 2006; Allendorf et al., 2010). Moreover, it continues to shape our understanding of human

evolution.

The analysis of genetic variation in human populations needs to take into account the history of our species. The Out-Of-Africa (OOA) reconstruction of this history provides the main framework: this widely established theory suggests a single African origin for modern humans, followed by global colonisation involving serial founder events. The first convincing genetic evidence was based on Restriction Fragment Length Polymorphism (RFLP) mapping of mitochondrial DNA (mtDNA) in individuals of African and non-African ancestry (Cann et al., 1987). This analysis was somewhat equivocal involving an unrooted tree, but the inference was subsequently supported by a rooted phylogeny of mtDNA sequences from individuals of African and non-African ancestry with chimpanzee as an outgroup (Vigilant, L. et al., 1991), corroborative evidence from polymorphic nuclear loci showing patterns of relatedness suggesting migration routes and loss of genetic diversity consistent with serial bottlenecks (Cavalli-Sforza, 1994).

As a result of this history, individuals of European descent have less genetic diversity than those of African descent (Jorde et al., 2000). Furthermore, linkage disequilibrium (LD), determined by the combination of alleles along a chromosome that are inherited together (haplotypes), is higher in non-African populations relative to African populations. Over time, recombination events break down LD within a haplotype. Therefore, lower LD would be expected in a large population with a more ancient origin (Reich et al., 2001). Admixture mapping exploits such patterns in LD especially to determine the origin of disease-susceptibility loci in admixed populations where the frequency of alleles may be very different in the parent populations (Chakraborty & Weis, 1988; Stephens et al., 1994; Falush et al., 2003).

Direct-to-consumer services such as 23andme, deCODEme and Navigenics have also exploited this broad approach to evaluate disease-susceptibility and ancestry, although disease risk should be interpreted with caution because outcomes can be different depending on the markers and reference panel used (Imai et al., 2010).

1.5. Signatures of selection from human population data

Selection acts on new deleterious mutations to eliminate alleles carrying them from the population; this type of selection is sometimes referred to as negative selection or purifying selection. It reduces genetic diversity within populations because the disadvantaged allele does not spread, and reduces divergence between species as long as the selection is sustained in both species. Purifying selection will also affect linked loci: this effect is known as as background selection. In effect, haplotypes containing a deleterious allele are unlikely to spread (unless they can be separated from the deleterious allele by recombination). This means the population size is effectively reduced, leading to greater genetic drift and hence loss of genetic diversity. The effect of a single site under purifying selection would be weak, because such deleterious mutations would be unlikely – however a region of DNA with many constrained sites will have tangible effects in reducing the effective population size at linked loci (Charlesworth et al., 1993).

The term ‘positive selection’ is often used for the form that increases the frequency of new and existing (Przeworski et al., 2005) mutations that increase fitness. The fixation of such alleles would be expected to increase divergence between species over and

above that expected under neutrality. At the same time, as frequency of haplotypes bearing the selected locus increases to fixation – a process known as a selective sweep, the genetic diversity will be reduced at linked loci (Maynard-Smith and Haigh, 1974), since neutral mutations in regions closely linked to the favoured allele also reach a high frequency (hitchhiking effect). Consequently, the level of LD (correlation between alleles inherited together) is increased resulting in unusually long haplotypes being maintained as seen at genes implicated in resistance to malaria G6PD (Sabeti et al., 2002) and the lactase gene associated with lactase persistence (Tishkoff et al., 2007). Because LD is broken down over time, high levels of LD are characteristic of recent selective sweeps, and some surveys have attributed such patterns in the human genome to sweeps arising after human-chimp divergence (Mikkelsen et al., 2005; Hernandez et al., 2011). Since an increase in LD can also result from low recombination rates, correlations between recombination rates and levels of diversity are used to distinguish between the effects (Nachman, 2001).

However, demographic effects can confound patterns of variation from different modes of evolutionary forces and can be a cause of departure from neutral-equilibrium, which is routinely detected by the use of statistics like Tajima's D (Tajima, 1989), Fay and Wu's H statistic (Fay and Wu, 2000). For example, an excess of high-frequency variants, expected after a recent population contraction can be misinterpreted as a signal of positive selection whereas an excess of low-frequency variants, expected after a recent population expansion, can be easily misinterpreted as the action of purifying selection. Hence, patterns of variation due to underlying demographic effects need to be extricated when inferring selection (Bamshad and Wooding, 2003; Stajich and Hahn, 2004). Since demography affects the entire genome whereas signatures of selection are more localised, comparing localised patterns of variations with genome-wide patterns from

the same individuals can help distinguish between the two (Nielsen, 2005; Nielsen et al., 2007). A departure from neutral expectation in F_{ST}, a measure of population differentiation widely used to detect selection (Beaumont and Balding, 2004) makes comparisons of genome-wide data from diverse populations.

One major source of information on allele frequencies throughout the human genome was generated by the HapMap project, an international effort to catalogue common variants (minor allele frequency >0.05) and define haplotype blocks to prioritise SNPs in association studies (International HapMap Consortium, 2005). Over 1 million SNPs in 270 individuals from four geographically diverse populations were genotyped with the aim of cataloguing at least one common SNP every 5 kb across the human genome. Ten ENCODE regions (The ENCODE Project Consortium, 2007) deemed functional based on evolutionary and computational analyses were also surveyed.

In a progressive phase of the HapMap project an additional 2.1 million SNPs ascertained from a number of genotyping platforms including Perlegen, Affymetrix and Illumina arrays were genotyped in the same individuals (International HapMap Consortium, 2007). However, since a growing number of studies seemed to suggest that rare variants played a more important role in common disease than previously assumed (Pritchard 2001; Pritchard & Cox, 2002; Li & Leal, 2008, Bodmer & Bonilla, 2008; Negentsev, 2009; Goldstein, 2009), Phase III of the HapMap project, which included ~ 1000 individuals across eleven geographically diverse populations (International HapMap Consortium, 2010), also imputed rare variants (MAF <5%) in addition to cataloguing copy number variants (CNVs). Since then, several studies have shown that rare variants in noncoding regions cannot be ignored in the manifestation of common

disease (Hirschhorn & Daly, 2005; Manolio et al., 2009; Haller et al., 2009).

Falling costs of sequencing and technological advances that reduced the time taken to sequence a human genome made it feasible to identify an even larger number of variants by surveying entire genomes from multiple individuals. Consequently, the 1000 Genomes Project focused their efforts on low coverage whole genome sequencing and high coverage exome sequencing in a greater number of populations including eight populations from the HapMap project (The 1000 Genomes Project Consortium, 2010). While the HapMap and 1000 Genomes Projects did not associate their samples with any phenotypic data based on disease and environmental conditions, the UK10K project aims to uncover even more rare variants by deep sequencing of phenotyped cohorts of ~ 10,000 individuals to further inform association studies. The unprecedented quantity of data generated from these ventures have led to the development of a vast array of new statistical methodologies (fineSTRUCTURE, IMPUTE, LDhat etc.), tools (GATK, VCFtools etc.) and analytical pipelines (Galaxy, CLCBio, IGV etc.) that have considerably advanced the field of genomics within a short span of time.

In this study I take advantage of sequence data from divergent vertebrate species and population genomic data from the HapMap and 1000 Genomes projects to explore the evolutionary dynamics in CNEs. In Chapter 2, I look for phylogenetic evidence of accelerated evolution within the CNE complement and also identify sites that may be potential candidates of adaptive regulatory evolution in the human lineage. In Chapter 3, I profile the patterns of purifying selection in CNEs and determine that the distribution of selective effects in CNEs is more consistent than in coding sequence. Finally in Chapter 4, I interrogate the patterns of variation in allele-frequency in CNEs and coding sequences across human populations to distinguish differences in the global distribution

of allele-frequencies at sites under purifying selection.

1.6. References

1. Abbasi AA (2011) Evolution of vertebrate appendicular structures: Insight from genetic and palaeontological data. Dev Dyn 240: 1005-1016.
2. Amores A, Catchen J, Ferrara A, Fontenot Q, Postlethwait JH (2011) Genome evolution and meiotic maps by massively parallel DNA sequencing: spotted gar, an outgroup for the teleost genome duplication. Genetics 188: 799-808.
3. Amores A, Force A, Yan YL, Joly L, Amemiya C, et al. (1998) Zebrafish hox clusters and vertebrate genome evolution. Science 282: 1711-1714.
4. Aparicio S, Chapman J, Stupka E, Putnam N, Chia JM, et al. (2002) Whole-genome shotgun assembly and analysis of the genome of Fugu rubripes. Science 297: 1301-1310.
5. Aparicio S, Hawker K, Cottage A, Mikawa Y, Zuo L, et al. (1997) Organization of the Fugu rubripes Hox clusters: evidence for continuing evolution of vertebrate Hox complexes. Nat Genet 16: 79-83.
6. Aparicio S, Morrison A, Gould A, Gilthorpe J, Chaudhuri C, et al. (1995) Detecting conserved regulatory elements with the model genome of the Japanese puffer fish, Fugu rubripes. Proceedings of the National Academy of Sciences 92: 1684-1688.
7. Bader, David A. "Computational biology and high-performance computing."Communications of the ACM 47.11 (2004): 34-41.
8. Bagheri-Fam S, Ferraz C, Demaille J, Scherer G, Pfeifer D (2001) Comparative genomics of the SOX9 region in human and Fugu rubripes: conservation of short regulatory sequence elements within large intergenic regions. Genomics 78: 73-82.
9. Bejerano G, Pheasant M, Makunin I, Stephen S, Kent WJ, et al. (2004) Ultraconserved elements in the human genome. Science 304: 1321-1325.
10. Benko S, Fantes JA, Amiel J, Kleinjan DJ, Thomas S, et al. (2009) Highly conserved non-coding elements on either side of SOX9 associated with Pierre Robin sequence. Nat Genet 41: 359-364.
11. Berezikov E, Guryev V, van de Belt J, Wienholds E, Plasterk RH, et al. (2005) Phylogenetic shadowing and computational identification of human microRNA genes. Cell 120: 21-24.
12. Bergman CM, Kreitman M (2001) Analysis of conserved noncoding DNA in Drosophila reveals similar constraints in intergenic and intronic sequences. Genome Res 11: 1335-1345.
13. Blomme T, Vandepoele K, De Bodt S, Simillion C, Maere S, et al. (2006) The gain and loss of genes during 600 million years of vertebrate evolution. Genome Biol 7: R43.
14. Boffelli D (2008) Phylogenetic shadowing: sequence comparisons of multiple primate species. Methods Mol Biol 453: 217-231.
15. Boffelli D, McAuliffe J, Ovcharenko D, Lewis KD, Ovcharenko I, et al. (2003) Phylogenetic shadowing of primate sequences to find functional regions of the human genome. Science 299: 1391-1394.
16. Bowes JB, Snyder KA, Segerdell E, Gibb R, Jarabek C, et al. (2008) Xenbase: a Xenopus biology and genomics resource. Nucleic Acids Res 36: D761-767.
17. Brenner S, Elgar G, Sandford R, Macrae A, Venkatesh B, et al. (1993) Characterization of the pufferfish (Fugu) genome as a compact model vertebrate genome. Nature 366: 265-268.
18. Brunet FG, Roest Crollius H, Paris M, Aury JM, Gibert P, et al. (2006) Gene loss and evolutionary rates following whole-genome duplication in teleost fishes. Mol Biol Evol 23: 1808-1816.
19. Cavalli-Sforza LL, Menozzi P, Piazza A. (1994). The History and Geography of Human Genes
20. Cock PJ, Antao T, Chang JT, Chapman BA, Cox CJ, et al. (2009) Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics 25: 1422-1423.
21. Collins FS, Morgan M, Patrinos A (2003) The Human Genome Project: lessons from large-scale biology. Science 300: 286-290.
22. Collins JR, Stephens RM, Gold B, Long B, Dean M, et al. (2003) An exhaustive DNA micro-satellite map of the human genome using high performance computing. Genomics 82: 10-19.
23. D'haene B, Attanasio C, Beysen D, Dostie J, Lemire E, et al. (2009) Disease-causing 7.4 kb cis-regulatory deletion disrupting conserved non-coding sequences and their interaction with the FOXL2 promotor: implications for mutation screening. PLoS Genet 5: e1000522.
24. Davidson EH, Erwin DH (2006) Gene regulatory networks and the evolution of animal body plans. Science 311: 796-800.
25. Davidson EH, Rast JP, Oliveri P, Ransick A, Calestani C, et al. (2002) A genomic regulatory network for development. Science 295: 1669-1678.

26. De Gobbi M, Viprakasit V, Hughes JR, Fisher C, Buckle VJ, et al. (2006) A regulatory SNP causes a human genetic disease by creating a new transcriptional promoter. Science 312: 1215-1217.
27. Dermitzakis ET, Clark AG (2002) Evolution of transcription factor binding sites in Mammalian gene regulatory regions: conservation and turnover. Mol Biol Evol 19: 1114-1121.
28. Dohm JC, Lottaz C, Borodina T, Himmelbauer H (2008) Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. Nucleic Acids Res 36: e105.
29. Donoghue PC, Purnell MA (2005) Genome duplication, extinction and vertebrate evolution. Trends Ecol Evol 20: 312-319.
30. Drake JA, Bird C, Nemesh J, Thomas DJ, Newton-Cheh C, et al. (2006) Conserved noncoding sequences are selectively constrained and not mutation cold spots. Nat Genet 38: 223-227.
31. Elgar G, Sandford R, Aparicio S, Macrae A, Venkatesh B, et al. (1996) Small is beautiful: comparative genomics with the pufferfish (*Fugu rubripes*). Trends Genet 12: 145-150.
32. Elgar G, Vavouri T (2008) Tuning in to the signals: noncoding sequence conservation in vertebrate genomes. Trends Genet 24: 344-352.
33. Ferris SD, Whitt GS (1979) Evolution of the differential regulation of duplicate genes after polyploidization. J Mol Evol 12: 267-317.
34. Force A, Lynch M, Pickett FB, Amores A, Yan YL, et al. (1999) Preservation of duplicate genes by complementary, degenerative mutations. Genetics 151: 1531-1545.
35. Frankel N, Wang S, Stern DL (2012) Conserved regulatory architecture underlies parallel genetic changes and convergent phenotypic evolution. Proc Natl Acad Sci U S A 109: 20975-20979.
36. Ganley AR, Kobayashi T (2007) Phylogenetic footprinting to find functional DNA elements. Methods Mol Biol 395: 367-380.
37. Garcia-Fernández J, Holland PW (1994) Archetypal organization of the amphioxus Hox gene cluster. Nature 370: 563-566.
38. Garcia-Fernàndez J, Holland PW (1996) Amphioxus Hox genes: insights into evolution and development. Int J Dev Biol Suppl 1: 71S-72S.
39. Goldstein DB, Chikhi L (2002) Human migrations and population structure: what we know and why it matters. Annu Rev Genomics Hum Genet 3: 129-152.
40. Goto N, Prins P, Nakao M, Bonnal R, Aerts J, et al. (2010) BioRuby: bioinformatics software for the Ruby programming language. Bioinformatics 26: 2617-2619.
41. Gumucio DL, Shelton DA, Bailey WJ, Slightom JL, Goodman M (1993) Phylogenetic footprinting reveals unexpected complexity in trans factor binding upstream from the epsilon-globin gene. Proc Natl Acad Sci U S A 90: 6018-6022.
42. Hahn MW, Rockman MV, Soranzo N, Goldstein DB, Wray GA (2004) Population genetic and phylogenetic evidence for positive selection on regulatory mutations at the factor VII locus in humans. Genetics 167: 867-877.
43. Harbison CT, Gordon DB, Lee TI, Rinaldi NJ, Macisaac KD, et al. (2004) Transcriptional regulatory code of a eukaryotic genome. Nature 431: 99-104.
44. Hardison RC (2000) Conserved noncoding sequences are reliable guides to regulatory elements. Trends Genet 16: 369-372.
45. Hellsten U, Khokha MK, Grammer TC, Harland RM, Richardson P, et al. (2007) Accelerated gene evolution and subfunctionalization in the pseudotetraploid frog Xenopus laevis. BMC Biol 5: 31.
46. Holland PW (1997) Vertebrate evolution: something fishy about Hox genes. Curr Biol 7: R570-572.
47. Holland PW, Garcia-Fernàndez J, Williams NA, Sidow A (1994) Gene duplications and the origins of vertebrate development. Dev Suppl: 125-133.
48. Hubbard T, Barker D, Birney E, Cameron G, Chen Y, et al. (2002) The Ensembl genome database project. Nucleic Acids Res 30: 38-41.
49. Hughes, A. L. (1999). Phylogenies of developmentally important proteins do not support the hypothesis of two rounds of genome duplication early in vertebrate history. Journal of molecular evolution, 48(5), 565-576.
50. Hughes MK, Hughes AL (1993) Evolution of duplicate genes in a tetraploid animal, Xenopus laevis. Mol Biol Evol 10: 1360-1369.
51. Hutchison CA (2007) DNA sequencing: bench to bedside and beyond. Nucleic Acids Res 35: 6227-6237.
52. Consortium IHGS (2004) Finishing the euchromatic sequence of the human genome. Nature 431: 931-945.
53. Keightley PD, Kryukov GV, Sunyaev S, Halligan DL, Gaffney DJ (2005) Evolutionary

constraints in conserved nongenic sequences of mammals. Genome Res 15: 1373-1378.
54. Kellis M, Patterson N, Endrizzi M, Birren B, Lander ES (2003) Sequencing and comparison of yeast species to identify genes and regulatory elements. Nature 423: 241-254.
55. Kleinjan DA, Seawright A, Childs AJ, van Heyningen V (2004) Conserved elements in Pax6 intron 7 involved in (auto)regulation and alternative transcription. Dev Biol 265: 462-477.
56. Kleinjan DA, van Heyningen V (2005) Long-range control of gene expression: emerging mechanisms and disruption in disease. Am J Hum Genet 76: 8-32.
57. Kleinjan DJ, van Heyningen V (1998) Position effect in human genetic disease. Hum Mol Genet 7: 1611-1618.
58. Krampis K, Booth T, Chapman B, Tiwari B, Bicak M, et al. (2012) Cloud BioLinux: pre-configured and on-demand bioinformatics computing for the genomics community. BMC Bioinformatics 13: 42.
59. Lang JM, Darling AE, Eisen JA (2013) Phylogeny of bacterial and archaeal genomes using conserved genes: supertrees and supermatrices. PLoS One 8: e62510.
60. Lettice LA, Daniels S, Sweeney E, Venkataraman S, Devenney PS, et al. (2011) Enhancer-adoption as a mechanism of human developmental disease. Hum Mutat 32: 1492-1499.
61. Lettice LA, Heaney SJ, Purdie LA, Li L, de Beer P, et al. (2003) A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. Hum Mol Genet 12: 1725-1735.
62. Málaga-Trillo E, Meyer A (2001) Genome Duplications and Accelerated Evolution of Hox Genes and Cluster Architecture in Teleost Fishes. American Zoologist 41: 676-686.
63. Mardis ER (2011) A decade's perspective on DNA sequencing technology. Nature 470: 198-203.
64. Matsunaga, A.; Tsugawa, M.; Fortes, J., "CloudBLAST: Combining MapReduce and Virtualization on Distributed Resources for Bioinformatics Applications," eScience, 2008. eScience '08. IEEE Fourth International Conference on , vol., no., pp.222,229, 7-12 Dec. 2008
65. Maxam AM, Gilbert W (1977) A new method for sequencing DNA. Proceedings of the National Academy of Sciences 74: 560-564.
66. Maynard-Smith, J. & Haigh, J. The hitch-hiking effect of a favorable gene. Genet. Res. 23, 23–35 (1974).
67. Metzker ML (2010) Sequencing technologies - the next generation. Nat Rev Genet 11: 31-46.
68. Meyer A, Van de Peer Y (2005) From 2R to 3R: evidence for a fish-specific genome duplication (FSGD). Bioessays 27: 937-945.
69. Morley M, Molony CM, Weber TM, Devlin JL, Ewens KG, et al. (2004) Genetic analysis of genome-wide variation in human gene expression. Nature 430: 743-747.
70. Moses AM, Chiang DY, Kellis M, Lander ES, Eisen MB (2003) Position specific variation in the rate of evolution in transcription factor binding sites. BMC Evol Biol 3: 19.
71. Moses AM, Pollard DA, Nix DA, Iyer VN, Li XY, et al. (2006) Large-scale turnover of functional transcription factor binding sites in Drosophila. PLoS Comput Biol 2: e130.
72. Nègre N, Brown CD, Ma L, Bristow CA, Miller SW, et al. (2011) A cis-regulatory map of the Drosophila genome. Nature 471: 527-531.
73. Nichols R (2001) Gene trees and species trees are not the same. Trends Ecol Evol 16: 358-364.
74. Nielsen R (2005) Molecular signatures of natural selection. Annu Rev Genet 39: 197-218.
75. Nielsen R, Hellmann I, Hubisz M, Bustamante C, Clark AG (2007) Recent and ongoing selection in the human genome. Nat Rev Genet 8: 857-868.
76. Ohno S (1999) Gene duplication and the uniqueness of vertebrate genomes circa 1970-1999. Semin Cell Dev Biol 10: 517-522.
77. Oleksiak MF, Churchill GA, Crawford DL (2002) Variation in gene expression within and among natural populations. Nat Genet 32: 261-266.
78. Pakendorf B, Stoneking M (2005) Mitochondrial DNA and human evolution. Annu Rev Genomics Hum Genet 6: 165-183.
79. Pastinen T, Hudson TJ (2004) Cis-acting regulatory variation in the human genome. Science 306: 647-650.
80. Postlethwait J, Amores A, Cresko W, Singer A, Yan YL (2004) Subfunction partitioning, the teleost radiation and the annotation of the human genome. Trends Genet 20: 481-490.
81. Prlić A, Yates A, Bliven SE, Rose PW, Jacobsen J, et al. (2012) BioJava: an open-source framework for bioinformatics in 2012. Bioinformatics 28: 2693-2695.
82. Przeworski M, Coop G, Wall JD (2005) The signature of positive selection on standing genetic variation. Evolution 59: 2312-2323.
83. Rzhetsky A (1995) Estimating substitution rates in ribosomal RNA genes. Genetics 141: 771-783.

84. Sandelin A, Bailey P, Bruce S, Engstrom PG, Klos JM, et al. (2004) Arrays of ultraconserved non-coding regions span the loci of key developmental genes in vertebrate genomes. BMC Genomics 5: 99.
85. Sanger F, Nicklen S, Coulson AR (1977) DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci U S A 74: 5463-5467.
86. Schadt EE, Linderman MD, Sorenson J, Lee L, Nolan GP (2010) Computational solutions to large-scale data management and analysis. Nat Rev Genet 11: 647-657.
87. Schatz MC (2009) CloudBurst: highly sensitive read mapping with MapReduce. Bioinformatics 25: 1363-1369.
88. Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, et al. (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res 15: 1034-1050.
89. Smith JJ, Kuraku S, Holt C, Sauka-Spengler T, Jiang N, et al. (2013) Sequencing of the sea lamprey (*Petromyzon marinus*) genome provides insights into vertebrate evolution. Nat Genet 45: 415-421, 421e411-412.
90. Spitz F, Gonzalez F, Duboule D (2003) A global control region defines a chromosomal regulatory landscape containing the HoxD cluster. Cell 113: 405-417.
91. Stajich JE, Block D, Boulez K, Brenner SE, Chervitz SA, et al. (2002) The Bioperl toolkit: Perl modules for the life sciences. Genome Res 12: 1611-1618.
92. Steinke D, Salzburger W, Braasch I, Meyer A (2006) Many genes in fish have species-specific asymmetric rates of molecular evolution. BMC Genomics 7: 20.
93. Taylor RC (2010) An overview of the Hadoop/MapReduce/HBase framework and its current applications in bioinformatics. BMC Bioinformatics 11 Suppl 12: S1.
94. Teichmann SA, Babu MM (2004) Gene regulatory network growth by duplication. Nat Genet 36: 492-496.
95. Van de Peer Y, Maere S, Meyer A (2009) The evolutionary significance of ancient genome duplications. Nat Rev Genet 10: 725-732.
96. Vigilant L, Stoneking M, Harpending H, Hawkes K, Wilson AC (1991) African populations and the evolution of human mitochondrial DNA. Science 253: 1503-1507.
97. Watson JD, Crick FH (1953) Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. Nature 171: 737-738.
98. Woolfe A, Goode DK, Cooke J, Callaway H, Smith S, et al. (2007) CONDOR: a database resource of developmentally associated conserved non-coding elements. BMC Dev Biol 7: 100.
99. Woolfe A, Goodson M, Goode DK, Snell P, McEwen GK, et al. (2005) Highly conserved non-coding sequences are associated with vertebrate development. PLoS Biol 3: e7.
100. Wray GA (2007) The evolutionary significance of cis-regulatory mutations. Nat Rev Genet 8: 206-216.
101. Yap, Tieng K., Ophir Frieder, and Robert L. Martino. High performance computational methods for biological sequence analysis. Kluwer academic publishers, 1996.
102. Zhang Z, Gerstein M (2003) Of mice and men: phylogenetic footprinting aids the discovery of regulatory elements. J Biol 2: 11.
103. Zöllner S, Pritchard JK (2005) Coalescent-based association mapping and fine mapping of complex trait loci. Genetics 169: 1071-1092.

Chapter 2

2. Prioritising Conserved Noncoding Elements with candidate adaptive changes for functional assays

2.1. Introduction

The sequencing of an increasing number of genomes has been the key driver behind comparative genomic studies. One use of this information has been to infer much more accurately the evolutionary relationships different organisms, resolving previously ambiguous phylogenies; an improvement made possible by comparison of multiple homologous DNA sequences from each species. The underlying logic of phylogenetic reconstruction is that a large number of differences are expected to have accumulated between sequences that are evolutionarily distant whereas fewer differences suggest the sequences under consideration are more closely related. If these changes accumulate as a result of effectively neutral mutations becoming fixed by genetic drift, then the rate of nucleotide substitution would be effectively constant (Kimura, 1979).

However, the rate at which changes accumulate is known to vary among sites within the genome: forces other than random genetic drift such as positive or negative (purifying) selection can cause a departure from the number of changes expected under neutrality

(Kimura, 1979). Of particular interest in this thesis is strong purifying selection, which minimises the number of differences that can accumulate in sequences under strong functional constraint. For example, Conserved Noncoding Elements (CNEs) that form putative *cis*-regulatory modules are often more highly conserved than coding sequences across similar evolutionary distances (Woolfe et al., 2005). Conversely, positive selection on beneficial mutations can enhance the number of changes that can accumulate between sequences.

In addition to differences from location to location within the genome, it is possible for the nature of selection to change with time. Discovering such changes would be of particular interest, as they may have shaped the different evolutionary trajectories of the different branches of the tree of life. With the recently increased resolution of phylogenetic trees, and the ability to infer selection on different parts of the genome, it is now possible to search for such changes in selective regime. This chapter attempts to do so. The phylogenetic information is compared with evidence from the genetic variation within populations, since both lines of evidence can reveal the action of selection; for example, a number of studies incorporate phylogenetic and population genetic evidence to demonstrate that *cis*-regulatory elements are targets for adaptive evolution in the human lineage (Hahn et al., 2004, Rockman et al., 2005).

2.1.1. Models of nucleotide substitution

Across large evolutionary distances the same site could have accumulated more than one change (multiple hits). If with multiple substitutions the nucleotide returns to the state it was in before the first substitution occurred, the distance between sequences is

underestimated resulting in a phenomenon called long branch attraction (reviewed in Bergsten, 2005). Existing nucleotide substitution models use a transition probability matrix to get around this problem. Therefore such models are best suited for comparing DNA sequences across highly divergent species. The transition probability is the probability that base *i* changes to base *j* after time *t*. A simple model of nucleotide substitution such as the Jukes-Cantor (JC69) (Jukes and Cantor, 1969) assumes that the rate at which a base changes into another is the same across all bases and that each base occurs at the same frequency (i.e. 25%) at equilibrium making it suitable for neutrally evolving noncoding regions.

The Kimura two-parameter (K80) model allows for different substitution rates for transitions and transversions, which is a better representation of the dynamics in real data (Kimura, 1980). Both models assume a symmetric substitution matrix, i.e. given the two allelic states *i* and *j*, the rate at which *i* -> *j* is equal to the rate at which *j* -> *i* (Ziheng Yang, 2006). In contrast, the Felsenstein (F84, Felsenstein and Churchill, 1996), Hasegawa-Kishono-Yano (HKY85) (Hasegawa et al., 1985) and general time reversible (GTR) (Tavaré, 1986) models assume an asymmetric substitution matrix while allowing for different base frequencies at equilibrium. Additionally, the models also incorporate the property of time-reversibility, which means that the transition probability matrix will not be affected by the directionality of the substitution, i.e. distances between sequences can be estimated irrespective of whether one sequence is ancestral relative to the other or not. Unlike the F84 and HKY85 models, the GTR model also allows for relative substitution rates of nucleotides to be calculated.

A GTR model of nucleotide substitution can be combined with a model of rate

heterogeneity that allows for pairs of nucleotide substitutions to have different rates of change. In a gamma model of rate heterogeneity the alpha parameter reflects the differences in substitution rate across different sites in the genome. A small value of alpha gives a highly leptokurtic gamma distribution that describes a situation where a majority of sites have low substitution rates while a small proportion have high substitution rates. As alpha increases to infinity all sites are modeled as evolving at the same rate. A number of software programs for phylogeny inference implement maximum-likelihood methods that estimate the GTR and gamma model parameters from the data [PhyML (Guindon and Gascuel, 2003); RAxML (Stamatakis, 2006); MetaPIGA (Helaers and Milinkovitch, 2010)]. This feature is especially useful when estimating the rate of evolution across noncoding DNA sequences, because substitution rates estimated for non-synonymous and synonymous sites in coding sequences based on codon usage are not appropriate in the context of CNEs.

2.1.2. Differences in evolutionary rates in coding sequences

Accelerated evolution in coding sequences can be detected by an increase in the proportion of synonymous substitutions in one lineage over others using an outgroup species as reference. This approach is implemented in the relative rates test of Li et al. (1985), for example in rodent-human comparisons. The relative rates test exploits our knowledge of the genetic code to distinguish mutations which are more and less likely to be subjected to purifying selection: mutations at non-synonymous sites (mutations that change the amino acid encoded by the sequence) are considered more likely to be deleterious; mutations at synonymous sites, being more likely to be neutral, are considered to have a higher probability of drifting to high frequencies and eventually

reaching fixation. The ratio of non-synonymous substitutions per non-synonymous site to synonymous substitutions per synonymous site (dN/dS) between homologous DNA sequences is therefore a metric used to detect signals of adaptive evolution. An increase in the proportion of non-synonymous substitutions (higher dN/dS) could be attributed to adaptive changes that are beneficial to the organism, or a relaxation of purifying selection.

The dN/dS ratio has been used to show that genes involved in early development (at least in vertebrates) are under stronger constraint than genes required at later developmental stages (Roux and Robinson-Rechavi, 2008). In contrast, the most rapidly evolving genes are immune system genes that constantly have to adapt upon exposure to infections. Additionally, genes with tissue-specific expression have been shown to evolve faster than genes expressed ubiquitously (Zhang and Li, 2004). The dN/dS ratio must be interpreted with some caution. For example a secondary effect of selection is seen on linked sites (Charlesworth, 1993): the action of selection can reduce the effective population size at adjacent sites. In theory reduced effective population size does not affect the rate of substitution of neutral alleles (Kimura, 1979), but weakly-selected non-synonymous mutations may behave as though they were neutral – due to the more vigorous genetic drift in smaller populations – and hence there will be an elevated rate of non-synonymous substitution.

Unfortunately, a metric such as the dN/dS ratio cannot be used to dissect the rate of regulatory sequence evolution, because there is no comparable syntax in non-coding DNA (no distinction between synonymous and non-synonymous sites). Therefore, relatively little is known about the rate of evolution across *cis*-regulatory modules that

govern patterns of gene expression in vertebrates.

2.1.3. Determining evolutionary rates in regulatory elements

The most widely used test for noncoding DNA is Tajima's relative rates test, a $\chi 2$-based test adapted from the relative rates test (Tajima, 1993) that is independent of mutation rate variation in the different lineages and makes no assumptions about the underlying substitution models making it ideal for noncoding DNA. This feature was exploited by Bird et al. (2007) to identify Accelerated NonCoding (ANC) sequences in the human genome using pan-mammalian sequence comparisons. However, with an increase in the divergence times between the species being compared, a simple counts based test is no longer appropriate because any back mutations that could have occurred are not accounted for. Therefore, several likelihood-based test statistics that incorporate substitution models have been developed for exploring accelerated evolution in conserved noncoding DNA (Prabhakar et al, 2006; Kim and Pritchard, 2007).

In this chapter I compare the distribution of expected number of substitutions per site in CNEs against a null model of constrained evolution obtained from simulations of conserved noncoding DNA evolving under a single substitution model. I compare *c.* 2400 CNEs across seven vertebrate species. These species were chosen to represent the major vertebrate lineages, and have an array of divergence times in which mutations could have accumulated. I develop a method to distinguish between CNEs that have undergone accelerated evolution and those that have remained more strongly constrained than others throughout the vertebrate lineages. Having identified a subset of CNEs with accelerated rates of evolution, I then use the allele frequencies in humans to evaluate evidence for adaptive evolution within the human lineage.

2.2. Methods

CNE alignments

The CNEs in the CONDOR database were defined by MLAGAN and SLAGAN alignments between orthologous regions of human, mouse, rat/dog and fugu genomes. Because of the very high sequence identity

within the CNEs, the choice of alignment algorithm has negligible effects as they

always align extremely easily and in the same way. Elements in other species with sequence similarity to the CNEs thus defined were identified by BLAST matches as the whole genomes of additional species were made available. FASTA sequences of the CNEs present in human, macaque, mouse, chicken, frog, zebrafish and fugu were extracted from the CONDOR database (Woolfe et al., 2007) and aligned using default parameters in Clustalw (Larkin et al., 2007) to generate the set of 2419 CNEs used in this analysis.

Generating expected range of tree length from simulated sequences

The concatenated alignment of 2419 CNEs was run through the standard RAxML phylogenetic software (Stamatakis, 2007), a program that infers phylogenies from sequence data by maximum likelihood. A GTR-Gamma model of nucleotide substitution was used to obtain the consensus tree. Values of the alpha shape parameter of the Gamma model of rate heterogeneity, nucleotide frequencies and the transition rates from the GTR matrix that generated the consensus tree were fed into INDELible (Fletcher and Yang, 2009) in order to run a simulation of the sequence evolution expected under this model. INDELible uses as input a phylogenetic tree (the consensus tree in this case) and models the evolution of the sequence along the different branches through the process of substitution according to the parameters of the substitution model specified.

The following consensus tree (in Newick format) generated from the concatenated alignment of CNEs was used as the input tree for the simulation:

(human:0.0040,((chicken:0.0217((zfish:0.0892, fugu:0.1193):0.1753, frog:0.0927):0.0099):0.0274,mouse:0.0344):0.0060,macaque:0.0036):0.0;

In total, 10^6 sequences of length 50-500bp in 50bp increments were simulated to span the variation in the length of CNEs.

A modified version of RAxML (Karen Siu-Ting and Dr. Chris Creevey) was used to force the topology and GTR matrix of the consensus tree obtained from the concatenated CNE alignment to optimise the branch length of the trees in the simulated sequences. Maximum likelihood values obtained from the modified version of RAxML are not comparable to those from the original RAxML program. Therefore the modified version of RAxML was used to re-optimise the branch length of the individual CNEs in a partitioned analysis.

GTR-GAMMA model parameters from the consensus tree used in the simulation.

Developed by: Christopher Creevey Teagasc, Grange, Dunsany, Co. Meath, Ireland. and Karen Siu-Ting Bioinformatics and Molecular Evolution Lab, NUI Maynooth, Ireland.

Transition type	a<->c	a<->g	a<->t	c<->g	c<->t	g<->t
Transition rate	0.98	3.14	0.57	2.16	3.11	1

Alpha parameter = 0.736172

Calculating distances

Values for total tree length were parsed from the RAxML output files. Branch length (distance between leaf and internal node of the Newick trees) was calculated using the BioPerl modules Bio::TreeIO, Bio::Tree::TreeFunctionsI, and Bio::Tree::TreeI. Perl Script used can be found in Appendices. Statistical analyses were performed in R (R Development Core Team, 2012).

CNE-gene associations

Clusters of CNEs associated with genes involved in transcriptional regulation and development are available for download from the CONDOR database (http://condor.nimr.mrc.ac.uk/, Woolfe et al., 2007). Extending the regions containing CNEs, defined by conserved synteny in whole genome comparisons of noncoding DNA between human and fugu as well as presence in at least two of mouse, rat or dog genomes, to the nearest annotated genes identified an over-representation of genes associated with transcriptional regulation and development based on Gene Ontology and InterPro domains. A table of CNE-gene associations defined in this manner was extracted from the CONDOR database and used in the analyses.

2.3. Results

2.3.1. Simulating expected variance with a consistent rate of evolution

The distance from a leaf node to the internal node (the branch length) in a phylogenetic tree represents the expected number of nucleotide substitutions per site, given the substitution model. The total tree length is representative of the total number of substitutions per site that have occurred across the different lineages. If all CNEs are evolving in a similar manner, a single substitution model should be able to explain the observed heterogeneity in the number of nucleotide substitutions per site across the CNE complement. Parameters of the GTR-Gamma model of nucleotide substitution were estimated from a concatenated alignment of 2419 CNEs from human, macaque, mouse, chicken, frog, zebrafish and fugu genomes. The parameter estimates were then used to optimise the branch length in one million (10^6) simulated noncoding sequences to determine the variance in the number of substitutions per site we can expect to see under a single substitution model. The resulting distribution of total tree length obtained from the simulated sequences (Figure 1) represents the null model of constrained evolution against which the total tree length in CNEs can be compared to identify elements with accelerated rates of evolution. The length of simulated sequences range from 50bp-500bp to encompass the variation in length of a majority of CNEs in the analysis because substitutions in short sequences exacerbate the branch length relative to longer sequences, rendering comparisons inappropriate.

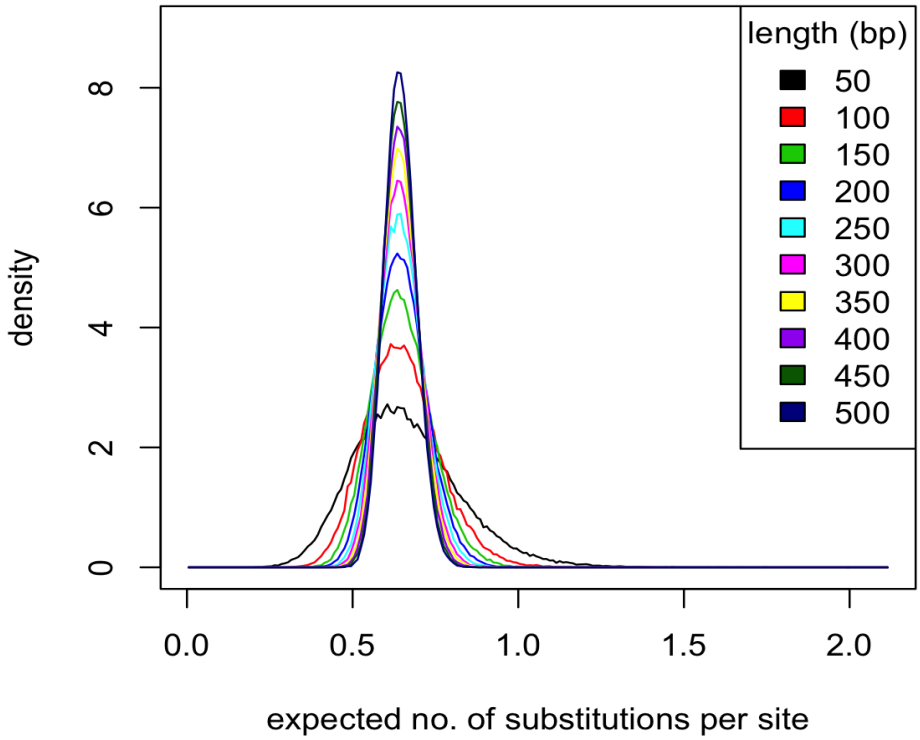

Figure 1: Distribution of the total tree length from simulated sequences. The probability density of the expected number of substitutions per site for each category of simulated sequence based on a single substitution model derived from the consensus tree of a concatenated alignment of 2419 CNEs.

2.3.2.Rate of evolution is not consistent across CNEs

The scatter observed in the number of nucleotide substitutions per site in Figure 2 illustrate a variation in the strength of selection in CNEs. The distribution of the number of nucleotide substitutions per site in the simulated conserved noncoding sequences evolving under a single substitution model (GTR-gamma) derived from the concatenated alignment of CNEs set a boundary to the variance we can expect to see under the assumption that all CNEs are evolving in a consistently constrained manner. The parameter estimates of the same substitution model were then used to optimise the branch length of each CNE in the dataset given the consensus tree. Approximately 30% of CNEs lie outside of the 0.005 and 0.995 quantiles obtained from simulated sequences where 1% would be expected to lie outside if the CNEs were evolving at a consistent rate, according to the fitted model. 18.6% of CNEs lie below the lower quantile (slow evolving) indicating atypically strong selective constraint in these CNEs across all vertebrate lineages; 10.8% lie above the upper quantile (fast evolving) indicating either relaxed constraint within the CNE complement and/or adaptive evolution of these CNEs in some or all lineages.

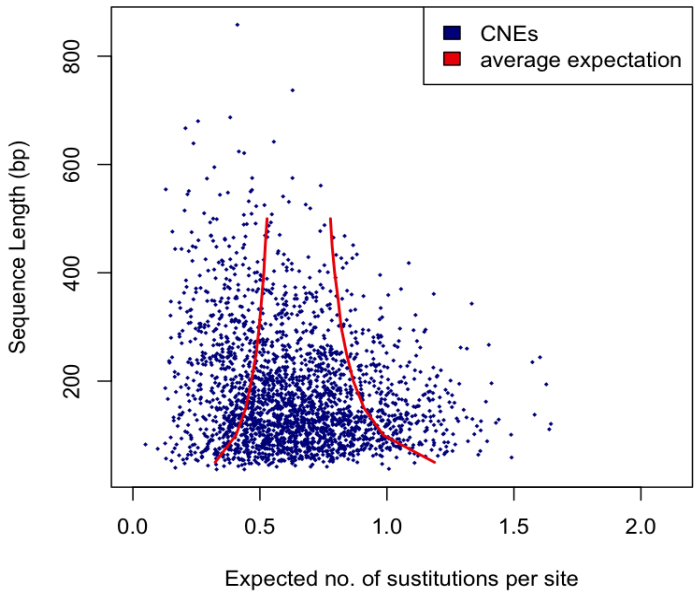

Figure 2: The distribution of total tree length across 2419 CNEs. The expected variance based on sequences evolving at a consistent rate is in red.

According to the observed distribution of substitutions per site across the vertebrate species, CNEs evolving slower than average should be under stronger evolutionary constraint than those evolving faster than average. CNEs evolving at a slower rate are expected to be subject to stronger purifying selection relative to CNEs that are evolving faster. Strong purifying selection has the effect of depressing derived allele-frequencies at selected sites (see Chapter 3), hence one way of determining whether CNEs evolving at a slower rate are indeed more strongly selected than those identified to be evolving faster is to compare the patterns of human polymorphism. A significant excess of rare derived alleles was observed in slow evolving CNEs relative to the faster evolving CNEs when comparing the derived allele frequencies of polymorphic sites in the human CNEs using data from the 1000 Genomes Project (Figure 3). A more detailed analysis of intra-specific variation in human CNEs follows in Chapter 3.

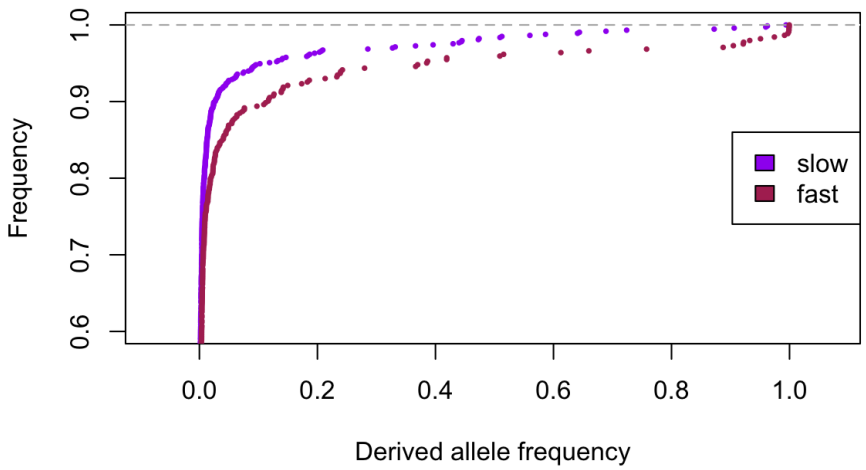

Figure 3: Cumulative frequency distribution of derived alleles in slow and fast evolving CNEs. Slow evolving CNEs have an excess of rare derived alleles relative to faster evolving CNEs (Kolmogorov-Smirnov Test, *p*-value = 0.008).

2.3.3.Identifying putative adaptive changes in CNEs

The total tree length does not distinguish between CNEs with high substitution rate over the whole tree and CNEs from the case where specific lineages (branches) have accumulated more differences than others. A separate comparison was therefore carried out (Figure 4), comparing the distribution of each terminal branch length to identify CNEs that have accumulated exceptionally high or low numbers of substitutions. Of particular interest were the 40 CNEs in the human lineage that have accumulated a higher than average number of differences. Of them, 22 also lie outside the simulation quantiles in at least one other species. The 18 CNEs with high rate in the human lineage (from the human- macaque ancestor to humans) were considered candidates for harbouring human-specific substitutions that might be adaptive. The numbers of human-specific substitutions and polymorphic sites in the CNE outliers are in Table 1. Note that the longest branch length estimates in the human lineage were for CRCNE00011085 and CRCNE00011098 associated with the SHOX gene. This long branch-length was attributed to the duplicated CNEs aligning with paralogous, rather than orthologous, sequences in the other species. Therefore, duplicated CNEs were excluded in further analyses.

Because the alignment did not include more closely related primates I aligned the CNE sequences across human, chimpanzee, orangutan and macaque to identify human-specific substitutions. Six of the CNEs did not have any human-specific substitutions. Nonetheless they are examples of CNEs where changes have become fixed in the human-chimp-orangutan ancestor after the split from macaque. When the 12 remaining CNE outliers were aligned with CNE sequences in all available vertebrate sequences in CONDOR, substitutions identified as human-specific by the 4-way primate alignment could also be found in other non-primate species. Five of the 12 CNEs have such sites that have been subject to multiple hits throughout the course of evolution and often the

base has reverted to the state in ancient vertebrates such as the teleosts. Substitutions that are truly human-specific in the pan-vertebrate alignment are only present in 7 CNE outliers and are potentially the most interesting ones to explore further. The alignments of these seven CNEs are in Appendix 6.4.

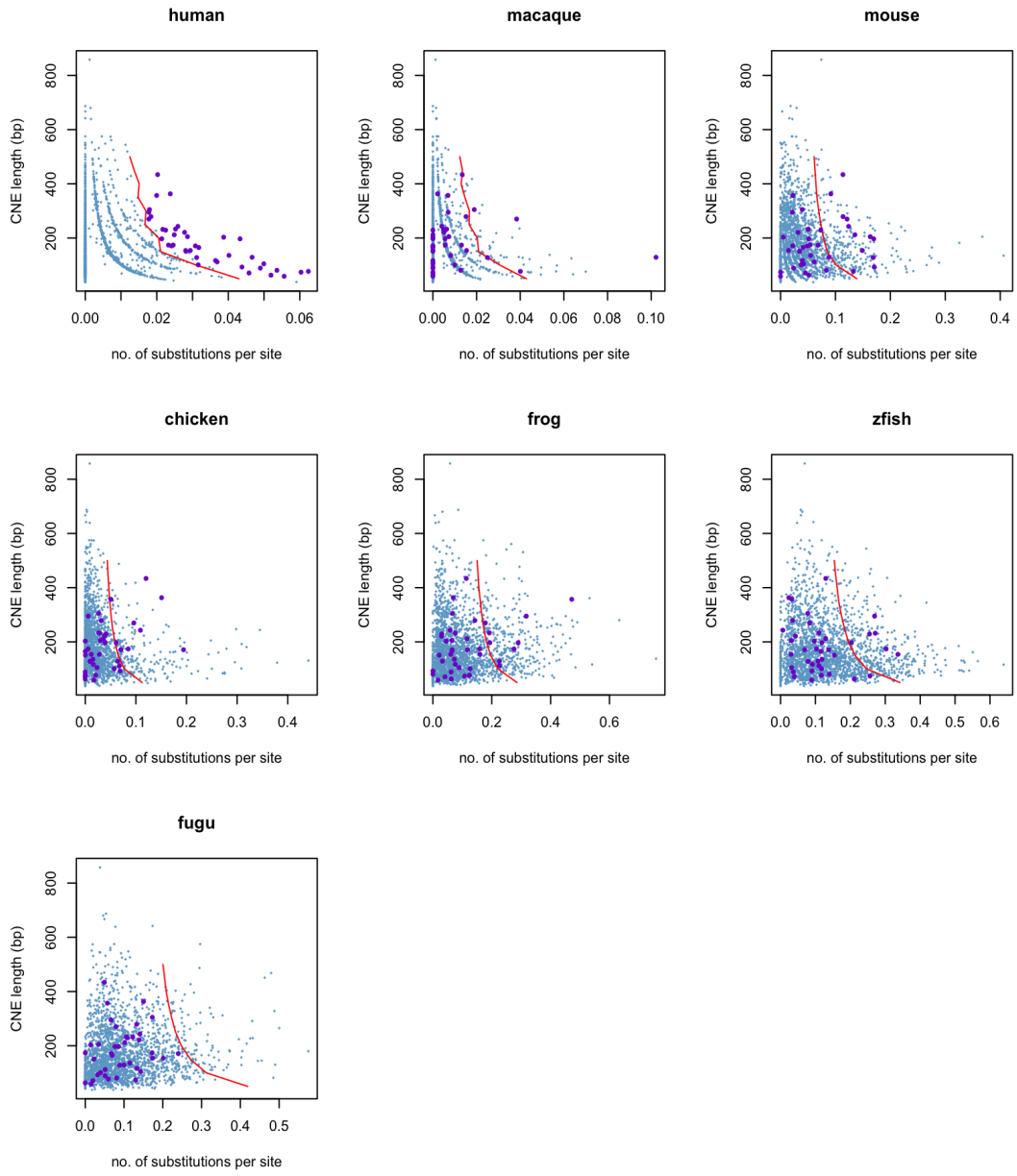

Figure 4: CNEs that have accumulated a large number of substitutions in the human lineage relative to the phylogenetic tree across seven species. The red boundary is obtained from simulating sequence evolution from parameters of the substitution model inferred from the CNEs themselves and facilitates classification of outliers. Data points in purple track fast-evolving human CNEs across the other lineages. It is evident that the pattern is not consistent across the other lineages.

CNE ID	CNE Length (bp)	Branch length	Gene association	No. of human-specific substitutions (4-way primate alignment)	No. of human-specific substitutions (pan-vertebrate)	No. of polymorphic substitutions	Total no. of polymorphic sites in CNE
CRCNE00003541	81	0.05360099	FOXP2	2	2	0	2
CRCNE00001579	165	0.03174884	POU6F2	3	1	1	3
CRCNE00005971	154	0.02938899	SATB1	1	1	1	3
CRCNE00008385	151	0.02814392	NR2F2	1	1	1	1
CRCNE00005870	117	0.0364222	ATBF1	3	1	0	0
CRCNE00008599	71	0.04579545	BNC2	2	1	0	0
CRCNE00003046	221	0.0277654	SALL3	2	1	0	4
CRCNE00004980	197	0.02139997	ZNF503.2	2	0	1	4
CRCNE00008371	203	0.03870938	NR2F2	7	0	0	1
CRCNE00005714	74	0.06033682	ZIC1	1	0	0	0
CRCNE00005408	89	0.04895919	TCF7L2	1	0	0	1
CRCNE00002159	136	0.04018667	GLI3	1	0	0	1
CRCNE00011098	***165***	***0.3194425***	***SHOX***	***24***		***1***	***2***
CRCNE00000270	101	0.03159111	DACH1	0		0	0
CRCNE00001691	59	0.05562112	MEIS1	0		0	0
CRCNE00004482	105	0.04994101	BCL11B	0		0	1
CRCNE00007924	77	0.06	POU3F2	0		0	2
CRCNE00001875	63	0.05193596	SOX1.1	0		0	2
CRCNE00010271	174	0.02459489	PAX1	0		0	4
CRCNE00011085	*120*	*0.3372965*	*SHOX*	*0*		*0*	*4*

Table 1: The most divergent CNEs in the human lineage and their gene associations.

If the substituted sites were evolving neutrally we would expect to see a large proportion of them to be polymorphic in human populations and their derived alleles to be drifting at intermediate frequencies. Positively selected mutations would be expected to have risen rapidly towards fixation whereas negatively selected mutations would have been lost. An example of two closely linked sites evolving under different selective forces is observed in CRCNE00003541 associated with FOXP2. CRCNE00003541 has two human-specific substituted sites neither of which is polymorphic in humans although

two polymorphic sites exist in close proximity (Figure 5). The derived allele **G** is near fixation in humans at the **A>G** polymorphic site (rs186789231, DAC = 2181/2184) upstream of the first substituted site with two copies and a singleton of the ancestral allele **A** segregating in the GBR and CHB populations respectively. The ancestral allele is retained in all species except for orangutan (derived allele = **G**) and bat (derived allele = **C**). In contrast, the **C>T** polymorphism (rs139000268) downstream of the first substituted site is rare in humans (DAC=29/2184) and is only substituted in bat; its absence in orangutan suggests it is a relatively recent mutation in the human lineage. However, given the low coverage of the bat genome, both instances of derived mutations may be limited to the orangutan and human lineages.

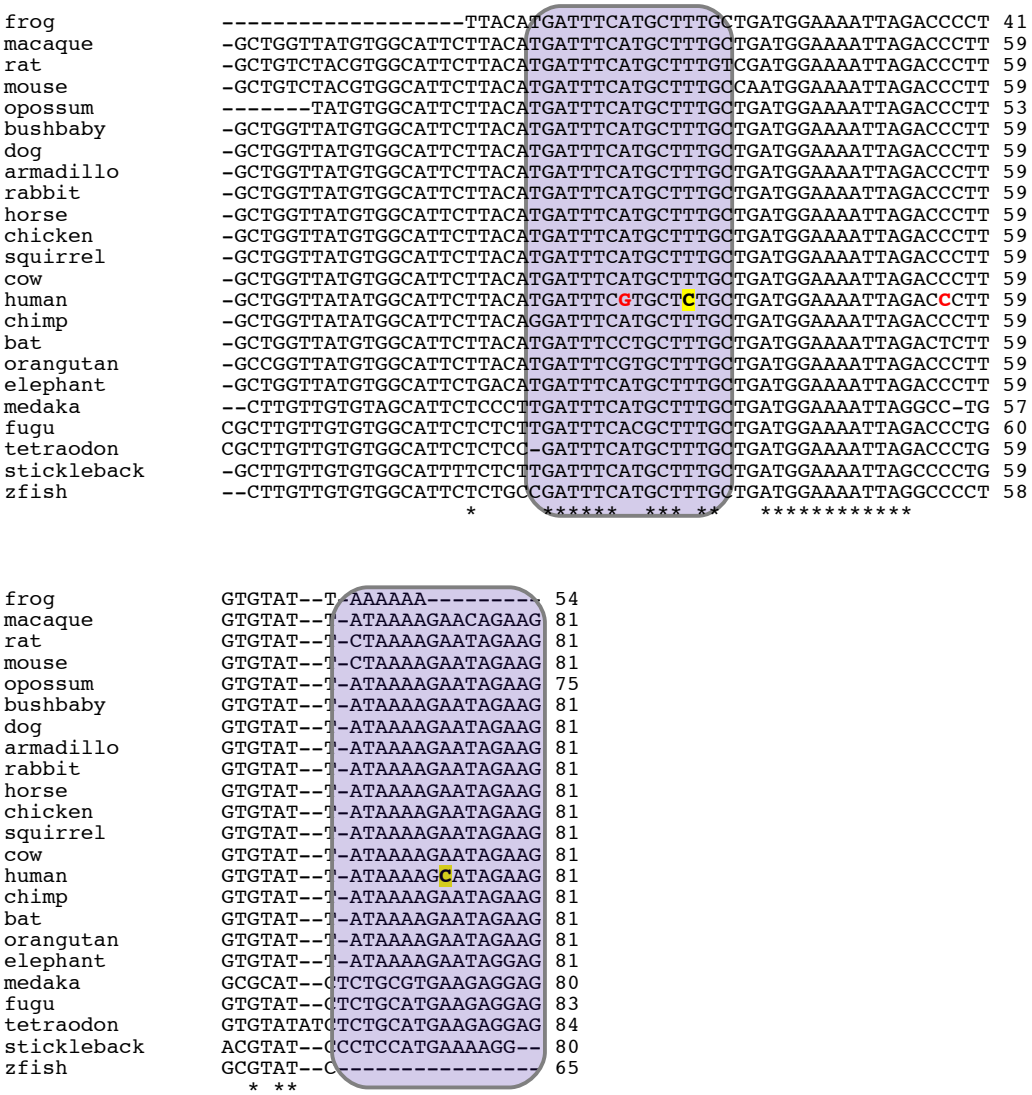

Figure 5: CRCNE00003541 (FOXP2 Region) alignment across vertebrates. Polymorphic sites rs186789231 and rs139000268 flanking the human-specific substitution (highlighted) are in red.

Figure 6: CRCNE00005971 (SATB1 Region) alignment across vertebrates. (T>C) polymorphisms rs182302130 and rs145059587. rs145059587 (highlighted) is close to fixation in the human lineage (DAC = 2176/2178).

In a second example, CRCNE00005971 associated with SATB1 has just one human-specific substitution that is also polymorphic out of a total of two polymorphic sites in the element (Figure 6). The **T>C** polymorphism rs145059587 is close to fixation in the human lineage (DAC = 2176/2178) with only two copies of the ancestral allele **T**, both singletons in GBR and ASW populations, segregating in a phylogenetically invariant position, which suggests an adaptive role. In contrast, although four copies of the ancestral allele **T** in closely linked rs182302130 (**T>C**) are segregating in the ASW (DAC=3) and YRI (DAC=1) populations; it is in a phylogenetically variant position and is more likely under relaxed constraint.

2.4. Discussion

2.4.1.Outlier detection

The excess of rare derived alleles in the slow evolving CNEs relative to fast evolving CNEs supports the interpretation differences that the shorter tree lengths are observed at more strongly constrained loci. The excess of expected number of substitutions per site in fast evolving CNEs could be attributed to either relaxed constraint on these elements or adaptive changes within one or more lineages; of these, the CNEs with substituted sites that are not polymorphic are good candidates to look for adaptive changes.

2.4.2.Putative adaptive changes

If the elevated substitution rate were due to adaptive changes that have been selected for and are being maintained by purifying selection, few polymorphisms would be expected at those substituted sites. For example, the human-specific substitution in CRCNE00003541 associated with FOXP2 is closely flanked by two polymorphic sites but is not polymorphic itself. The FOXP2 gene that encodes a transcription factor has been a target of recent selection in humans (Enard, 2002). Therefore it is likely that associated regulatory elements are also subject to similar novel selective forces. Given the implication of FOXP2 mutations in speech and language disorders (Zhang, J. et al. 2002), experimental investigation of mutations in CRCNE00003541 to gauge any effects on gene expression patterns is warranted.

The CNEs associated with HOXD9, IRX5, NR2F1, FOXP1 and ZNF503 have the highest frequency of elevated substitution rate (Figure 7). However some rapidly evolving elements may have escaped detection owing to the ascertainment criterion that

required the elements to be aligned across all seven species in the study; this requirement means that many known CNEs have been excluded from the analysis: this dataset represents only ~ 40% of the full CNE complement associated with each gene, (Table 2). There is no direct evidence of such a bias – there is no correlation between the proportions of fast-evolving CNEs in a cluster and the percentage of CNEs that could not be aligned. In fact, the low number of CNEs that could be aligned across all seven species is most likely attributed to the low coverage of the zebrafish and frog genomes available. As the number and quality of sequenced genomes improves, it will be worthwhile constructing alignments and repeating the analyses with the latest sequences available.

Given the roles and evolutionary history of the HOX and IRX genes for instance, the observations may be of heuristic value in understanding the rapid rate of evolution in associated elements. For example, the HOX gene clusters, which encode highly conserved transcription factors that guide spatio-temporal gene expression during early development, have been extensively studied as a model of morphological innovation in vertebrates due to their diversification in the vertebrate lineages following tandem gene duplications (Duboule and Dolle 1989; Holland et al., 1992; Amores et al., 1998; Duboule 2007). Furthermore, evidence of adaptive evolution (Lynch et al., 2006; Liang Lu et al., 2013) and recently acquired regulatory control (Spitz at al., 2001) in HOX genes suggest that CNEs associated with these genes may be potential sources of adaptation in some or all lineages. In support of this scenario, a recent study elegantly demonstrated that a single lineage-specific substitution in a conserved *cis*-regulatory element played a major role in the evolution of the vertebrates body plan (Guerreiro et al., 2013). The authors demonstrated that the substitution (specific to snakes and other

animals with extended ribcages such as elephants and manatees) in an otherwise phylogenetically invariant position of a HOX/PAX enhancer, promotes the growth of extra ribs in transgenic mice (a mammalian system where the phenomenon of an extended ribcage is not normally observed).

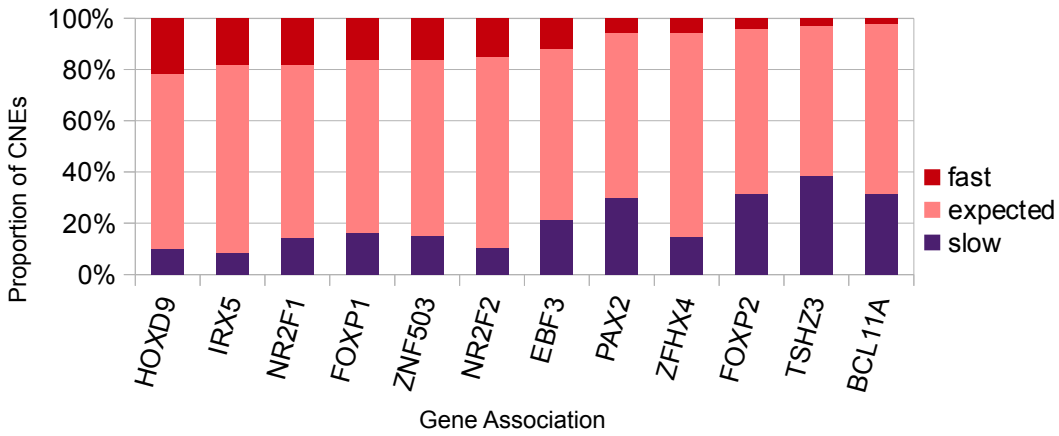

Figure 7: CNE evolution by gene association. The relative proportions of fast and slow evolving CNEs across each gene cluster using a subset of the data where at least 50 CNEs are associated with a gene.

Gene cluster	Total no. of CNEs in gene cluster	No. of CNEs in dataset	% of total
HOXD9	143	51	36
IRX5	344	82	24
NR2F1	242	55	23
FOXP1	111	55	50
ZNF503	293	87	30
NR2F2	165	87	53
EBF3	209	75	36
PAX2	113	70	62
ZFHX4	128	54	42
FOXP2	152	73	48
TSHZ3	238	101	42
BCL11A	160	51	32

Table 2: The CNEs surveyed in each gene cluster.

2.4.3. Fate of duplicated CNEs - SHOX as an example

CRCNE00011085 and CRCNE00011098 located ~ 4kb downstream of the SHOX gene in the pseudo-autosomal region (PAR1) of human Chr X are two extreme outliers: they had the fastest rate of evolution detected in the human lineage. However, this extreme result depends on an alignment obtained using the 'best-hits' settings from blastn; this method aligned the sequences to paralogous CNEs associated with SHOX2 in the other species. The SHOX gene is present in the chicken genome in a syntenic position to the human SHOX gene, suggesting that the duplication of SHOX2 is ancient. However, the SHOX gene has been lost in the rodents genomes (Zhong and Holland, 2011), but was also absent in the chimpanzee genome currently available on Ensembl, most likely due to low coverage in the PAR regions of the sex chromosomes in chimpanzee.

Despite the long evolutionary period since duplication in which substantial functional divergence could have occurred, the paralogous genes express in similar domains; both SHOX2 and SHOX genes are expressed in the limb bud tissue during embryonic development although SHOX2 is mainly expressed in the proximal regions whereas SHOX is expressed at a later stage in the middle part of the limb bud (Clement-Jones et al., 2000; Tiecke et al., 2006). However, the regulatory sequences governing the expression of these genes have since acquired considerable differences. For instance, CRCNE00011098 (165 bp) located downstream of the SHOX gene (on human Chr X) consists of a 57bp core fragment that is also found downstream of the SHOX2 gene (on human Chr 3) (Figure 8). Deletion of the core CNE fragment on Chr 3, which bears 100% identity to the corresponding 57bp region within the full length CNE (165bp) in chimp and macaque, results in a loss of SHOX2 expression (Sabherwal et al., 2007). On the other hand, the same 57bp core in CRCNE00011098 downstream of SHOX has

eight substitutions compared to the ancestral element downstream of SHOX2. While these substitutions do not disrupt a deeply conserved PBX-HOX site common to both elements they do however disrupt a MEIS site upstream of the PBX-HOX sites in the SHOX element (Figure 9), which may contribute to aforementioned SHOX-specific spatio-temporal expression patterns. Changes to expression patterns may well be observed in such scenarios because binding sites in proximity to each other often form heterodimers/trimers (motif clusters) that recruit transcription factors; PBX transcription factors cooperatively bind DNA with MEIS and HOX proteins (Shanmugam, 1999) and disrupting such interactions may have a loss-of-function effect (Parker et al., 2011). Therefore it was not surprising that a variant analysis across ~50 CNEs with co-occurring PBX-HOX and MEIS binding sites revealed just five polymorphic sites in the binding motifs (Table 3). All five variants segregate at low allele frequencies; they are in the 3' end of the binding motifs at otherwise phylogenetically invariant positions (Figures 10 & 11), suggesting that mutations in the 5' part of the motifs are more likely to have strongly deleterious consequences while 3' changes maybe tolerated.

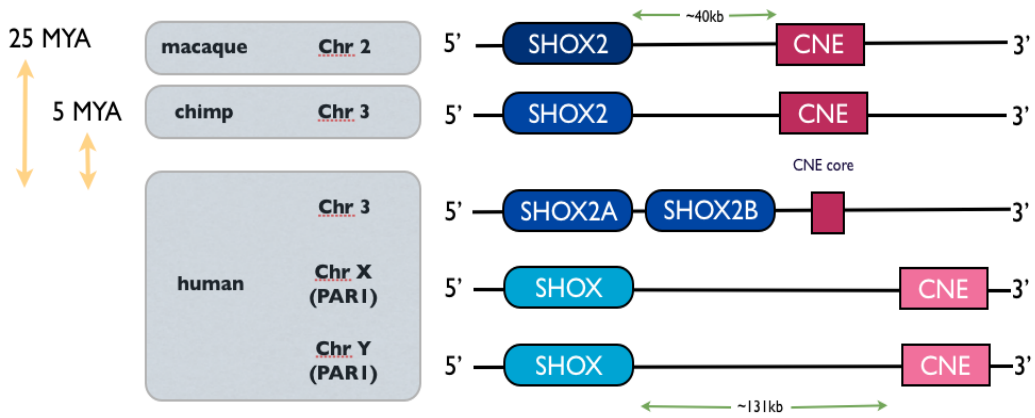

Figure 8: The arrangement of CRCNE00011098 relative to SHOX and SHOX2 genes. The SHOX2 gene, a homolog of SHOX, has undergone a human-specific duplication event after human-chimp divergence and exists in two isoforms.

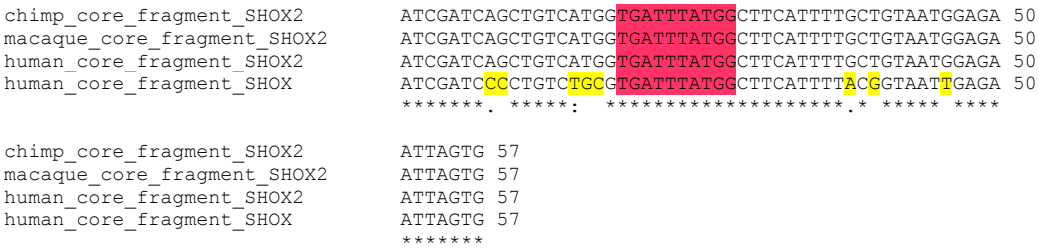

Figure 9: Partial alignment of CRCNE00011098. A PBX-HOX site (red) in the core fragment of the CNE is not disrupted by CRCNE00011098-specific substitutions (highlighted). However, a MEIS site (grey) has one substitution in the SHOX core fragment.

CNE ID	Site	Motif	Derived Allele	Derived Allele Count
CRCNE00003213	MEIS	CTGT**C**A	T	3/2108
CRCNE00004548	MEIS	CTGTC**A**	G	2/2080
CRCNE00009711	MEIS	CTGT**C**A	A	6/708
CRCNE00010260	MEIS	CTGTC**A**	G	2/2148
CRCNE00000750	PBX-HOX	TGATGGATG**G**	T	261/1946
CRCNE00005966	PBX-HOX	TGATAAATC**G**	T	4/2134

Table 3: Frequency of variants from the 1000 Genomes Low coverage data in MEIS and PBX-HOX sites. Of 50 CNEs surveyed, 5 variants from the 1000 Genomes low coverage data mapped to PBX-HOX and MEIS binding motifs. Only one variant has reached a derived allele frequency of 13% while the frequency of the derived allele in the rest is $< 1\%$.

Figure 10: Variant in MEIS site CRCNE00009711 proximal to HMX2 and ZNFN1A5 genes.

Figure 11: Variant in PBX-HOX sites of CRCNE00000750 (reverse strand) associated with the SOX14 gene.

2.5. Future work

Future work can focus on the following aspects:

1. Using a single substitution model to compare the patterns of evolution of the CNE complement in each cluster with the patterns of evolution of the gene they are associated with. Such comparisons may distinguish between the following scenarios:

   a) Both the CNE cluster and its associated gene have undergone accelerated evolution in concert in some or all lineages.

   b) The associated gene has undergone accelerated evolution while the CNEs have not.

   c) Some or all CNEs associated with the gene have undergone accelerated evolution, suggesting regulatory subfunctionalisation may have occurred.

2. Exploring the relative contribution from each of the lineages to the fast evolving CNEs. Because the focus of this study has been to detect outlier CNEs based on the total tree length (i.e. the expected number of substitutions accumulated across the entire tree) and outliers in the human lineage, potential adaptive changes in the other lineages could be explored.

3. Co-injecting human DNA carrying the ancestral and derived states of human-specific substitutions in FOXP2 associated CRCNE00003541 in zebrafish embryos to detect any differences in expression patterns. Allele-specific co-injections can be carried out using the *Tol2* functional assay that is currently being used in the Elgar Lab. *Tol2* or Transposable element of *Oryzias latipes*, number 2 from Medaka is a 4.7kb autonomous element encoding a transposase gene with cut and paste transposition activity (Koga, 2001). This feature has been exploited to develop the

Tol2 transposon-mediated transgenesis of putative enhancers to assay their functionality via Green Fluorescence Protein (GFP) reporter gene expression (Fisher, 2006).

4. It must be noted that only positive regulatory activity can be detected using the *Tol2* vector system. Random integration of the GFP reporter constructs in different regions of the genome can give rise to inconsistent expression patterns (position effects). For example, ectopic gene expression in the sonic hedgehog (SHH) gene involved in human limb formation is caused by a change in the genomic context of SHH that brings it under the influence of a set of enhancers different from its own (Lettice 2011). Therefore, it would be necessary to screen a large number of embryos before a consistent tissue-specific expression pattern can be identified. Nevertheless, differences in expression patterns between the ancestral and derived alleles may still be too subtle to be detected. For example, in sequence-specific DNA-protein interactions, mutations in non-recognition sites within the motif can be tolerated resulting in reduced affinity to the protein manifesting in reduced transcript abundance. This is in contrast to mutations in the recognition site that abolish binding altogether (Weiher 1983, Clark 1988), which are easier to detect due to complete lack of transcription. To overcome this, simultaneous injection of Red Fluorescence Protein (RFP) with GFP has been used to distinguish between different enhancers active in different types of tissues in the same zebrafish embryo (Wan 2002). A similar strategy could be employed to detect differences in expression patterns driven by the two allelic states of the CNE - one tagged with GFP and the other with RFP.

5. Exploring gain/loss of known transcription factor binding sites in fast evolving CNEs. For example, potential candidates could be transcription factor binding sites with high proportions of adaptive substitutions in humans (listed in Arbiza et al., 2013).

6. Combining information on the presence of p300 transcriptional coactivators and monomethylation signatures characteristic of active enhancers identified by the ENCODE project (Ref), to locate enhancer elements in CNEs.

2.6. Acknowledgements

I thank members of the Exelixis lab (Heidelberg Institute for Theoretical Studies, Germany) consisting of Alexandros Stamatakis, Pavlos Pavlidis, Andre Aberer, Fernando Izqueirdo, Nikos Alachiotis, Kassian, and Solon for their help with high performance computing methods during my internship. I would also like to thank Karen Siu-Ting and Dr. Chris Creevey from the Bioinformatics and Molecular Evolution Lab, NUI Maynooth, Ireland for her prompt help with the modified version of RAxML that she helped develop.

2.7. References

1. Amores A, Force A, Yan YL, Joly L, Amemiya C, et al. (1998) Zebrafish hox clusters and vertebrate genome evolution. Science 282: 1711-1714.
2. Arbiza L, Gronau I, Aksoy BA, Hubisz MJ, Gulko B, et al. (2013) Genome-wide inference of natural selection on human transcription factor binding sites. Nat Genet 45: 723-729.
3. Bird CP, Stranger BE, Liu M, Thomas DJ, Ingle CE, et al. (2007) Fast-evolving noncoding sequences in the human genome. Genome Biol 8: R118.
4. Charlesworth B, Morgan MT, Charlesworth D (1993) The effect of deleterious mutations on neutral molecular variation. Genetics 134: 1289-1303.
5. Clark L, Pollock RM, Hay RT (1988) Identification and purification of EBP1: a HeLa cell protein that binds to a region overlapping the 'core' of the SV40 enhancer. Genes Dev 2: 991-1002.
6. Clement-Jones M, Schiller S, Rao E, Blaschke RJ, Zuniga A, et al. (2000) The short stature homeobox gene SHOX is involved in skeletal abnormalities in Turner syndrome. Hum Mol Genet 9: 695-702.
7. Duboule D (2007) The rise and fall of Hox gene clusters. Development 134: 2549-2560.
8. Duboule D, DollÈ P (1989) The structural and functional organization of the murine HOX gene family resembles that of Drosophila homeotic genes. EMBO J 8: 1497-1505.
9. Enard W, Przeworski M, Fisher SE, Lai CS, Wiebe V, et al. (2002) Molecular evolution of FOXP2, a gene involved in speech and language. Nature 418: 869-872.
10. Felsenstein J (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach. J Mol Evol 17: 368-376.
11. Felsenstein J, Churchill GA (1996) A Hidden Markov Model approach to variation among sites in rate of evolution. Mol Biol Evol 13: 93-104.
12. Fisher S, Grice EA, Vinton RM, Bessling SL, Urasaki A, et al. (2006) Evaluating the biological relevance of putative enhancers using Tol2 transposon-mediated transgenesis in zebrafish. Nat Protoc 1: 1297-1305.
13. Fletcher W, Yang Z (2009) INDELible: a flexible simulator of biological sequence evolution. Mol Biol Evol 26: 1879-1888.
14. Guerreiro I, Nunes A, Woltering JM, Casaca A, NÛvoa A, et al. (2013) Role of a polymorphism in a Hox/Pax-responsive enhancer in the evolution of the vertebrate spine. Proc Natl Acad Sci U S A 110: 10682-10686.
15. Guindon S, Gascuel O (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol 52: 696-704.
16. Hahn MW, Rockman MV, Soranzo N, Goldstein DB, Wray GA (2004) Population genetic and phylogenetic evidence for positive selection on regulatory mutations at the factor VII locus in humans. Genetics 167: 867-877.
17. Hasegawa M, Kishino H, Yano T (1985) Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J Mol Evol 22: 160-174.
18. Helaers R, Milinkovitch MC (2010) MetaPIGA v2.0: maximum likelihood large phylogeny estimation using the metapopulation genetic algorithm and other stochastic heuristics. BMC Bioinformatics 11: 379.
19. Holland PW, Holland LZ, Williams NA, Holland ND (1992) An amphioxus homeobox gene: sequence conservation, spatial expression during development and insights into vertebrate evolution. Development 116: 653-661.
20. Kim SY, Pritchard JK (2007) Adaptive evolution of conserved noncoding elements in mammals. PLoS Genet 3: 1572-1586.
21. Kimura M (1979) The neutral theory of molecular evolution. Sci Am 241: 98-100, 102, 108 passim.
22. Kimura M (1980) A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J Mol Evol 16: 111-120.
23. Koga A, Hori H (2001) The Tol2 transposable element of the medaka fish: an active DNA-based

element naturally occurring in a vertebrate genome. Genes Genet Syst 76: 1-8.
24. Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, et al. (2007) Clustal W and Clustal X version 2.0. Bioinformatics 23: 2947-2948.
25. Lettice LA, Daniels S, Sweeney E, Venkataraman S, Devenney PS, et al. (2011) Enhancer-adoption as a mechanism of human developmental disease. Hum Mutat 32: 1492-1499.
26. Li WH, Wu CI, Luo CC (1985) A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. Mol Biol Evol 2: 150-174.
27. Liang L, Shen YY, Pan XW, Zhou TC, Yang C, et al. (2013) Adaptive evolution of the Hox gene family for development in bats and dolphins. PLoS One 8: e65944.
28. Lynch M, Conery JS (2000) The evolutionary fate and consequences of duplicate genes. Science 290: 1151-1155.
29. Lynch VJ, Roth JJ, Wagner GP (2006) Adaptive evolution of Hox-gene homeodomains after cluster duplications. BMC Evol Biol 6: 86.
30. Parker HJ, Piccinelli P, Sauka-Spengler T, Bronner M, Elgar G (2011) Ancient Pbx-Hox signatures define hundreds of vertebrate developmental enhancers. BMC Genomics 12: 637.
31. Prabhakar S, Noonan JP, P‰‰bo S, Rubin EM (2006) Accelerated evolution of conserved noncoding sequences in humans. Science 314: 786.
32. Rockman MV, Hahn MW, Soranzo N, Zimprich F, Goldstein DB, et al. (2005) Ancient and recent positive selection transformed opioid cis-regulation in humans. PLoS Biol 3: e387.
33. Roux J, Robinson-Rechavi M (2008) Developmental constraints on vertebrate genome evolution. PLoS Genet 4: e1000311.
34. Sabherwal N, Bangs F, Rˆth R, Weiss B, Jantz K, et al. (2007) Long-range conserved non-coding SHOX sequences regulate expression in developing chicken limb and are associated with short stature phenotypes in human patients. Hum Mol Genet 16: 210-222.
35. Shanmugam K, Green NC, Rambaldi I, Saragovi HU, Featherstone MS (1999) PBX and MEIS as non-DNA-binding partners in trimeric complexes with HOX proteins. Mol Cell Biol 19: 7577-7588.
36. Spitz F, Gonzalez F, Peichel C, Vogt TF, Duboule D, et al. (2001) Large scale transgenic and cluster deletion analysis of the HoxD complex separate an ancestral regulatory module from evolutionary innovations. Genes Dev 15: 2209-2214.
37. Stamatakis A (2006) RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 22: 2688-2690.
38. Tajima F (1993) Simple methods for testing the molecular evolutionary clock hypothesis. Genetics 135: 599-607.
39. Tiecke E, Bangs F, Blaschke R, Farrell ER, Rappold G, et al. (2006) Expression of the short stature homeobox gene Shox is restricted by proximal and distal signals in chick limb buds and affects the length of skeletal elements. Dev Biol 298: 585-596.
40. Tschopp P, Duboule D (2011) A genetic approach to the transcriptional regulation of Hox gene clusters. Annu Rev Genet 45: 145-166.
41. Wan H, He J, Ju B, Yan T, Lam TJ, et al. (2002) Generation of two-color transgenic zebrafish using the green and red fluorescent protein reporter genes gfp and rfp. Mar Biotechnol (NY) 4: 146-154.
42. Weiher H, Kˆnig M, Gruss P (1983) Multiple point mutations affecting the simian virus 40 enhancer. Science 219: 626-631.
43. Woolfe A, Goode DK, Cooke J, Callaway H, Smith S, et al. (2007) CONDOR: a database resource of developmentally associated conserved non-coding elements. BMC Dev Biol 7: 100.
44. Woolfe A, Goodson M, Goode DK, Snell P, McEwen GK, et al. (2005) Highly conserved non-coding sequences are associated with vertebrate development. PLoS Biol 3: e7.
45. Zhang J, Webb DM, Podlaha O (2002) Accelerated protein evolution and origins of human-specific features: Foxp2 as an example. Genetics 162: 1825-1835.
46. Zhang L, Li WH (2004) Mammalian housekeeping genes evolve more slowly than tissue-specific genes. Mol Biol Evol 21: 236-239.
47. Zhong YF, Holland PW (2011) The dynamics of vertebrate homeobox gene evolution: gain and loss of genes in mouse and human lineages. BMC Evol Biol 11: 169.

Chapter 3

3. Purifying selection in Conserved Noncoding Elements (CNEs) is more consistent than in coding sequences

In the previous chapter, I distinguished between CNEs that show different rates of divergence through vertebrate evolution and, in particular, identified sites that appear to have become fixed in the human lineage through positive selection. In this chapter I make use of data on human polymorphism in combination with the analysis of phylogenetic conservation to profile the effects of purifying selection on CNEs.

3.1. Introduction

The effect of mutations in coding sequence is a well-characterised phenomenon whereby non-synonymous changes alter the resulting protein product. Therefore, mutations causing non-synonymous changes tend to be under stronger purifying selection than synonymous mutations. There are only rare instances where synonymous changes have been demonstrated to have phenotypic effects (Todorova et al., 2003). This difference between synonymous and non-synonymous sites can be exploited in the analysis of genomes; for example, since synonymous changes are assumed to be largely neutral in effect, their genetic diversity can be used as a null distribution against which

to test for evidence to identify loci at which non-synonymous changes have been established by positive selection (e.g. Macdonald-Kreitman test, (Cai et al., 2009)). Conversely, non-synonymous changes are treated as being more likely to be the cause of an altered phenotype. Indeed, genome-wide scans for causative mutations often concentrate on exons (the 'exome') and have been successful in detecting non-synonymous changes responsible for numerous genetic diseases, (Singleton, 2011). By contrast, there is scant data on the effect of mutations in regulatory sequences; in part because regulatory regions can be difficult to identify, although the ENCODE project (The ENCODE Project Consortium, 2011) has successfully identified a large number of regulatory elements from patterns of DNAse hypersensitive sites and post-translational histone modifications in the human genome.

Nevertheless, a few studies have been able to identify *cis*-regulatory mutations implicated in human diseases. For example, beta thalassemia and haemophilia B are both instances of human diseases caused by mutations affecting transcription-factor binding sites in regulatory sequences (reviewed in Epstein, 2009). However, not all transcription-factor binding motifs are well characterised, and other types of sequence may also have regulatory activity. In these cases evolutionary conservation provides a means to identify sequences of functional importance to the organism. If a non-coding sequence has been conserved across large evolutionary periods, so that it is found in a number of diverse species, that pattern suggests the sequence is undergoing selection against mutations that would modify it. Occasionally such mutations have been identified within a species and are indeed found to be deleterious. For example, Lettice et al. found heterozygous dominant point mutations in a highly conserved enhancer 1Mb away from the Shh locus in humans that segregate with pre-axial polydactyly (Lettice et

al., 2003). Similarly, Benko et al., (2009) found a heterozygous point mutation in a highly conserved noncoding element flanking the SOX9 locus that alters binding of the transcription factor MSX1 associated with the Pierre Robin syndrome. More recently, two rare variants in the 5' UTR and an intron of the *RBM8A* gene were found to segregate in individuals with Thrombocytopenia (reduced platelet count) with Absent Radii (TAR) syndrome, whereas no exonic mutations were found in affected individuals at that locus (Albers et al., 2013). Generally it has only been when exome sequencing fails to identify any causative mutations that non-coding DNA is surveyed. Now that whole genomes are being sequenced in greater numbers, such as in the 1000 Genomes Project (Abecasis et al., 2010), there is the opportunity to identify more non-coding variants. However, there are still hurdles to assessing their importance. Firstly, large re-sequencing projects have tended to neglect non-exonic regions resulting in much lower coverage. Consequently, the stringent quality control measures used in SNP calling from next-generation sequencing data mean that rare variants in noncoding DNA may be filtered out as putative sequencing errors.

Secondly, it is much harder to predict the consequence of mutation/variation in regulatory sequences because their grammar is poorly understood. An effective way of identifying putative regulatory sequences, in particular those that are under strong selection, has been to use cross species comparisons, often across large evolutionary distances - a method referred to as phylogenetic footprinting (Tagle et al., 1988). The various methods used to predict *cis*-regulatory modules such as CONREAL (Berezikov et al., 2004), VISTA (Dubchak and Ryaboy, 2006), Modulefinder (Philippakis et al., 2005), often depend on a signal of evolutionary conservation. In contrast, phylogenetic

shadowing of sequences using species that are more closely related helps identify regions that have diverged recently (Bofelli et al., 2003), and may have acquired a lineage-specific role.

In this thesis, the focus is on conserved noncoding elements (CNEs) identified through multiple alignments of mammalian and Fugu genomes (Woolfe et al., 2007). CNEs differ from other sets of conserved non-coding sequence that have been identified by comparative analyses, for example Ultra-Conserved Elements identified by human-mouse-rat genomic comparisons (Bejerano et al., 2004) and Highly Conserved Elements (Siepel et al., 2005), in not overlapping known exons. CNEs tested experimentally in zebrafish embryos show tissue-specific enhancer activity at various stages of embryonic development. These sequences are conserved across all jawed vertebrates and likely define a set of developmental regulatory elements. For example, SOX21 and PAX6-associated CNEs enhance GFP-expression in the developing eye (Woolfe et al., 2005) while FOXP1/FOXP2-associated CNEs up-regulate GFP-expression in the hindbrain (McEwen et al., 2006).

The enrichment of vertebrate CNEs for conserved binding site motifs such as the Pbx-Hox hetero-dimer (Parker et al., 2011) and the over-representation of several transcription factor position weight matrices in mammalian conserved noncoding sequences (Minovitsky et al., 2007) suggest that conservation of noncoding sequences is likely due, at least in part, to the presence of common transcription factor binding sites. Some transcription factors are highly sequence-specific and only bind to genomic regions with the exact transcription factor binding sequence (Stormo et al., 2010). In such highly specific interactions, any variation in the transcription factor binding sequence might have an effect on the transcription factor binding and subsequent gene

expression. For example,

a bias in ChIP-seq reads mapping preferentially to one of two alleles at a heterozygous locus indicates allele-specific binding of CTCF, a transcriptional and chromatin regulator, resulting in varying levels of gene expression at nearby genes (McDaniell et al., 2010). Conversely, those positions that are less strongly conserved may be less important for sequence function. This logic is used in the construction of the Position Weight Matrix (PWM), which reflects the affinity of transcription factors to their preferred binding sites (Spivakov et al., 2012).

On average CNEs are about 200bp in length (maximum being *c.* 800bp), yet their conservation cannot be explained by our current knowledge of transcription-factor binding sites, since most are only 4-10bp long. In fact the rate of evolution of known binding sites is faster than that of CNEs. One possibility would be if the binding sites overlap each other and the order of overlap is necessary to retain the proper function of the *cis*-regulatory module (as discussed in Elgar and Vavouri, 2008). Another hypothesis is that conserved noncoding sequences (CNSs) represent mutational 'coldspots', however this explanation has been rejected (Drake et al., 2006) because of the excess of rare derived alleles observed within CNSs relative to polymorphisms outside CNSs that cannot be explained by population bottleneck effects or background selection. CNEs can be defined by a large number of completely (evolutionarily) conserved sites (Non Variable Regions), as well as a number of more variable sites (Restricted Variable Regions) based on their conservation across seven divergent vertebrate species. In this study, I evaluate the hypothesis that restricted variable regions in CNEs have been accumulating substitutions in the human lineage due to relaxed evolutionary constraint, resulting in more within-species polymorphism than non-variable regions. Using the

occurrence and allele frequencies of SNPs from both the HapMap and 1000 Genomes Projects in CNEs I show evidence that a) non-variable regions within CNEs are under stronger selective constraint than restricted variable regions, b) the distribution of selective effects in CNEs are different to that in non-synonymous sites and c) there are discrepancies between the results obtained from HapMap and 1000 Genomes Project datasets.

3.2. Methods

Generating multiple sequence alignments

CNE sequences in FASTA format were downloaded from the CONDOR database (Woolfe et al., 2007) for the following species: *Homo sapiens, Macaca mulatta*, *Mus musculus*, *Gallus gallus*, *Xenopus tropicalis, Danio rerio* and *Takifugu rubripes.* Out of ~7000 human-fugu CNEs spanning a combined length of ~ 800,000 bp, I identified a subset of 1809 CNEs spanning 318286 bp that could be aligned across all seven species. ClustalW (Larkin et al., 2007) with default parameters was used to align the FASTA sequences that were originally identified in a phylogenetically sensitive manner using MLAGAN (Woolfe et al., 2005). Because of the very high sequence identity within the CNEs, the choice of alignment algorithm becomes irrelevant as they always align extremely easily and in the same way. FASTA sequences for the seven vertebrate species can be downloaded from the CONDOR website using the list of CNE IDs provided in supplementary text S1.

Classifying NVRs and RVRs in CNEs

I defined two classes of sites within CNEs based on their conservation across the seven vertebrate species; custom Perl scripts (see Appendices) were used to distinguish between bases that are similar in all species in the alignments (NVRs) and those that are different in at least one species (RVRs). In instances where the presence of the alternate allele at a polymorphic site in the human reference sequence made a non-variable region a variable region, the site was reclassified as a Non-Variable Region that is polymorphic. 161361 bp across 1809 CNEs were non-variable in all seven species while 156925 bp were variable in at least one species.

Control regions

a) Coding sequences

The Biomart tool (Kasprzyk, 2011) on Ensembl 71 (Flicek et al., 2013) was used to retrieve the exon coordinates of all transcripts on forward strand genes on human chromosomes 1-22 and X. The transcript with the longest coding sequence was retained. The number of 0-fold, 2-fold and 4-fold degenerate sites in the full transcript with the longest coding sequence was obtained by using the software MEGA 5.1 (Tamura et al., 2011). The relative proportion of non-synonymous sites (0-fold degenerate) in the genic sequences was thus determined to be 64% whereas the proportion of synonymous sites (2 and 4-fold degenerate) was determined to be 36%. In each transcript, exons with 5' and 3' UTRs were not included when extracting variants so that only variants in coding sequences were obtained. The consequence of the variant to the transcript was determined from Ensembl using the Transcript ID as a query. Variants with more than one consequence within the same transcript were excluded from the analysis. 59277 non-synonymous SNPs and 47687 synonymous SNPs were thus retained for analysis.

b) Noncoding regions

Five chromosomal regions (Table 1) that did not overlap known exons and CNEs were randomly chosen from five different chromosomes to constitute the noncoding control. Five chromosomes were chosen solely for the ease of computation with which contiguous length-matched regions (for ~ 320,000 bp) could be queried and manipulated. Any bases in the noncoding control that overlapped annotated regulatory features and/or GERP elements were excluded from the analysis. 56540 non-conserved noncoding SNPs were thus identified.

Chromosome	Start	End
Chr4	32431865	33236944
Chr5	30127241	30771304
Chr8	137999178	138819687
Chr9	29516650	30326809
Chr12	87269143	88091174

Table 1: Coordinates of noncoding control regions (Hg19 GRCh37 Assembly).

Extracting allele frequencies from the HapMap Project

I used the marker IDs of SNPs reported from Biomart to extract the allele frequencies of variants from the HapMap Release #27 dataset with a custom XML query. 721 SNP in ~ 800,000 bp of CNE regions were reported from HapMap Release #27. 182 non-synonymous, 400 synonymous and 982 nonconserved non-coding variants were obtained from length-matched control regions. The number of variants that mapped to 318286 bp is 70 and 176 in CNE NVRs and CNE RVRs respectively. The allele frequencies were averaged across all the populations to obtain a global allele frequency for a given variant .

Extracting variants from the 1000 Genomes Project

Variants were extracted from the file "ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20110521/supporting/ALL.wgs.project_consensus_vqsr2b.20101123.snps.low_coverage.sites.vcf.gz" using tabix (Li, 2011). The resulting vcf file was parsed using custom Perl scripts (see Appendices) to obtain derived allele frequencies of variants. After filtering out any variants that reported an alternate allele count of zero, a total of 1227 and 1960 SNPs were retained for analysis in CNE NVRs and CNE RVRs respectively.

Determining the ancestral state of variants

The ancestral state of all variants in the analysis was obtained from Ensembl73 (Flicek et al., 2013). All genomic coordinates were based on Hg19 Feb 2009 assembly of the human genome, Genome Reference Consortium GRCh37.

Significance testing

P-values in the Kolmogorov-Smirnov and Chi-squared significance tests were carried out in R (R Development Core Team, 2012).

Site-frequency spectra analyses

The derived allele counts was extracted from the VCF file using custom Perl scripts to generate the site frequency spectrum for each category of site. The population expansion model in PRFREQ software (Boyko et al., 2008) was used to fit the neutral distribution for synonymous sites under the demographic parameters in Table 5. The same demographic parameters were used to fit the gamma distribution of fitness effects on non-synonymous sites and both classes of CNE sites. A mutation rate per site per generation of 1.8×10^{-8} was used.

3.3. Results and Discussion

At the time of conceptualisation of the study, the number of vertebrate genomes for which good genome-wide coverage was available was limited. Hence, I built alignments of CNEs from the human, macaque, mouse, chicken, frog, zebrafish and fugu genomes, which are a good representation of vertebrates in key stages of our evolution. The combined divergence time between these species represents approximately 2,900 million years of evolution and given the mutation rate variation in the different species, especially the high mutation rate in fish that reflects their short generation time, the conservation of these sequences to such a high degree (> 85% between human and fugu in some instances) is remarkable. The divergence time since the last common ancestor with the human lineage for each of the vertebrate lineages in the alignments is in Table 2.

Organism	Divergence time since last common ancestor (Mya)
Macaque	29.2
Mouse	92.4
Chicken	301.7
Frog	371.2
Zebrafish	400.1
Fugu	400.1

Table 2 : Divergence time since last common ancestor. The divergence time since last common ancestor with human obtained from Time Tree (Hedges *et al*., 2006).

3.3.1.Defining NVRs and RVRs

From nearly 7000 CNEs identified from alignments including human, mouse, rat, dog and fugu (Woolfe et al., 2005), ~1800 CNEs could be aligned across all seven chosen species. The original multiple alignments to identify CNEs were carried out in a phylogenetically sensitive glocal alignment algorithm (MLAGAN) but subsequent alignments have been made using CLUSTALW. Using these alignments I distinguished between two categories of sites within CNEs (Figure 1). CNE sites that are invariant across all seven species in the alignments are defined Non-Variable Regions (hereafter referred to as NVRs). Sites where at least one substitution is present in any of the species are defined Restricted Variable Regions (hereafter referred to as RVRs) because often, the differences at such sites are restricted to certain subgroups such as mammals only or primates only. Therefore, depending on the depth of the phylogeny and choice of species in building alignments, it is acknowledged that RVRs may well be reclassified as NVRs and vice versa.

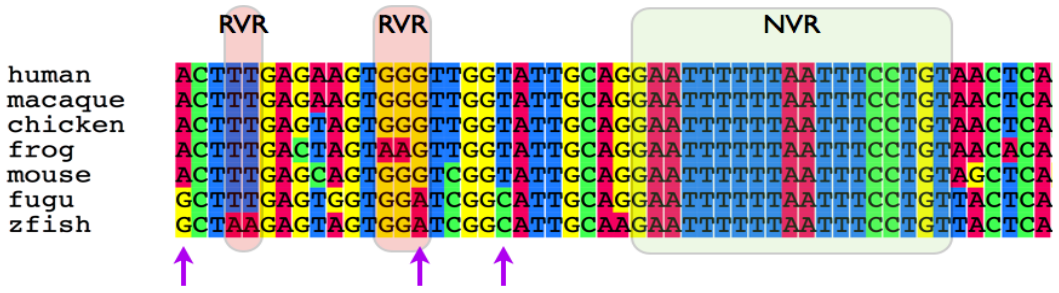

Figure 1 : The two categories of CNE sites. Nonvariable sites (NVRs) are invariant in all species in the alignment. RVRs have at least one substitution in at least one of the species in the alignment. However, some RVRs have substitutions that are restricted to one clade. The purple arrows indicate a few examples of substitutions that are unique to the two teleosts in this alignment.

I compared the proportion of polymorphic sites and the derived allele-frequency spectrum of single nucleotide polymorphisms (SNPs) in CNEs with other regions of the

human genome, for the purpose of exploring selective constraint in CNEs, and in a wider context understanding the evolutionary forces that define *cis*-regulatory modules in vertebrate genomes. Comparisons were made with three categories of region; non-synonymous sites (those resulting in amino acid changes), synonymous sites from coding regions, and non-exonic sequences that do not overlap any CNEs or other annotated regulatory features. The non-synonymous sites act as a positive control, since it is known that a subset of non-synonymous changes is counter-acted by relatively strong purifying selection (Hughes et al., 2003). Conversely, on the assumption that they are under negligible selection, I have used synonymous and non-coding sites as the negative controls. I obtained SNPs that map to CNEs, coding and noncoding regions from the public databases of both the HapMap Project and the 1000 Genomes Project and used the derived allele frequencies of SNPs to compare selective constraint in these sequences. The proportion of polymorphic sites found in 1809 CNEs spanning a length of 318,286 bp, coding and noncoding regions are given in Table 3.

		DAC >= 1		**DAC >= 6**	
Type of site	**No. of sites surveyed**	**No. of SNPs**	**% sites with SNPs**	**No. of SNPs**	**% sites with SNPs**
CNEs (Total)	318286	3187	1	1119	0.35
a) NVR	161361	1227	0.76	367	0.23
b) RVR	156925	1960	1.25	752	0.48
Coding sequences (Total)	12051470	106964	0.89	40946	0.34
c) Non-synonymous (64%)	7712941	59277	0.77	19680	0.26
d) Synonymous (36%)	4338529	47687	1.1	21266	0.49
Noncoding	3611144	56540	1.57	28395	0.79

Table 3: Variants from 1000 Genomes Low Coverage Data across the different categories. The proportion of sites that is polymorphic in each category using SNPs where at least one derived allele is reported (DAC >=1) and those SNPs where at least six derived alleles are reported (DAC >= 6).

3.3.2. Sites within CNEs are subject to different levels of constraint

The nucleotide differences in Restricted Variable Regions of CNEs are sometimes specific to a single evolutionary lineage: a pattern which could be explained if a mutation was fixed in that lineage by positive selection and is being maintained by purifying selection. In that case purifying selection might be of comparable strength in the NVRs and RVRs. In order to investigate whether selective constraint in both classes of CNE sites is comparable I looked at both the proportion of polymorphic sites and spectra of derived allele frequencies. The imputation accuracy for low coverage imputed SNPs in the 1000 Genomes Project was highest for SNPs with an allele count of at least six (Abecasis et al., 2012). Any biases introduced by imputation should affect both classes of CNE sites equally. Therefore, I chose to use this cutoff in the derived allele frequency spectra analyses of CNEs in subsequent comparisons with coding sequences. This cutoff does not affect the patterns observed when derived alleles of allele count less than six are used (see Figure 2B).

When all observed SNPs in the two CNE categories are considered, there are significantly fewer polymorphisms in NVRs relative to RVRs (Table 3). This observation indicates stronger selective constraint at sites that have been conserved across all lineages, most likely reflecting the importance of such sites as functionally indispensable. Mutations in these regions may have functional consequences. A significant difference between the derived allele frequency spectra between the two classes of CNE sites, where NVRs have an excess of rare derived alleles compared to RVRs, is observed (Figure 2). This distinction between the two classes of sites within a CNE indicate that CNEs are composed of sites that are subject to different levels of evolutionary constraint and may have different roles in a regulatory context.

3.3.3. Purifying selection is strongest at NVRs

Previous studies have shown that UltraConserved Elements (UCEs) defined by human-mouse-rat comparisons experience stronger purifying selection than coding regions (Katzman et al., 2007). UCEs are 100% conserved across at least 200bp in the three mammalian genomes used to define them. However, they often overlap exons where non-synonymous mutations can contribute to the excess of rare derived alleles. By contrast, CNEs do not overlap any known exons and represent a larger set of sequences conserved across all jawed vertebrates and with a broader range of sequence identity. Nevertheless, NVRs within CNEs exhibit a lower proportion of sites that are polymorphic relative to non-synonymous sites (Table 3). The reduction in diversity at NVRs in CNEs is accompanied by a derived allele frequency spectrum that shows a significant excess of rare derived alleles relative to non-synonymous sites (Figure 2A, Table 4), indicating stronger purifying selection at NVRs. A comparable pattern of polymorphism in mouse ultraconserved elements was interpreted in a similar manner (Halligan et al., 2011).

Categories of sites (DAC >=6)		**χ^2 *P* value**	**K-S *P* value**
CNE - NVR	CNE - RVR	< 2.20E-016	0
CNE - NVR	Non-synonymous	0.03	3.43E-007
CNE - NVR	Synonymous	< 2.20E-016	4.44E-015
CNE - NVR	Noncoding	< 2.20E-016	< 2.20E-016
CNE - RVR	Non-synonymous	< 2.20E-016	0.08
CNE - RVR	Synonymous	0.54	1.30E-008
CNE - RVR	Noncoding	< 2.20E-016	1.06E-013
Non-synonymous	Synonymous	< 2.20E-016	< 2.20E-016
Non-synonymous	Noncoding	< 2.20E-016	< 2.20E-016
Synonymous	Noncoding	< 2.20E-016	4.44E-016

Table 4: *P* values from Chi-square and K-S tests. *P*-values from the χ^2 test (df=1) to detect differences in proportion of observed polymorphic sites between the different categories and the Kolmogorov-Smirnov test to detect differences in the derived allele-frequency spectra between categories.

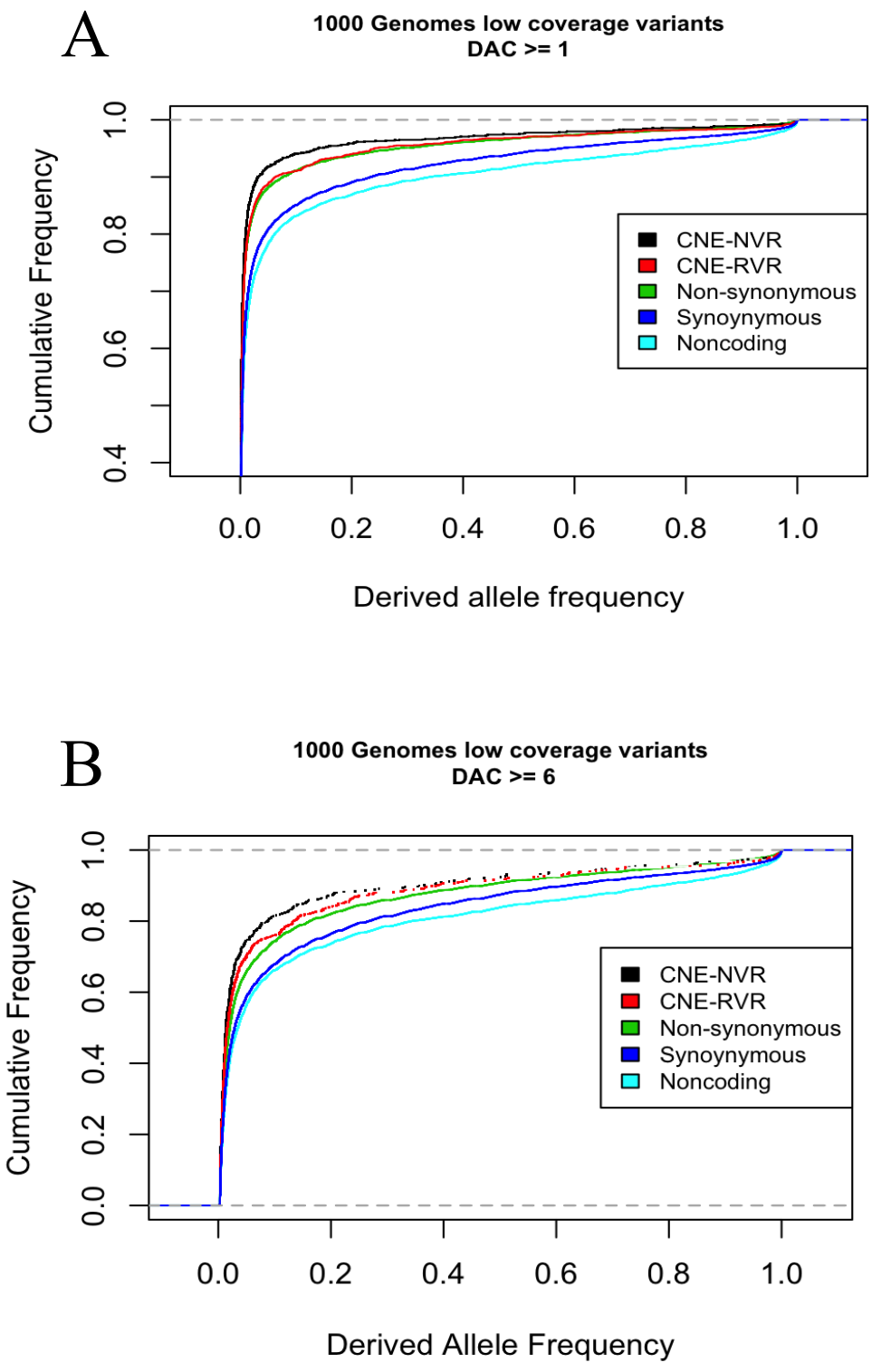

Figure 2: Cumulative derived allele-frequency in CNEs and control regions from 1000 Genomes Project. A) An excess of rare derived alleles is observed in CNE-NVRs, CNE-RVRs and Non-synonymous sites relative to Synonymous and Non-coding controls. B) The pattern remains unchanged when SNPs with a derived allele count of less than six are used.

3.3.4. Reduced diversity in CNEs is not due to a bias in variant calling

A reduced level of variation in CNEs relative to synonymous sites is observed (Table 3), suggesting that mutations in CNEs are subject to continuing purifying selection in the human genome. Assuming that synonymous mutations occur under a neutral model, variation at synonymous sites should be comparable to that at nonconserved noncoding sites. However, the proportion of synonymous sites that are polymorphic is significantly lower than that at nonconserved noncoding sites. These two nearly-neutral categories are expected to differ for a number of reasons, including the effects of hitchhiking and background selection in coding regions (Stephan, 2010) and the effects of epigenetic modification (Keller et al., 2007).

A second consideration is the possible differences in SNP ascertainment bias between coding and non-coding regions. Variants in the 1000 Genomes Project are called using the Variant Quality Score Recalibrator (VQSR) algorithm implemented in the Genome Analysis Toolkit (GATK) (DePristo et al., 2011). VQSR includes HapMap 3 sites as "true sites" to train a Gaussian mixture model, which then evaluates the probability of known and novel variants in the call set being real, as opposed to being an artifact of sequencing or data processing. Because the HapMap sites are predominantly common variants in which coding variants are most likely to have been validated by Sanger sequencing, this could mean that fewer variants in noncoding regions are being reported resulting in the low levels of variation observed in CNEs. If this was indeed the case, then the same bias should extend to nonconserved noncoding sites. However, the proportion of CNE sites that are polymorphic is significantly lower than that at nonconserved noncoding sites demonstrating that the observed low levels of variation at CNE sites is not an artifact of the variant calling procedure.

3.3.5. Purifying selection in CNEs is more consistent than in coding sequences

I employed the approach of interpreting unfolded site-frequency spectra to infer the distribution of selective effects, as developed for comparisons between synonymous and non-synonymous sites in protein coding regions (Piganeau et al., 2003; Eyre-Walker et al., 2006; Boyko et al., 2008). The underlying logic is that mutations under weaker purifying selection have a higher probability of segregating in the population and drifting to higher frequencies than mutations with non-neutral effects. I explored the site-frequency spectra of different classes of sites in the Yoruba population of the 1000 Genomes Project to explain the observed levels of heterozygosity, combined with lower frequency of common alleles found at NVRs, relative to non-synonymous sites.

Initially, I sought to investigate the distribution of selective effects using the full unfolded site-frequency spectra from the YRI population in the 1000 Genomes project using the population expansion model in the PRFREQ software with published parameters (Table 5) inferred from African-American data (Boyko et al., 2008).

Demographic parameters for African Americans under a population expansion model	
Ancestral effective population size (Nanc)	7778
No. of generations since population dynamics (Ngen)	6809
Effective population size of current population (Ncurr)	25636
Scaled time since non-stationary population dynamics (tau = Ngen/2Ncurr)	0.13
Ratio of ancestral to current effective population size (omega = Nanc/Ncurr)	0.3
Selection parameters for gamma distributed fitness effects	
shape	0.18
rate	6.25

Table 5: Parameters inferred from African-American data. Demographic and selection parameters used to fit the distribution of selective effects as published in Boyko et al., 2008.

PRFREQ works in two stages. First, it estimates the demographic parameters based on neutrally evolving loci (the synonymous site frequency spectrum), then the demographic parameters are fixed to estimate the parameters of the gamma distribution of fitness effects in the non-synonymous site frequency spectrum. However, the demographic parameters of tau (scaled time since non-stationary population dynamics) and omega (ratio of ancestral to current effective population size) estimated from African-American data in Boyko et al. (2008) were not good predictors of the synonymous site-frequency spectra in YRI because of European admixture in the African-American data (Figure 3).

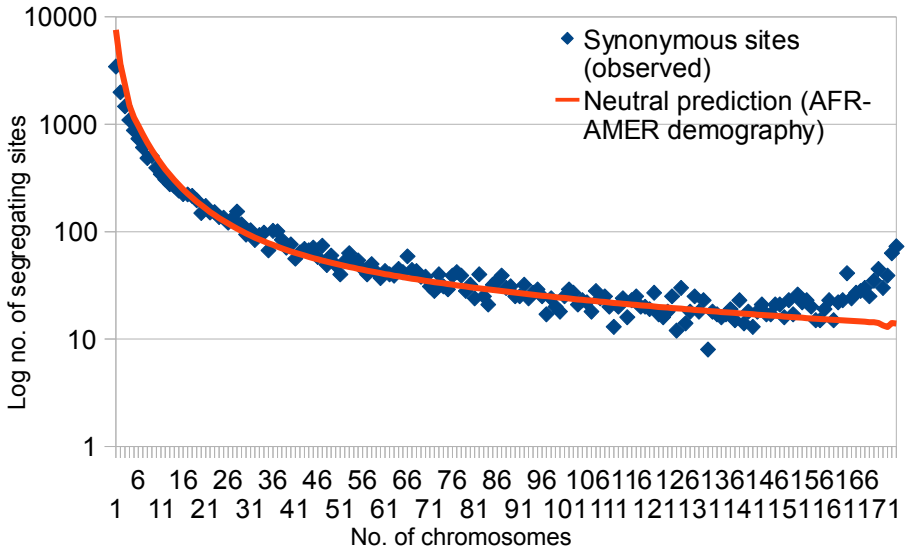

Figure 3: Site Frequency Spectra for synonymous sites in YRI vs their neutral prediction. Using the African-American demographic parameters (Table 4) overestimates heterozygosity in the YRI as seen in the excess of rare alleles simulated, as would be expected after a recent population expansion. The excess of high frequency derived alleles in the spectrum is most likely attributable to ancestral misclassification (Hernandez et al., 2007).

A better fit for the observed synonymous site frequency spectrum when the current effective population size for YRI reduced by half (Table 6), reflecting their relatively small isolated population free of European admixture (Figure 4).

Demographic parameters for YRI under a population expansion model	
Ancestral effective population size (Nanc)	7778
No. of generations since population dynamics (Ngen)	6809
Effective population size of current population (Ncurr)	12818
Scaled time since non-stationary population dynamics (tau = Ngen/2Ncurr)	0.3
Ratio of ancestral to current effective population size (omega = Nanc/Ncurr)	0.61
Selection parameters for gamma distributed fitness effects	
shape	0.1
rate	6.25

Table 6: Adjusted parameters for YRI data. Demographic and selection parameters used to fit the distribution of selective effects in YRI.

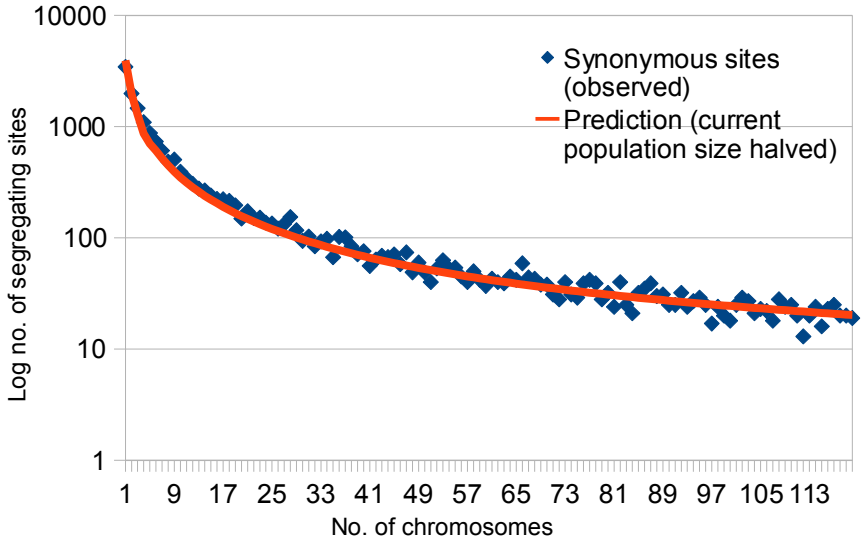

Figure 4: Site Frequency Spectra for synonymous sites in YRI vs their neutral prediction after adjustment. Using the adjusted demographic parameters (Table 2) provided a better fit for the neutral prediction for synonymous sites. The spectrum was truncated at 70% of the total number of chromosomes in the sample to avoid the effect of ancestral misclassification (x-axis now goes up to N=119 not N=175).

The site-frequency spectrum at non-synonymous sites in the data is best explained by a gamma distribution of selective effects with a high variance (σ^2 =2.6x10^{-3}) (Figure 5), consistent with previous findings using human polymorphism data (Boyko et al., 2008).

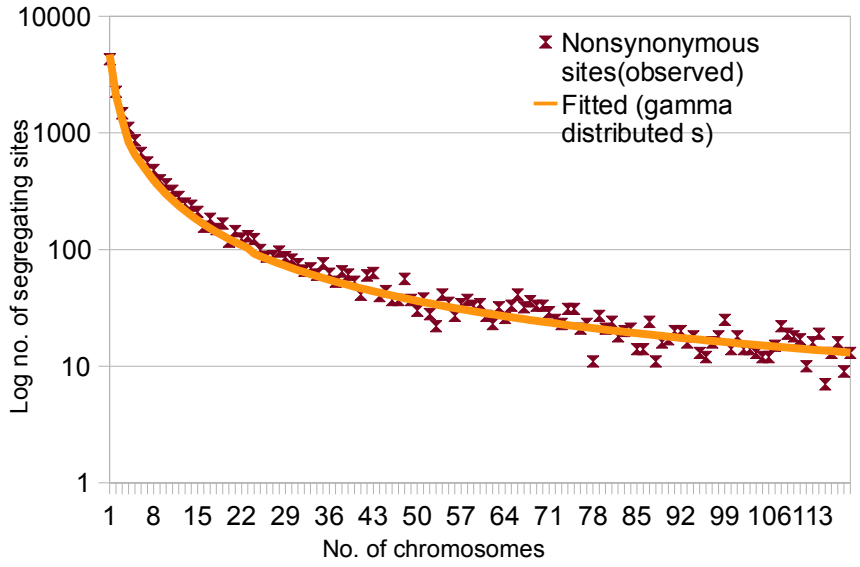

Figure 5: Non-synonymous site-frequency spectrum in YRI and prediction under a gamma distribution of selective effects. The mean of the gamma distribution used to fit the YRI data (shape =0.1, rate = 6.25) was lower than that of the gamma distribution (shape = 0.184, rate =6.25) used to fit the African-American data in Boyko et al. (2008) most likely reflecting the high proportion of high-frequency derived alleles in the YRI data.

The spectrum for NVRs is consistent with a much lower variation in the effects of selection (σ^2 = 3.8 x 10 $^{-9}$), with selection sufficient to keep mutations predominantly at lower frequencies (Figure 6).

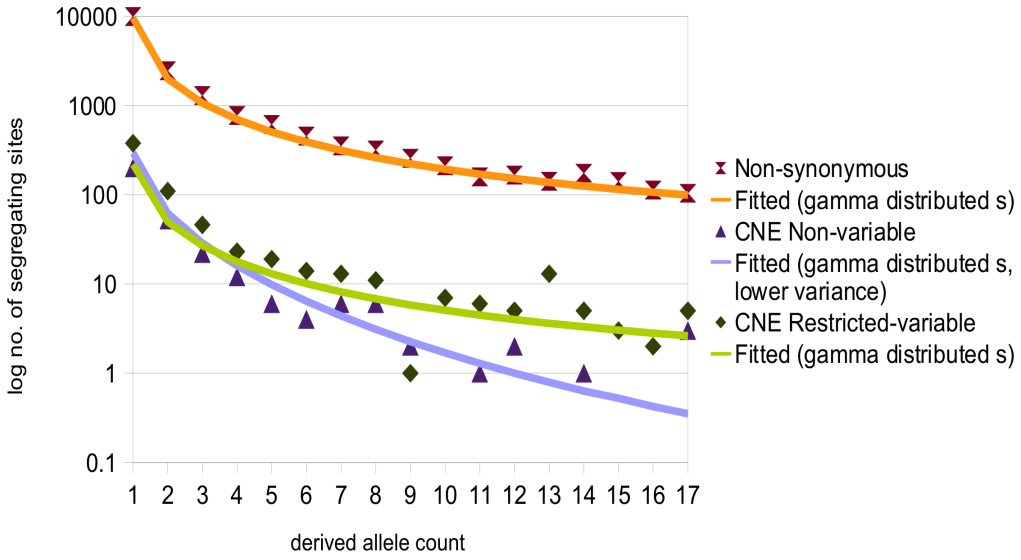

Figure 6: Site-frequency spectra in CNEs and non-synonymous sites. Site-frequency spectra in YRI binned into 5 units where the derived allele is the minor allele. The observed non-synonymous site-frequency spectrum fits a gamma distribution of selective effects (shape = 0.1, rate = 6.25). The observed site-frequency spectrum in CNE NVRs fits a gamma distribution of selective effects with lower variance (shape=12, rate=56250).

Although the derived allele frequency spectrum at RVRs in CNEs is not significantly different from the spectrum at non-synonymous sites, a significantly higher proportion of sites in RVRs are polymorphic relative to non-synonymous sites (Tables 3 and 4). This higher level of heterozygosity in RVRs suggests a relatively smaller proportion of strongly deleterious mutations than at non-synonymous sites. A similar pattern in HapMap and Environmental Genome Project data at a different set of conserved non-coding sites (CNSs), has been attributed to weaker selective effects on CNSs (Asthana et al., 2007).

In a functional context, the different selective regimes acting on non-synonymous sites and CNEs might be attributed to the differences in mechanism of action between protein-coding sequences and noncoding regulatory sequences. Non-synonymous

changes can have major effects on protein structure and function, particularly through truncation. One might expect that there will be a specific number of non-synonymous changes in any coding sequence that might render the resultant protein completely non-functional, whereas other more conservative changes might have little or no effect on protein function. This would be reflected in a wide spectrum of selective pressures at a limited number of non-synonymous sites, with some being essentially immutable (thus no variant alleles) and others having relatively high derived allele frequencies. CNEs represent an entirely different form of functional unit, mediating their action through the binding of large numbers of transcription factors. In general, transcription factor binding sites are highly redundant with a rapid turnover rate (reviewed in Dowell, 2010), but it has been proposed that large *cis*-regulatory modules such as CNEs might be composed of overlapping sets of binding sites thereby imposing a greater evolutionary constraint at each nucleotide position (reviewed in Elgar and Vavouri, 2008).

Nevertheless the activity of regulatory sequences is generally tissue-specific. Mutations in regulatory motifs are consequently more subtle and result in a reduced affinity of a transcription-factor for a motif bearing the mutant allele and altered levels of expression in the genes they regulate. For example, allele-specific differential binding of RNA polymerase II and nuclear factor κB have been associated with SNPs in binding regions (Kasowski et al., 2010). Proteome-Wide Analysis of SNPs (PWAS) have also identified functional differences in transcriptional activity in the presence of SNPs at several transcription factor binding sites implicated in immune response (Butter et al., 2012). In *Drosophila*, reduced levels of polymorphism have been observed at functional transcription factor binding sites (i.e. transcription factor bound motifs) relative to instances of the same motif outside of the bound region that are not deemed functional

(Spivakov et al., 2012). Bound transcription factor-motifs have also been shown to be under stronger purifying selection than unbound motifs (Mu et al., 2011). Consequently, a majority of sites within CNEs are likely to tolerate mutation (excluding those that might lead to a deleterious phenotype), as they are unlikely to have as dramatic an effect on the individual as complete ablation of a protein. However, these sites would still be under very strong purifying selection given their roles in multiple transcription factor binding, hence their extremely low derived allele frequencies.

The site frequency spectrum in CNE RVRs is consistent with a pervasive selective effect that is weaker than at CNE NVRs reflected by the higher level of heterozygosity in RVRs and the higher proportion of derived alleles that drift to high frequencies. The relative relaxation in the selective effect at RVRs implies a difference in functionality; it is possible that a proportion of RVRs function as spacers that maintain the functional binding sites in NVRs. There is evidence that the length of sequence separating functional binding sites is more important than the composition of sequences in some DNA-protein interactions. For example, transcription factor p73 interacts with its half-sites differently in the presence of spacers of various lengths (Ethayathulla et al., 2011). Alternatively, CNEs could consist of overlapping transcription factor binding sites where the more degenerate positions within a binding site are concentrated in RVRs resulting in a higher tolerance to mutations.

3.3.6.Discrepancies between the HapMap and 1000 Genomes datasets

Derived allele frequency spectra from HapMap (International HapMap Consortium, 2005) genotype data have previously been used to compare selective constraint between different types of sequences. For example, Conserved Non-Coding sequences defined by human-mouse and human-dog comparisons, which reflect much smaller divergence times than across the CNEs in this study, have been shown to be under stronger selective constraint than nonconserved regions and under similar constraint to non-synonymous mutations (Drake et al., 2006). Before the 1000 Genomes data was publicly available I also looked at the derived allele-frequency spectra from HapMap Release #27. The derived allele frequency spectra of both categories of CNE SNPs obtained from the HapMap Project (Figure 7) is not significantly different to the spectrum at synonymous sites.

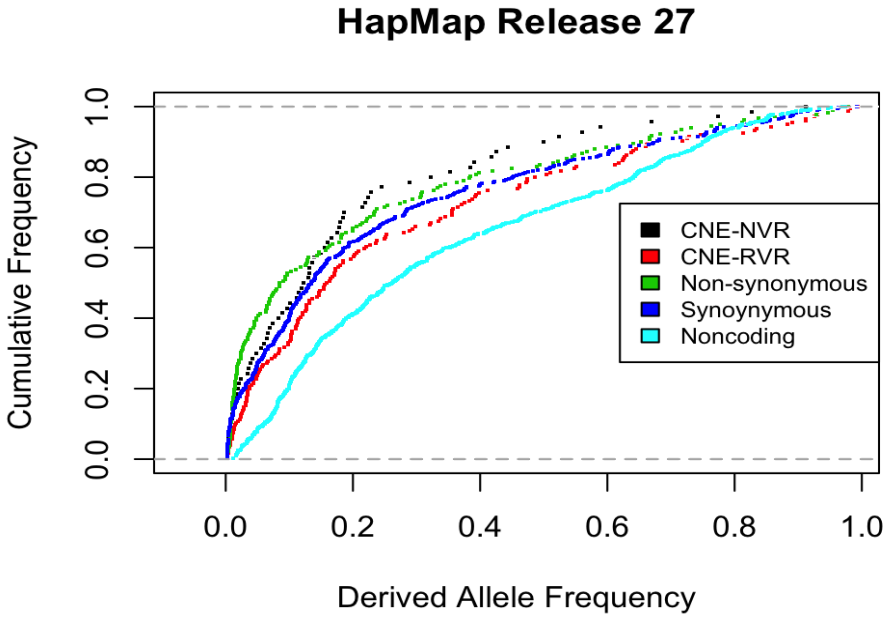

Figure 7: Cumulative derived allele-frequency in CNEs and control regions from the HapMap Project. An excess of rare derived alleles is observed in Non-synonymous sites relative to Synonymous and Non-coding controls. CNE-NVRs have an excess of rare derived alleles compared to CNE-RVRs. The derived allele-frequency spectra in CNEs resemble that at synonymous sites as a result of ascertainment bias in the HapMap dataset.

With the HapMap dataset it is still possible to observe that non-synonymous sites have an excess of rare derived alleles relative to synonymous sites. Similarly, NVRs in CNEs have an excess of rare derived alleles relative to RVRs. However it is impossible to determine that NVRs in CNEs are under stronger purifying selection than non-synonymous sites because the derived allele frequency spectra at both types of noncoding sequence (conserved and nonconserved) are biased downward relative to the spectra at both types of coding sequence.

These discrepancies could be explained by a bias in the data toward coding SNPs due to the ascertainment procedure in the HapMap Project, given that it was designed to capture common variants. In the HapMap Project, variants with a minor allele frequency of >0.05 in a panel of individuals with African, European and Asian ancestry were given preference. A majority of rare variation that is private to a specific population is lost in this way resulting in a large number of rare variants in non-exonic regions being excluded (reviewed in Teo et al., 2010). Preference was also given to validated SNPs in the HapMap Pilot Project (Hammer et al., 2013). Most noncoding variants are unlikely to be validated by other studies given the focus on mutation screening in exomes and there may still be a great number of singletons - i.e. present only in a single copy in the sampled population - being discarded as false positives by rigorous filtering. Furthermore, re-sequencing was done in select ENCODE regions with rare variation in these regions being captured better. CNEs do not overlap any of the 15 ENCODE regions resequenced in Phases I and III of the HapMap Project. Various methods for ascertainment correction are employed because the bias in the sampling strategy including choice of variant discovery panel affects association studies pertaining to complex disorders (Clark et al., 2005).

Although in comparison to the HapMap data, the low coverage data of the 1000 Genomes Project reveal a less biased pattern in the derived allele frequency spectra (as described earlier in this text), the number of individuals genotyped is still heavily biased toward variants in coding regions (Figure 8). Therefore a large proportion of the patterns in rare alleles outside of protein-coding regions may only be as good as the accuracy of the imputation algorithms used to fill in the missing data.

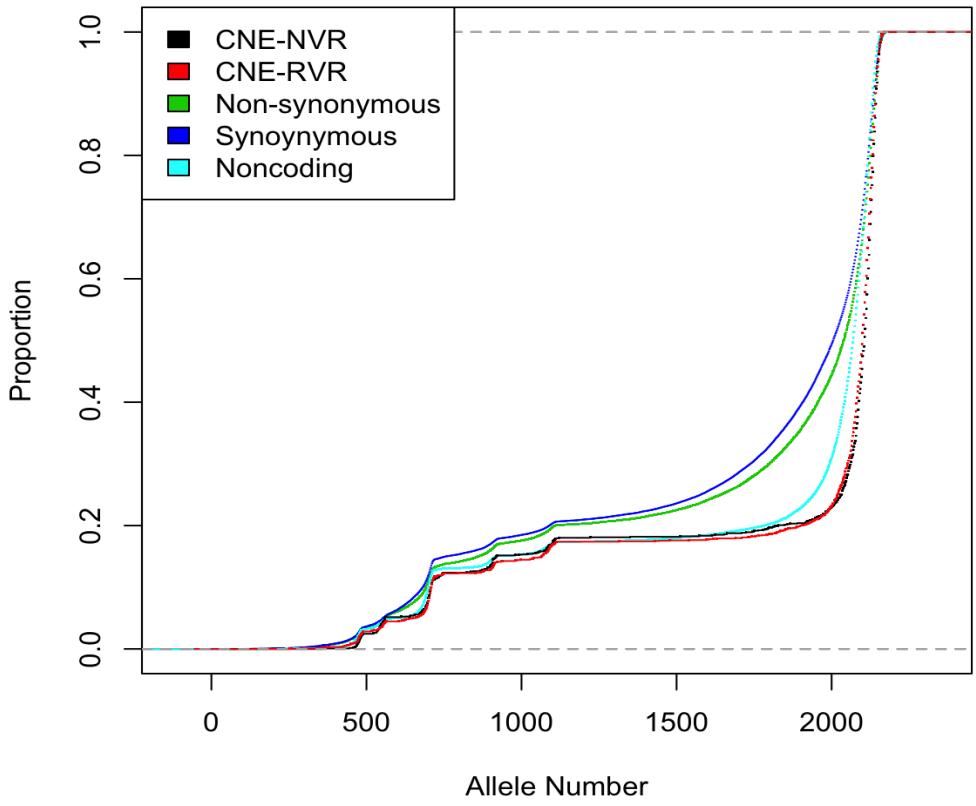

Figure 8: Cumulative Distribution of Number of Individuals Surveyed. The 1000 Genomes Project low coverage dataset shows a bias toward coding variants in terms of the total number of individuals genotyped.

The various cutoffs used by different genotyping platforms to call variants affect conclusions on population stratification human genomic studies (Albrechsten et al., 2010). Since data from the different HapMap populations have been pooled together in

this study, population stratification should not have an effect because any ascertainment bias affects all genomic regions equally. An effect is seen in the different genomic regions because rare alleles especially outside of coding regions were less likely to be called. The results are clearer with the 1000 Genomes dataset because, albeit at low coverage, the whole genome sequencing approach captures a greater number of rare variants outside of the exome.

3.4. Conclusion

The current focus on exome-wide sequencing (rather than genome-wide) may be justified in research projects which are attempting to identify causal variants in diseases/disorders that follow Mendelian patterns of inheritance (Bras et al., 2011; Guerreiro et al., 2012; Hammer et al., 2013). However, the study of complex genetic diseases/disorders, for example developmental disorders determined by perturbation of regulatory networks, warrants either whole genome sequencing or targeted re-sequencing of putative regulatory regions to identify alleles that contribute to an increased risk of occurrence. Variant discovery pipelines in many genome-wide re-sequencing projects discard noncoding variants altogether resulting in potentially important data being lost. Because evolutionarily conserved noncoding DNA represents a small fraction of the vast noncoding landscape, the addition of loci spanning such regions to existing exome selection strategies may be particularly valuable.

I have combined deep, historical phylogenetic footprinting with the occurrence of SNPs and their derived allele frequencies in human populations to identify two classes of sites (NVRs and RVRs) in CNEs that experience different effects of selective pressure. The approach of combining phylogenetic footprinting with population genomics is more

effective than either method alone in identifying evolutionarily conserved sites that are more important than others within a conserved region and that may be vital to defining and maintaining functionality of the element.

3.5. Future work

1. Correcting for ancestral misclassification in the site-frequency spectra by using the method of Hernandez et al., (2011) might result in an estimate of the mean of the gamma distribution of fitness effects in YRI that is closer to the value observed in Boyko et al., (2008).
2. Future work should also focus on mapping transcription factor binding sites from ChIP-Seq and other data to CNEs to further explore the relationship between deep evolutionary conservation and binding site degeneracy, paving the way for a better understanding of the role of mutations in regulatory regions in genetic disease. A clearer picture of whether overlapping known transcription factor binding sites might explain the seemingly contradictory presence of NVRs in CNEs may also emerge.

3.6. References

1. Abecasis GR, Altshuler D, Auton A, Brooks LD, Durbin RM, et al. (2010) A map of human genome variation from population-scale sequencing. Nature 467: 1061-1073.
2. Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, et al. (2012) An integrated map of genetic variation from 1,092 human genomes. Nature 491: 56-65.
3. Albers CA, Newbury-Ecob R, Ouwehand WH, Ghevaert C (2013) New insights into the genetic basis of TAR (thrombocytopenia-absent radii) syndrome. Curr Opin Genet Dev.
4. Albrechtsen A, Nielsen FC, Nielsen R (2010) Ascertainment biases in SNP chips affect measures of population divergence. Mol Biol Evol 27: 2534-2547.
5. Altshuler DM, Gibbs RA, Peltonen L, Dermitzakis E, Schaffner SF, et al. (2010) Integrating common and rare genetic variation in diverse human populations. Nature 467: 52-58.
6. Asthana S, Noble WS, Kryukov G, Grant CE, Sunyaev S, et al. (2007) Widely distributed noncoding purifying selection in the human genome. Proc Natl Acad Sci U S A 104: 12410-12415.
7. Bejerano G, Pheasant M, Makunin I, Stephen S, Kent WJ, et al. (2004) Ultraconserved elements in the human genome. Science 304: 1321-1325.
8. Benko S, Fantes JA, Amiel J, Kleinjan DJ, Thomas S, et al. (2009) Highly conserved non-coding elements on either side of SOX9 associated with Pierre Robin sequence. Nat Genet 41: 359-364.
9. Berezikov E, Guryev V, Plasterk RH, Cuppen E (2004) CONREAL: conserved regulatory elements anchored alignment algorithm for identification of transcription factor binding sites by phylogenetic footprinting. Genome Res 14: 170-178.
10. Birney E, Stamatoyannopoulos JA, Dutta A, GuigÛ R, Gingeras TR, et al. (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature 447: 799-816.
11. Boffelli D, McAuliffe J, Ovcharenko D, Lewis KD, Ovcharenko I, et al. (2003) Phylogenetic shadowing of primate sequences to find functional regions of the human genome. Science 299: 1391-1394.
12. Boyko AR, Williamson SH, Indap AR, Degenhardt JD, Hernandez RD, et al. (2008) Assessing the evolutionary impact of amino acid mutations in the human genome. PLoS Genet 4: e1000083.
13. Bras JM, Singleton AB (2011) Exome sequencing in Parkinson's disease. Clin Genet 80: 104-109.
14. Butter F, Davison L, Viturawong T, Scheibe M, Vermeulen M, et al. (2012) Proteome-wide analysis of disease-associated SNPs that show allele-specific transcription factor binding. PLoS Genet 8: e1002982.
15. Cai JJ, Macpherson JM, Sella G, Petrov DA (2009) Pervasive hitchhiking at coding and regulatory sites in humans. PLoS Genet 5: e1000336.
16. Clark AG, Hubisz MJ, Bustamante CD, Williamson SH, Nielsen R (2005) Ascertainment bias in studies of human genome-wide polymorphism. Genome Res 15: 1496-1502.
17. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, et al. (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet 43: 491-498.
18. Dowell RD (2010) Transcription factor binding variation in the evolution of gene regulation. Trends Genet 26: 468-475.
19. Drake JA, Bird C, Nemesh J, Thomas DJ, Newton-Cheh C, et al. (2006) Conserved noncoding sequences are selectively constrained and not mutation cold spots. Nat Genet 38: 223-227.
20. Dubchak I, Ryaboy DV (2006) VISTA family of computational tools for comparative analysis of DNA sequences and whole genomes. Methods Mol Biol 338: 69-89.
21. Elgar G, Vavouri T (2008) Tuning in to the signals: noncoding sequence conservation in vertebrate genomes. Trends Genet 24: 344-352.
22. Epstein DJ (2009) Cis-regulatory mutations in human disease. Brief Funct Genomic Proteomic 8: 310-316.
23. Ethayathulla AS, Tse PW, Monti P, Nguyen S, Inga A, et al. (2012) Structure of p73 DNA-binding domain tetramer modulates p73 transactivation. Proc Natl Acad Sci U S A 109: 6066-6071.
24. Eyre-Walker A, Woolfit M, Phelps T (2006) The distribution of fitness effects of new deleterious amino acid mutations in humans. Genetics 173: 891-900.
25. Flicek P, Ahmed I, Amode MR, Barrell D, Beal K, et al. (2013) Ensembl 2013. Nucleic Acids

Res 41: D48-55.
26. Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, et al. (2007) A second generation human haplotype map of over 3.1 million SNPs. Nature 449: 851-861.
27. Guerreiro RJ, Lohmann E, Kinsella E, Br·s JM, Luu N, et al. (2012) Exome sequencing reveals an unexpected genetic cause of disease: NOTCH3 mutation in a Turkish family with Alzheimer's disease. Neurobiol Aging 33: 1008.e1017-1023.
28. Halligan DL, Oliver F, Guthrie J, Stemshorn KC, Harr B, et al. (2011) Positive and negative selection in murine ultraconserved noncoding elements. Mol Biol Evol 28: 2651-2660.
29. Hammer MB, Eleuch-Fayache G, Gibbs JR, Arepalli SK, Chong SB, et al. (2013) Exome sequencing: an efficient diagnostic tool for complex neurodegenerative disorders. Eur J Neurol 20: 486-492.
30. Hedges SB, Dudley J, Kumar S (2006) TimeTree: a public knowledge-base of divergence times among organisms. Bioinformatics 22: 2971-2972.
31. Hughes AL, Packer B, Welch R, Bergen AW, Chanock SJ, et al. (2003) Widespread purifying selection at polymorphic sites in human protein-coding loci. Proc Natl Acad Sci U S A 100: 15754-15757.
32. Consortium IH (2005) A haplotype map of the human genome. Nature 437: 1299-1320.
33. Kasowski M, Grubert F, Heffelfinger C, Hariharan M, Asabere A, et al. (2010) Variation in transcription factor binding among humans. Science 328: 232-235.
34. Kasprzyk A (2011) BioMart: driving a paradigm change in biological data management. Database (Oxford) 2011: bar049.
35. Katzman S, Kern AD, Bejerano G, Fewell G, Fulton L, et al. (2007) Human genome ultraconserved elements are ultraselected. Science 317: 915.
36. Keller I, Bensasson D, Nichols RA (2007) Transition-transversion bias is not universal: a counter example from grasshopper pseudogenes. PLoS Genet 3: e22.
37. Lettice LA, Heaney SJ, Purdie LA, Li L, de Beer P, et al. (2003) A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. Hum Mol Genet 12: 1725-1735.
38. Lettice LA, Hill AE, Devenney PS, Hill RE (2008) Point mutations in a distant sonic hedgehog cis-regulator generate a variable regulatory output responsible for preaxial polydactyly. Hum Mol Genet 17: 978-985.
39. Li H (2011) Tabix: fast retrieval of sequence features from generic TAB-delimited files. Bioinformatics 27: 718-719.
40. McDaniell R, Lee BK, Song L, Liu Z, Boyle AP, et al. (2010) Heritable individual-specific and allele-specific chromatin signatures in humans. Science 328: 235-239.
41. McEwen GK, Woolfe A, Goode D, Vavouri T, Callaway H, et al. (2006) Ancient duplicated conserved noncoding elements in vertebrates: a genomic and functional analysis. Genome Res 16: 451-465.
42. Minovitsky S, Stegmaier P, Kel A, Kondrashov AS, Dubchak I (2007) Short sequence motifs, overrepresented in mammalian conserved non-coding sequences. BMC Genomics 8: 378.
43. Mu XJ, Lu ZJ, Kong Y, Lam HY, Gerstein MB (2011) Analysis of genomic variation in non-coding elements using population-scale sequencing data from the 1000 Genomes Project. Nucleic Acids Res 39: 7058-7076.
44. Parker HJ, Piccinelli P, Sauka-Spengler T, Bronner M, Elgar G (2011) Ancient Pbx-Hox signatures define hundreds of vertebrate developmental enhancers. BMC Genomics 12: 637.
45. Philippakis AA, He FS, Bulyk ML (2005) Modulefinder: a tool for computational discovery of cis regulatory modules. Pac Symp Biocomput: 519-530.
46. Piganeau G, Eyre-Walker A (2003) Estimating the distribution of fitness effects from DNA sequence data: implications for the molecular clock. Proc Natl Acad Sci U S A 100: 10335-10340.
47. Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, et al. (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res 15: 1034-1050.
48. Singleton AB (2011) Exome sequencing: a transformative technology. Lancet Neurol 10: 942-946.
49. Spivakov M, Akhtar J, Kheradpour P, Beal K, Girardot C, et al. (2012) Analysis of variation at transcription factor binding sites in Drosophila and humans. Genome Biol 13: R49.
50. Stephan W (2010) Genetic hitchhiking versus background selection: the controversy and its implications. Philos Trans R Soc Lond B Biol Sci 365: 1245-1253.
51. Stormo GD, Zhao Y (2010) Determining the specificity of protein-DNA interactions. Nat Rev Genet 11: 751-760.

52. Tagle DA, Koop BF, Goodman M, Slightom JL, Hess DL, et al. (1988) Embryonic epsilon and gamma globin genes of a prosimian primate (Galago crassicaudatus). Nucleotide and amino acid sequences, developmental regulation and phylogenetic footprints. J Mol Biol 203: 439-455.
53. Tamura K, Peterson D, Peterson N, Stecher G, Nei M, et al. (2011) MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol Biol Evol 28: 2731-2739.
54. Teo YY, Small KS, Kwiatkowski DP (2010) Methodological challenges of genome-wide association analysis in Africa. Nat Rev Genet 11: 149-160.
55. Todorova A, Halliger-Keller B, Walter MC, Dabauvalle MC, Lochm¸ller H, et al. (2003) A synonymous codon change in the LMNA gene alters mRNA splicing and causes limb girdle muscular dystrophy type 1B. J Med Genet 40: e115.
56. Woolfe A, Goode DK, Cooke J, Callaway H, Smith S, et al. (2007) CONDOR: a database resource of developmentally associated conserved non-coding elements. BMC Dev Biol 7: 100.
57. Woolfe A, Goodson M, Goode DK, Snell P, McEwen GK, et al. (2005) Highly conserved non-coding sequences are associated with vertebrate development. PLoS Biol 3: e7.

Chapter 4

4. Patterns of genetic differentiation in CNEs

In Chapter 2 I explored patterns of substitutions between species to identify sites that may have adaptive functions within Conserved Noncoding Elements (CNEs). In Chapter 3 I used a combination of phylogenetic footprinting and phylogenetic shadowing to distinguish two types of sites within CNEs that are subject to different levels of constraint: phylogenetically invariant sites (CNE NVRs) and phylogenetically variant sites (CNE RVRs). Further, I exploited patterns of derived allele frequencies within a global human population to determine that purifying selection is most consistent in phylogenetically invariant positions in CNEs by contrasting their allele frequency spectrum with synonymous (nearly neutral) and non-synonymous sites (under purifying selection). However, there may be more to be gleaned about the evolution of CNEs by interrogating patterns of variance in allele frequencies at selected sites across the different human subpopulations. In this chapter I investigate the patterns of genetic differentiation among populations at polymorphic sites in CNEs and contrast them with patterns observed at synonymous and non-synonymous sites.

4.1. Introduction

The study of genetic diversity between and within populations is of great interest for a number of reasons. The patterns can reveal clues about the evolutionary history of a species and also about ongoing processes – such as gene flow and adaptation to different environments. For example, differences in allele frequencies at coding loci in geographically distinct populations of *Drosophila* have been exploited to explore the genetics underlying morphology and the process of speciation (Ayala et al., 1974). These approaches can be extended to the analysis of other types of genetic variation including single nucleotide polymorphisms (SNPs), microsatellite markers and copy number variants (CNVs) between and within populations.

Under the assumption of neutrality, patterns of genetic differentiation between populations are determined by demography, in particular the balance between genetic drift and gene flow. Consider a series of populations that are isolated: that is, there is no gene flow between them. Under the action of genetic drift, different alleles would attain high frequencies in the different populations, and with time would become fixed. This process would establish a certain amount of differentiation between the populations. On the other hand, gene flow between populations would tend to reduce this differentiation by evening out the differences in allele frequency. Hence the relative influence of genetic drift and gene flow are reflected in the observed genetic differentiation.

One measure of this differentiation is Wright's F_{ST} (Wright, 1951). It can be understood as a correlation: consider drawing two alleles from different individuals in the same population. If there is genetic differentiation among populations, the second individual is more likely to carry a matching allele, because of the shared ancestry unique to that

population. This genetic correlation is quantified by the parameter F_{ST}. The value of the parameter can be estimated from the allele frequency data from the populations, for example by the Weir and Cockerham's statistic Theta (Weir and Cockerham, 1984).

One valuable property of this statistic is that its expected value is the same for different neutral loci (with some caveats, see below). This common value of F_{ST} reflects the common demographic history shared by the loci: if there has been a high degree of migration/gene flow between the two populations being compared, a low value of F_{ST} is expected, and *vice versa*. This expectation of a constant F_{ST} has been exploited to detect the action of selection, since different types of selection can either increase or decrease the differentiation between populations (reviewed in Novembre and Rienzo, 2009).

Selection tending to maintain the same allele frequency in different populations will reduce genetic differentiation among populations. One example of such balancing selection occurs at the beta-globin gene, where an allele confers partial resistance to malaria in heterozygotes, but sickle cell anemia in homozygotes (Pasvol et al., 1978). The key characteristic of this type of selection is not that one allele is favoured everywhere, or selected against, but rather that there is an equilibrium frequency. If genetic drift displaces the allele frequency from this equilibrium, selection will tend to return it. In the case of the sickle cell polymorphism, the selection is sufficiently strong that it can also be detected in the genotypes: an excess of heterozygotes.

This combination of heterozygote excess and even allele frequencies has been seen at a few other loci across the genome (Andres et al., 2009), particularly the human histocompatibility system (HLA) locus (Hedrick and Thomson, 1983). However, even

in the case of balancing selection, genetic differentiation need not be small. There can actually be atypically large differentiation between populations if the selective regime (and hence the equilibrium frequency) differs from place-to-place. This type of pattern is seen over larger spatial scales at the sickle-cell locus; it can be convincingly explained, since the frequency of the resistant allele is high only in populations with a history of exposure to malaria, such as tropical Africa and India (Pasvol et al., 1978).

Direct selection in favour of different alleles in different populations (selection for local adaptations) can also result in increased differentiation among populations, For example, alleles at the LCT locus, associated with the ability to digest lactose into adulthood (lactase persistence), are at high frequencies in European-derived populations, while being rare in African and East Asian populations. This pattern shows a plausible correlation with the geographical distribution of dairy farming (Bersaglieri et al., 2004). In another example, the allele frequency in genes associated with variation in skin pigmentation among populations, show stark differences in populations of African, European and Asian descent (Lao et al., 2007) – a pattern that may have evolved in response to varying levels of UVB radiation globally (Jablonski and Chaplin, 2000). Additionally, genetic differentiation is a means for exploring the variation in disease-susceptibility in different populations to better understand the genetic basis of human disease (Tishkoff and Verelli, 2003).

Beaumont and Nichols (1996) developed a method to identify loci that might be affected by selection, based on high or low values of F_{ST} that would be unlikely if they were neutral. Their method simulates the distribution of F_{ST} expected under neutral evolution. This outlier detection method against a backdrop of F_{ST} at neutrally evolving loci is

implemented in the software LOSITAN (Antao et al., 2008) and an example using a subset of human non-synonymous sites is illustrated in Figure 1.

However, the distribution of F_{ST} is conditional on heterozygosity in a number of ways: the variance in F_{ST} values is higher at intermediate heterozygosities because genetic drift has a larger effect on intermediate allele frequencies, whereas there are more subtle effects on the mean value of F_{ST} for a given demographic history at high and low heterozygosities. These include a downward bias in the estimate of F_{ST} when one allele is very rare, and the effects of mutation in depressing F_{ST} at highly heterozygous loci (Beaumont and Nichols, 1996).

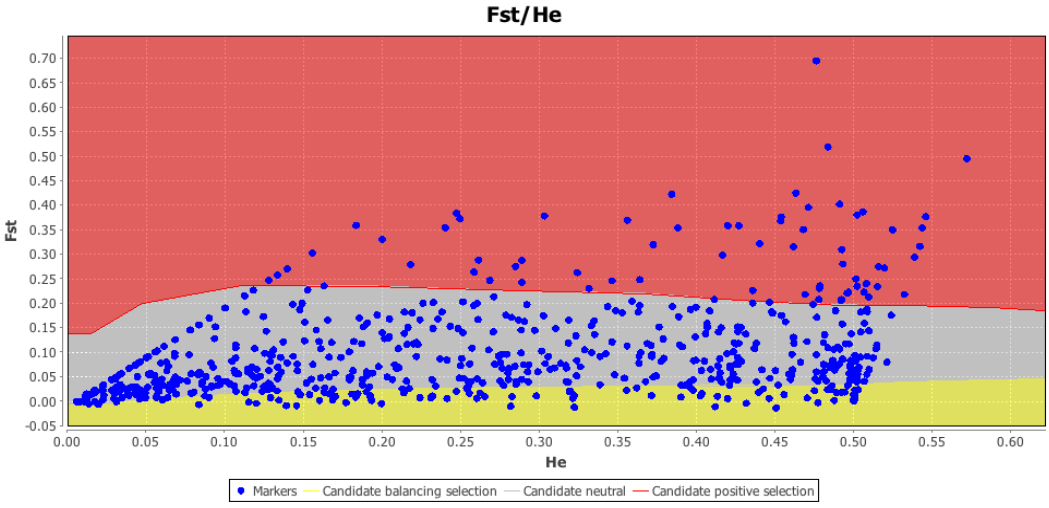

Figure 1: Beaumont-Nichols plot (in LOSITAN) to detect outlier loci in a subset of human non-synonymous sites. The F_{ST} value for each site (blue symbols) lies against a backdrop of the quantiles generated by the null distribution. Loci with low F_{ST} are usually interpreted as subject to balancing selection whereas loci with high F_{ST} are interpreted as subject to disparate directional selection in different populations.

More recent methods of quantifying genetic differentiation are based on a likelihood function which is appropriate down to very low allele frequencies, these are founded on genetic models (Balding and Nichols 1995) in which the variation of allele frequencies are beta-binomial distributed (or multinomial-Dirichlet distributed in the case of multi-allelic loci). A recent implementation (Foll and Gaggiotti, 2006) also detects outlying loci in a manner analogous to LOSITAN although it does not separate the effects of mutation and drift.

While the above mentioned methods are capable of identifying outlier loci that are under balancing and directional selection, it is unclear where in the spectrum sites under purifying selection and hence at low allele-frequencies lie and whether there is a systematic difference in the pattern of differentiation at negatively selected sites and neutral sites at similar allele-frequencies (perhaps useful in prioritising variants in disease association). Sites thought to be under purifying selection have been reported to have an excess of low F_{ST} variants (Barreiro et al., 2008). The lack of strong population differentiation at such sites, relative to that expected under neutrality, was attributed by these authors to the effects of purifying selection that keep deleterious alleles at low frequencies. Barreiro et al. (2008) noted this pattern, in the appendix to their paper, but even so state 'negative selection has globally reduced population differentiation at amino acid–altering mutations, particularly in disease-related genes'. However, there is a little-appreciated bias in the estimation of F_{ST}, which depresses estimates at loci with very low frequencies of the rarest allele (it can be seen in the graphs of expected F_{ST} in Beaumont & Nichols 1996 at low heterozygosity values). Therefore, an excess of rare variants in a sample can give rise to an excess of low F_{ST} values.

In this chapter I explore the patterns of differentiation in allele frequency among human populations at CNEs for evidence of distinctive patterns of F_{ST} at sites under purifying selection; comparing the patterns with other loci thought to be essentially neutral (synonymous sites in protein coding loci and non-coding sites) and with loci thought to be under purifying selection (non-synonymous sites in protein coding loci). I use two broad approaches: one based on the distribution of Weir and Cockerham's Theta (an estimator of F_{ST}), and a second approach based on parameters fitted by beta-binomial and quasi-binomial models.

4.2. Methods

Extracting variants

The genotypes of the variants in the different categories were parsed from the 1000 Genomes low coverage vcf file "ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20110521/supporting/ALL.wgs.project_consensus_vqsr2b.20101123.snps.low_coverage.sites.vcf.gz", using the tabix tool with variant coordinates as input. The variants were subdivided into the different populations using the 'vcf-subset' script in the Vcftools package (Danecek et al., 2011) using the individual IDs from the population panel for the 1000 Genomes Project (see Appendix). The genomic sites were those chosen for the analyses in Chapters 2 and 3.

Weir and Cockerham's Theta

An algorithm for calculating Weir and Cockerham Theta, as described in Weir and Cockerham, (1984), was coded in R (R Core Development Team, 2012).

Statistical analyses

The analysis of the allele counts in each population was carried out in the following steps:

1. For each site in the analysis, Weir and Cockerham's Theta was calculated to characterise the variation in allele frequencies about the expectation given by the global mean (i.e. not accounting for demographic history).

Accounting for demographic history in calculating expected frequency

2. For subsequent analyses an improved estimate of the expected allele frequencies in a subpopulation, taking into account demographic history, was used. This estimate was obtained by considering that the allele frequencies in any one population will be closer to some populations (typically nearby populations), and more differentiated from others (typically more distant populations). This relationship was characterized for each population in turn, by a multiple regression of its derived allele frequencies against the frequencies in all other populations. The model fitted one set of regression coefficients for the whole genome (a single coefficient for each other population). These coefficients were then used to calculate the expected frequency of the derived allele at each site in contrast to using an average across all populations (the global average). The logistic regression was carried out using the 'glm' function of R (using a quasi-binomial error distribution to accommodate fluctuations in allele frequency due to drift; the covariates were the Laplace estimates of derived allele frequency in each other population). The expected values were extracted from the model using the 'fitted()' function of R.

Exploring deviation from expected frequencies by site category

3. In order to detect any systematic deviation of derived allele frequencies from

these expectations in the different categories of sites (CNEs vs synonymous vs non-synonymous etc.), the observed derived allele frequencies were regressed onto the expected frequencies using the type of site as a predictor (again using the glm function of R, assuming a quasi-binomial error). Note that using only type of site as a predictor assumes that each subpopulation deviates from the expected frequency by the same degree. The model fits an intercept and slope for each variant in each site category. Significant differences in the fitted coefficients for each site category were obtained from an analysis of variance.

Matching allele-frequency distributions across categories of sites

4. Because both Steps 1 and 3 might be affected by the allele frequency distribution, they were repeated on a specially selected subset of the data, in which the allele frequency distributions were identical for the five categories of site: the global derived allele counts and number of variants were matched to the derived allele counts and number of variants in CNE-NVRs as these comprised the smaller dataset.
5. Although the quasi-binomial distribution accommodates variation in the allele frequency due to drift, it does not take into account the smaller variance observed at low allele frequencies. In contrast, the over dispersion of the allele frequencies around the expectation due to genetic drift where a smaller variance is observed at low allele frequencies, is expected to follow a beta-binomial distribution which is implemented in the vgam package in R (Yee, 2013). The function is less stable than the quasi-binomial option of glm; and the analysis did not converge for the full dataset, but it was suitable for the reduced dataset produced in step 4. Additionally, the beta-binomial option of the vglm package in R fits a 'correlation parameter', which is directly equivalent to logit(F_{ST}).

Exploring deviation from expected frequencies by site category and subpopulation

6. Since the observed allele frequency in each population and type of locus might potentially have a different degree of differentiation from the expected frequency, separate estimates of the regression coefficients (intercept, slope and logit(Fst)) were obtained by analyzing each combination in turn.

ANOVA

7. An analysis of variance using the 'anova()' function in R was carried out to test for significant differences in logit(F_{ST}) among categories of site, and among subpopulations. Using the 'lm()' function of R, the parameter estimates of logit(Fst)) obtained from Step 6 were used to fit a linear model weighted by the inverse of their standard error given the type of site and subpopulation combination as predictors.
8. The analysis was repeated forcing the regression between observed and expected allele frequencies to have a slope of exactly 1.0 using the ‘offset’ option in the vglm package, which is equivalent to characterizing the variation around the green 1:1 line (i.e. a perfect fit between the observed and expected frequencies).

4.3. Results and Discussion

4.3.1. Exploring population differentiation using Weir and Cockerham's Statistic Theta

When all variants are included in estimating F_{ST} using Weir and Cockerham's Theta, an excess of sites showing low F_{ST} is seen in my results for the categories thought to be under strong purifying selection (Figure 2A). This pattern is similar to that observed at non-synonymous sites by Barreiro et al. (2008).If I restrict my analysis to rare alleles (frequency <0.05), the depression of F_{ST} remained significant at CNEs and non-synonymous sites relative to synonymous sites (Kolmogorov-Smirnov test: *p*-values 0.0002631 and < 2.2e-16 respectively). However, if I matched the global allele frequencies more precisely, there is no evidence of a depression in F_{ST} (Figure 2B, 'subset') at sites under strong purifying selection. In this case a subset of sites from each category was randomly selected (without replacement) to exactly match the derived allele counts in the CNE-NVR dataset (given that it was the smallest dataset).

This result suggests that any difference in F_{ST} values, as estimated by the Weir and Cockerham formula, can actually be attributed to differences in the global allele frequencies. For example, the profile of F_{ST} in each category reflects the proportion of derived alleles present in the sample (Figure 3). Thus, there is no evidence for a difference in the variation of the allele frequencies around the global average in any of the categories found using the standard estimates of F_{ST}.

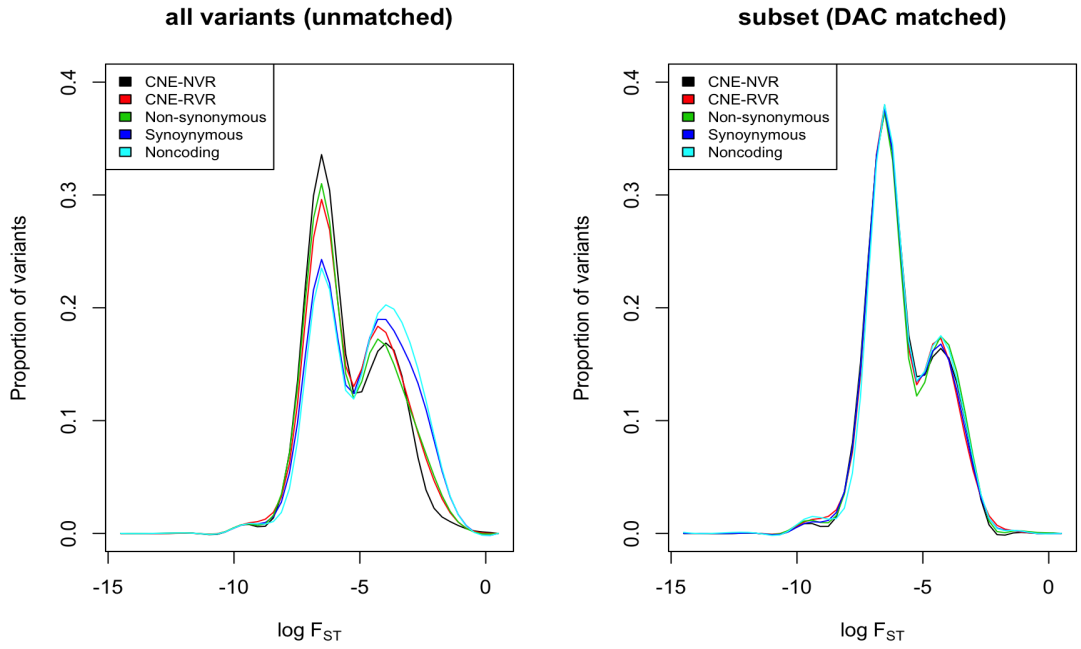

Figure 2: Comparison of F_{ST} estimated using Weir and Cockerham Theta in the different categories. A) Sites thought to be under purifying selection exhibit an excess of low F_{ST} sites when the samples are not matched for global allele-frequency. B) An excess of low F_{ST} sites is not observed at negatively selected sites when samples are matched for the exact derived allele count at CNE-NVR sites.

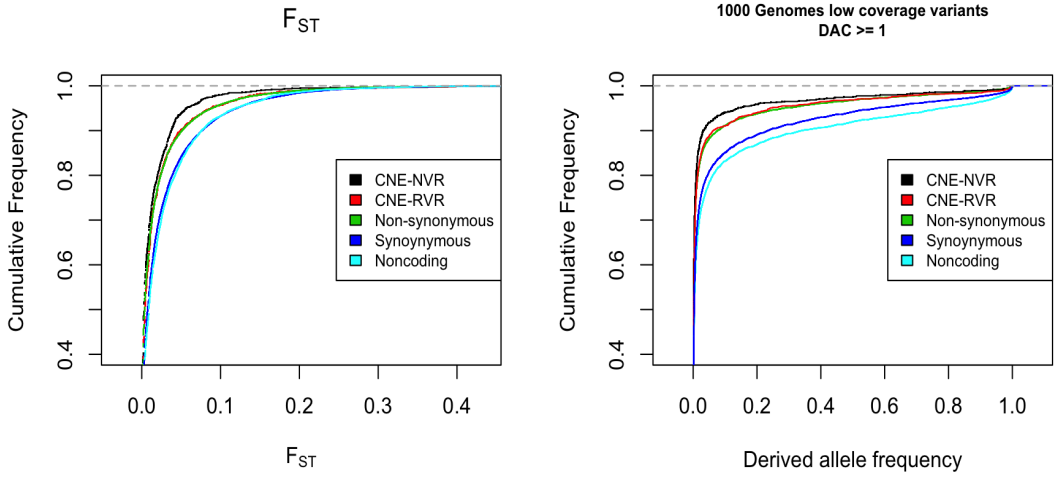

Figure 3: Comparison of derived allele-frequency and F_{ST} profiles. A) The cumulative frequency distribution of absolute F_{ST} in the different categories of sites. Sites thought to be under purifying selection exhibit an excess of low F_{ST} sites. B) The cumulative frequency distribution of derived allele-frequency in the different categories of sites. Sites thought to be under purifying selection exhibit an excess of low frequency derived alleles, resulting in a similar profile in F_{ST}.

4.3.2.Exploring population differentiation using statistical models

The conventional estimate of the expected frequency of an allele in a subpopulation is the global allele-frequency. The conventional method of estimating population differentiation relies on how much the observed allele frequency in a subpopulation deviates from the global mean (as in the case of Weir and Cockerham's estimate of F_{ST}). However, the history of human colonisation around the globe, and the current migration between subpopulations has resulted in spatial autocorrelation: nearby populations can be expected to be more similar in allele frequency at a given locus. To account for this pattern, I obtained estimates of allele frequency for each subpopulation by regression against the Laplace estimates of allele frequencies across all sites in all other populations (See Step 2 in Methods: Statistical analyses). As anticipated, these estimates are much closer to the observed allele frequencies than the global allele frequencies (Figure 4), showing that the approach allowed for the spatial autocorrelation with just a single set of regression coefficients (one for each other population) for the whole genome.

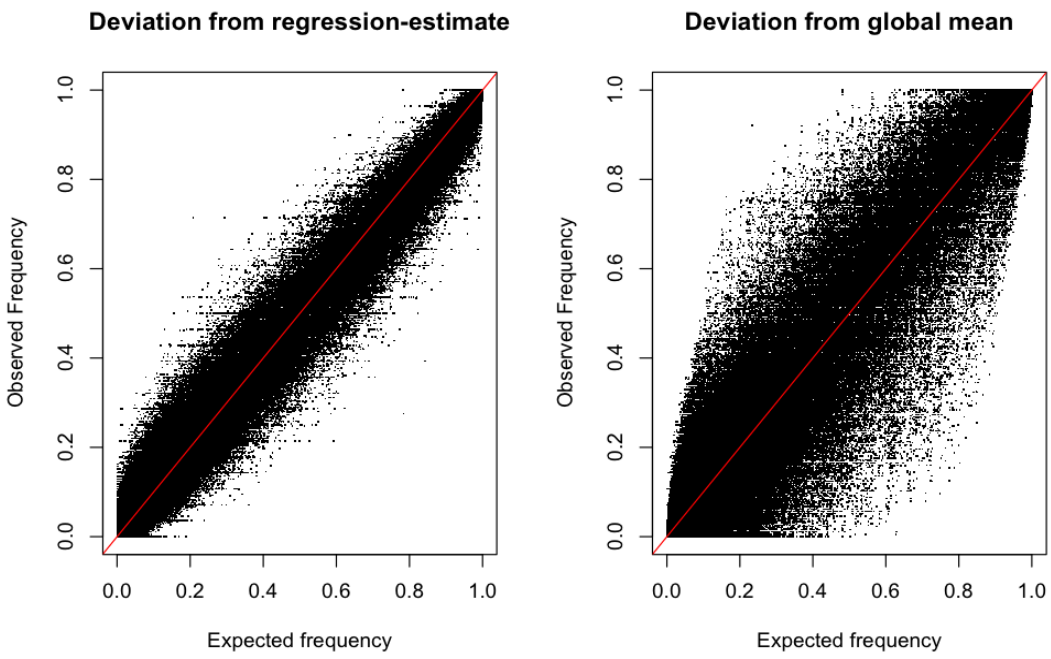

Figure 4: Comparison of regression-derived estimate and global average as predictors of expected subpopulation allele-frequencies.

Deviation from expected frequencies by site category

Next, I looked for systematic differences in the pattern of observed allele frequency in in the different classes of sites using a logistic regression (See Step 3 in Methods: Statistical analyses).

Significant differences were observed between the different classes of sites (Figure 5): all sites thought to be under purifying selection deviate the most from the expected allele-frequencies, with CNE-NVRs (Nonvariable sites) showing the largest deviation from expectation. In contrast, sites deemed to be evolving neutrally (noncoding and synonymous sites) show little deviation from expected allele-frequencies. This result indicates that differences in site-specific selection may have an effect on the underlying distribution of derived alleles in the subpopulations.

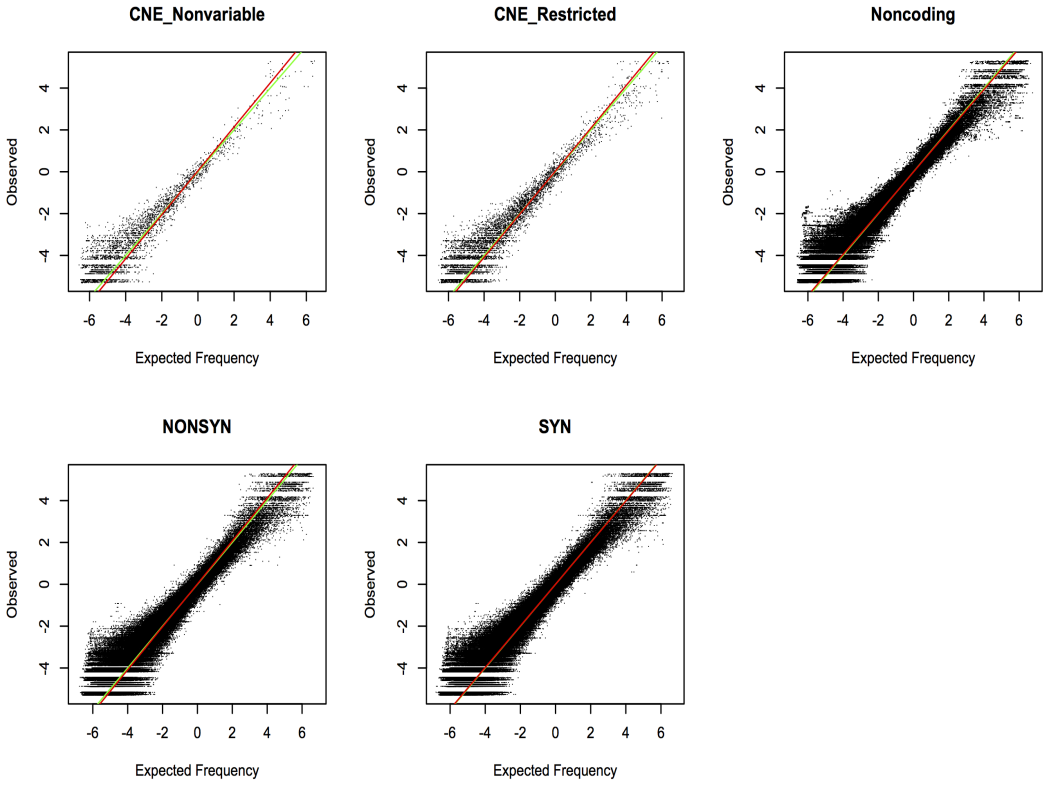

Figure 5: Deviation of observed subpopulation allele-frequencies from the expectation in the different categories. The allele frequencies on both x and y axes are on a logit scale. The green line represents the situation where the observed allele-frequencies exactly match the expectation. In red is the trend line obtained by regressing the observed allele-frequencies against the expectation. Allele-frequencies at CNE-NVRs (CNE_Nonvariable) (under strong purifying selection), tend to deviate the most from expected allele-frequencies whereas Noncoding and synonymous sites (SYN) (evolving neutrally), deviate the least.

We observed earlier that the distribution of Weir and Cockerham's F_{ST} estimate was influenced by the proportion of sites with a given allele-frequency in the data. Similarly, when the regression model was applied to a subset of the data (Step 4 in Methods: Statistical analyses) where the sample size and allele counts of variants in the different categories of sites were matched to the sample size and allele counts of CNE-NVR variants, there were only very minor differences, barely perceptible (Figure 6), albeit formally significant. These differences in regression coefficients were detected in a statistical model that assumed a quasi-binomial error to accommodate fluctuations in the expected allele-frequencies resulting from genetic drift. However, F_{ST} estimates are conditional on heterozygosity where estimates are depressed at low allele-frequencies. Therefore, the analysis was repeated using a beta-binomial error distribution around the expected allele-frequencies (see Step 5 in Methods: Statistical analyses), which takes into account the smaller variance in the effect of genetic drift at low allele-frequencies and the same pattern was observed (results not shown).

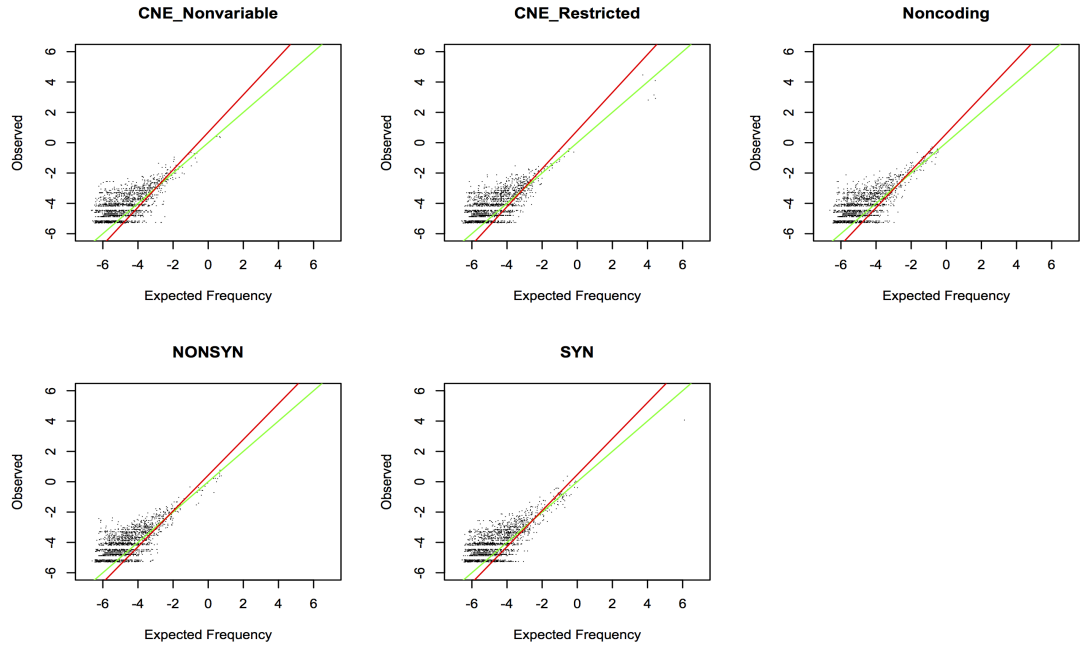

Figure 6: The deviation in allele-frequencies in different classes of sites using matched allele counts and sample sizes. The allele frequencies on both x and y axes are on a logit scale. The differences in the pattern of deviation of the fitted trend line from the expectation is barely perceptible across the different classes of sites.

Deviation from expected frequencies by site category and subpopulation

Both variations of the model so far assumed the same value of F_{ST} for each population i.e. the fluctuation from expected allele-frequencies is constant in each subpopulation. A better model would be to allow a different degree of fluctuation from expected allele-frequencies in each subpopulation. The beta-binomial regression in the vgam package estimates the 'correlation' of the beta binomial distribution which is exactly equivalent to the logit(F_{ST}). When different regressions were carried out for each site category and subpopulation using this model, small but formally significant differences were observed as before. However, the estimates of regression coefficients in the different

categories follow no logical pattern; for example, the slope increases in the order Non-synonymous (1.17), Synonymous (1.18), Noncoding (1.22), CNE-NVR (1.23) and CNE-RVR (1.25). The method detected a significant difference in F_{ST} values among populations, but no significant differences between the different categories of sites.

A different measure of any systematic differences among the site categories would be comparing the pattern of expected allele-frequencies in contrast to the deviation of the observed frequencies from the expectation. To achieve this, the analysis was repeated forcing the regression between observed and expected allele frequencies to have a slope of exactly 1.0 using the 'offset' function, which is equivalent to characterizing the variation around the green 1:1 line in Figure 6, rather than around the red line. Once again, although there was a significant difference in F_{ST} values among populations there was no significant difference in the F_{ST} among sites (results not shown).

In light of the above observations it can be concluded that patterns of deviation from expected allele-frequencies cannot be used to distinguish between variants that are under purifying selection and those that are evolving neutrally

although there are clear and significant differences in their global allele frequency spectra. This chapter has explored the additional information that can be obtained by looking at the variation among human subpopulations. There has been no well-developed population genetics theory to predict the effect of strong purifying selection, especially in non-equilibrium situations such as the human allele frequencies resulting from the serial bottlenecks as humans colonised the world (Cavalli-Sforza, 1994).

Here I have shown that the reported depression of F_{ST} (Barreiro et al., 2008) can be attributed to the allele frequencies alone (as also observed by Elhaik, 2012), and could be explained by biases attributable to the mathematical properties of the estimators, which depend on the allele-frequency distribution (Jakobsson et al., 2013). After making appropriate allowances for the allele frequencies, I find no depression of F_{ST} using conventional estimators of F_{ST}, nor using more recently-developed approaches based on the beta-binomial model (Balding and Nichols, 1995). There remain some very subtle significant differences in allele frequency distribution that are formally significant, but they show no consistent pattern (for example the pattern at synonymous and non-coding sites are dissimilar), and may warrant further investigation, although they are probably the result of the minor effects on biased variant calling even in the 1000 genomes data (outlined in Chapter 3).

4.4. References

1. AndrÈs AM, Hubisz MJ, Indap A, Torgerson DG, Degenhardt JD, et al. (2009) Targets of balancing selection in the human genome. Mol Biol Evol 26: 2755-2764.
2. Antao T, Lopes A, Lopes RJ, Beja-Pereira A, Luikart G (2008) LOSITAN: a workbench to detect molecular adaptation based on a Fst-outlier method. BMC Bioinformatics 9: 323.
3. Ayala FJ, Tracey ML (1974) Genetic differentiation within and between species of the *Drosophila willistoni* group. Proc Natl Acad Sci U S A 71: 999-1003.
4. Balding DJ, Nichols RA (1995) A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity. Genetica 96: 3-12.
5. Barreiro LB, Laval G, Quach H, Patin E, Quintana-Murci L (2008) Natural selection has driven population differentiation in modern humans. Nat Genet 40: 340-345.
6. Beaumont MA, Nichols RA (1996) Evaluating Loci for Use in the Genetic Analysis of Population Structure. Proceedings of the Royal Society of London Series B: Biological Sciences 263: 1619-1626.
7. Bersaglieri T, Sabeti PC, Patterson N, Vanderploeg T, Schaffner SF, et al. (2004) Genetic signatures of strong recent positive selection at the lactase gene. Am J Hum Genet 74: 1111-1120.
8. Cavalli-Sforza LL, Menozzi P, Piazza A. 1994. The History and Geography of Human Genes
9. Elhaik E (2012) Empirical distributions of F(ST) from large-scale human polymorphism data. PLoS One 7: e49837.
10. Hedrick PW, Thomson G (1983) Evidence for balancing selection at HLA. Genetics 104: 449-456.
11. Hinds DA, Stuve LL, Nilsen GB, Halperin E, Eskin E, et al. (2005) Whole-genome patterns of common DNA variation in three human populations. Science 307: 1072-1079.
12. Jablonski NG, Chaplin G (2000) The evolution of human skin coloration. J Hum Evol 39: 57-106.
13. Jakobsson M, Edge MD, Rosenberg NA (2013) The relationship between F(ST) and the frequency of the most frequent allele. Genetics 193: 515-528.
14. Kimura M, Weiss GH (1964) The Stepping Stone Model of Population Structure and the Decrease of Genetic Correlation with Distance. Genetics 49: 561-576.
15. Krithika S, Maji S, Vasulu TS (2007) Intertribal and temporal allele-frequency variation at the ABO locus among Tibeto-Burman-speaking Adi subtribes of Arunachal Pradesh, India. Hum Biol 79: 355-362.
16. Lao O, de Gruijter JM, van Duijn K, Navarro A, Kayser M (2007) Signatures of positive selection in genes associated with human skin pigmentation as revealed from analyses of single nucleotide polymorphisms. Ann Hum Genet 71: 354-369.
17. Marth G, Schuler G, Yeh R, Davenport R, Agarwala R, et al. (2003) Sequence variations in the public human genome data reflect a bottlenecked population history. Proc Natl Acad Sci U S A 100: 376-381.
18. Novembre J, Di Rienzo A (2009) Spatial patterns of variation due to natural selection in humans. Nat Rev Genet 10: 745-755.
19. Pasvol G, Weatherall DJ, Wilson RJ (1978) Cellular mechanism for the protective effect of haemoglobin S against P. falciparum malaria. Nature 274: 701-703.
20. Relethford JH, Blangero J (1990) Detection of differential gene flow from patterns of quantitative variation. Hum Biol 62: 5-25.
21. Thomas W. Yee (2013). VGAM: Vector Generalized Linear and Additive Models. R package version 0.9-3. URL http://CRAN.R-project.org/package=VGAM
22. Tishkoff SA, Verrelli BC (2003) Patterns of human genetic diversity: implications for human evolutionary history and disease. Annu Rev Genomics Hum Genet 4: 293-340.
23. Weir BS, Cockerham CC (1984) Estimating F-Statistics for the Analysis of Population Structure. Evolution 38: pp. 1358-1370.

5. Summary and future prospects

If a section of non-coding DNA is very similar in distantly related species, this pattern is a phylogenetic signal of strong purifying selection – identifying a candidate regulatory region. Such comparisons were used in Chapter 2, and in addition identified some related sequences and subsequences that were less similar, perhaps undergoing a combination of positive selection to adapt to a new role with continuing purifying selection. Phylogenetic comparisons alone provide ambiguous information about the cause of differences in rates of evolution; for example, higher rates of substitution might be due to adaptive evolution in favour of change, or relaxed selection. Additional information was obtained from polymorphism data across multiple individuals within a species – an approach developed in Chapter 3. This within-species data in the form of global allele frequencies at polymorphic sites in human populations indicate strong purifying selection in CNEs, more consistent across the sites within a CNE than selection on coding sequences. Despite reports to the contrary, I found in Chapter 4 no additional information could be gained by examining patterns of differentiation among populations at selected sites; methods for exploring such patterns are not well developed.

An accurate representation of the forces of selection in different regions of the genome can only be achieved with good sequencing coverage across the whole genome. For example, I observed in Chapter 3, that stronger selection in CNEs was only evident with the 1000 Genomes Project low coverage (~ 4X) dataset, a publicly available dataset that is less biased than the HapMap data. It is a more reliable dataset because of its whole-genome approach; it was therefore possible to identify a larger number of rare variants

in non-protein coding sequences. Nonetheless, a number of factors impede accurate identification of polymorphism data at non-protein coding loci. Firstly, obtaining uniform sequence coverage (distribution of reads) across the genome is challenging because the output of next generation sequencing is influenced by the properties of the sequence. For example, sequences with a high GC content (Dohm et al., 2008) and AT-rich repetitive sequences (Harismendy et al., 2009) are prone to low coverage. As a result, promoter/enhancer sequences, consisting of binding site motifs in their sequence that often have such properties, can be especially difficult to sequence. However, improvements in sequencing technologies such as an amplification-free process in the Pacific Biosciences sequencing platform can be used to lower such biases in sequence coverage (Ross et al., 2013); hence in the near future more informative analyses should be possible.

Secondly, platform-specific base-calling errors, inherent in next generation sequencing, result in heterozygotes being miscalled as homozygotes. Additional sources of errors in the 1000 Genomes data such as batch effects, i.e. biases introduced by the time of day, differences arising because of subtle variations in the methods of the technicians handling the samples, and downstream data processing, have been identified (Leek et al., 2010) and might have been exacerbated by employing heterogenous calling procedures across platforms. For instance, merged data from the 1000 Genomes was sequenced using Illumina and SOLiD platforms across four sequencing stations (Broad Institute, Michigan, Boston College and NCBI) each using its own processing pipeline. While this approach gives a high confidence that a variant is a true variant rather than a sequencing error because it has been identified in independent call sets, it is difficult to assess its quality because there is no cross-platform standard for base/variant quality.

Although such errors affect all loci, it may be less pronounced at coding loci because of prior information from a rich resource of validated variants at such sites routinely incorporated in processing pipelines.

I also found that the number of individuals genotyped at coding sites is far higher than at non protein-coding sites in the 1000 Genomes data. Probabilistic frameworks, such as the GATK Unified Genotyper used in the 1000 Genomes Project, incorporate prior information on allele-frequencies based on larger datasets (e.g. dbSNP) to estimate the most likely genotypes and allele-frequencies in a sample (Nielsen et al., 2011). Therefore rare mutations in non protein-coding loci are less likely to be genotyped. Improvements in algorithms, e.g. Cortex (Iqbal et al., 2012), that genotype known variants from low coverage sequence data (too low to be mapped to a reference sequence) can mitigate issues with genomic regions that are prone to low coverage but has the limitation that it cannot be applied to genotype novel variants.

Although pooled allele-frequency data across samples is sufficient for most analyses, determining the genotype of individuals is more informative in population studies and essential in disease-association studies that use linkage disequilibrium information in haplotypes, which require that genotypes are correctly phased (assigned to the chromosome of origin). For example, analytical pipelines that narrow genomic search space in causal variant discovery, such as PriVar (Zhang et al., 2012), predict deleterious scores based on several algorithms using logistic regression and in the absence of pedigree information use HapMap data to estimate haplotype frequencies in regions that are identical by descent (IBD regions). However, the false positive rate of genome-wide statistically significant scores increases as the number of unmatched samples between

the background and target populations increases (Yandell et al., 2011). Such error rates may be further amplified by inaccurate genotype calls that generate haplotype information used in downstream analyses.

Differences in gene expression, considered a (heritable) quantitative trait, are driven by DNA variation (polymorphism) and expression quantitative loci (eQTLs) mapping aims to identify the genomic position (DNA variants) to which that quantitative trait (transcript abundance in this case) is linked (although it does not identify the causal variant(s) in disease associations) (Gilad et al., 2008). Therefore, identifying eQTLs associated with allele-specific expression (differences in transcript abundance between two haplotypes of an individual) will benefit from accurate genotyping and better genome-wide coverage because genotypes are often imputed if not already known.

A number of resources to explore variation in genomes are available to researchers today (surveyed by Pabinger et al., 2013). They include analytical pipelines (e.g. ANNOVAR (Wang et al., 2010)), genome mining tools (e.g. GEMINI (Paila et al., 2013)) and workflow management systems that integrate annotation information from several public databases. There has been a shift from exomes to noncoding DNA in the search for causal variants; conventional focus on transcriptional regulation has also extended to post-translational regulation (e.g. in miRNA) (Bulik-Sullivan, 2013). However, advances in this area are hindered by limited knowledge of functional activity in noncoding regions and most current research is guided by phylogenetic conservation and ENCODE annotations. For example, the disease gene finder VAAST (Yandell et al., 2011) explores the cumulative impact of both coding and noncoding variants but uses primate alignments limited to regulatory regions annotated by ENCODE to predict

deleterious effects in noncoding variation. Hence, variant prioritisation pipelines that focus on greater evolutionary depth and information from deep re-sequencing in putative noncoding regulatory regions in case-control cohorts such as being developed by the Elgar lab are necessary.

Phase 3 of the 1000 Genomes Project now comprises of approximately 2500 individuals, including samples sequenced at a standard coverage of 40X (up to 80X at higher coverage) by the Complete Genomics sequencing platform (Peters et al., 2012). The strategy used by Complete Genomics where homologous chromosomes are physically separated (prior to sequencing by long fragment read (LFR) technology), facilitates diploid genome sequencing of an individual, which automates the phasing procedure at heterozygous loci (albeit made difficult by long runs of homozygosity in non-African populations due to bottleneck effects). The method also allows for genotyping in low coverage sequences using information from neighbouring phased heterozygous SNPs that originate from the same well. Such advances will better catalogue DNA variation in non protein-coding loci including *cis*-acting regulatory changes that affect gene expression in an allele-specific manner by potentially disrupting transcription factor binding sites.

The advent of RNA sequencing technology to quantify genome-wide expression levels at high resolution using read count data as opposed to traditional microarray technology has been key in identifying genes with different transcript usage resulting from allele-specific transcript structure among populations in the 1000 Genomes data (Lappalainen et al., 2013). The authors also identify functional regulatory variation in the human genome by using a combination of RNA sequencing and genome sequencing to

correlate transcript abundance with regulatory variants that are undetected by eQTL analysis due to their low frequencies. Furthermore, Lappalainen et al. (2013) develop an analysis pipeline to identify putative causal regulatory variants from eQTLs. Importantly, their approach does not rely on prior information from SNP arrays; in 81% of cases, the causal variant identified by their analysis pipeline was not present in the widely used Omni 2.5M array. In an example the authors identify a putative causal regulatory variant associated with differences in DGKD gene expression that affect calcium levels in serum (the homeostasis of which is required for healthy teeth and bones). Similar approaches can be applied to other biomarkers of disease phenotypes to aid efforts that guide personalised genomic medicine.

The increasing number of genome sequences from ethnically diverse populations will aid biomarker discovery and validation studies, and impact efforts to minimise the healthcare gap between Europeans and non-Europeans with regard to personalised interventions and therapies. However challenges remain, particularly in genomic data access/sharing across national and international borders and comprehensible workflows to translate human genomic research into clinical applications. The recent formation of the Global Alliance, an international effort aimed at establishing a regulatory framework and best practices for sharing genomic and clinical data securely, and increased investment to address these challenges are indicators of rapid advances in genomic medicine in the near future (e.g. by Silicon Valley venture capitalists). This funding has supported startup companies (e.g. Bina Technologies) developing innovative workflow management systems capable of integrating multiple large datasets. The recent (2013) launch of the company Genomics England by the Department of Health in the UK is evidence of the indispensable role genomic medicine will play in tackling cancer, rare,

and infectious diseases. Its flagship programme The 100K Genome Project aims to sequence 100,000 individuals to aid in research activities that will inform treatment/therapy in an attempt to integrate genomic research into the mainstream healthcare system. The project's Data Working Group (at EBI) in addition to highlighting a need to invest in state-of-the-art infrastructure to meet the project's aims also stresses the imminent shortage of trained bioinformaticians and clinical genomicists with the capacity to translate research findings where they are most needed (this is unsurprising because a researcher/clinician is faced with an overwhelming array of choices in sequencing platforms and analytical methodologies that are currently not standardised). Comprehensive partnerships, similar to those that exist within the biotechnology industry (e.g. Intercrossing, http://intercrossing.wikispaces.com/), between higher education institutions and clinics should be established to help meet the forecasted demand and expedite translational genomic research in future.

5.1. References

1. Bulik-Sullivan B, Selitsky S, Sethupathy P (2013) Prioritization of genetic variants in the microRNA regulome as functional candidates in genome-wide association studies. Hum Mutat 34: 1049-1056.
2. Dohm JC, Lottaz C, Borodina T, Himmelbauer H (2008) Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. Nucleic Acids Res 36: e105.
3. Gilad Y, Rifkin SA, Pritchard JK (2008) Revealing the architecture of gene regulation: the promise of eQTL studies. Trends Genet 24: 408-415.
4. Harismendy O, Ng PC, Strausberg RL, Wang X, Stockwell TB, et al. (2009) Evaluation of next generation sequencing platforms for population targeted sequencing studies. Genome Biol 10: R32.
5. Iqbal Z, Caccamo M, Turner I, Flicek P, McVean G (2012) De novo assembly and genotyping of variants using colored de Bruijn graphs. Nat Genet 44: 226-232.
6. Lappalainen T, Sammeth M, Friedl‰nder MR, 't Hoen PA, Monlong J, et al. (2013) Transcriptome and genome sequencing uncovers functional variation in humans. Nature 501: 506-511.
7. Leek JT, Scharpf RB, Bravo HC, Simcha D, Langmead B, et al. (2010) Tackling the widespread and critical impact of batch effects in high-throughput data. Nat Rev Genet 11: 733-739.
8. Nielsen R, Paul JS, Albrechtsen A, Song YS (2011) Genotype and SNP calling from next-generation sequencing data. Nat Rev Genet 12: 443-451.
9. Pabinger S, Dander A, Fischer M, Snajder R, Sperk M, et al. (2013) A survey of tools for variant analysis of next-generation genome sequencing data. Brief Bioinform.
10. Paila U, Chapman BA, Kirchner R, Quinlan AR (2013) GEMINI: integrative exploration of genetic variation and genome annotations. PLoS Comput Biol 9: e1003153.
11. Peters BA, Kermani BG, Sparks AB, Alferov O, Hong P, et al. (2012) Accurate whole-genome sequencing and haplotyping from 10 to 20 human cells. Nature 487: 190-195.
12. Ross MG, Russ C, Costello M, Hollinger A, Lennon NJ, et al. (2013) Characterizing and measuring bias in sequence data. Genome Biol 14: R51.
13. Wang K, Li M, Hakonarson H (2010) ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res 38: e164.
14. Yandell M, Huff C, Hu H, Singleton M, Moore B, et al. (2011) A probabilistic disease-gene finder for personal genomes. Genome Res 21: 1529-1542.
15. Zhang L, Zhang J, Yang J, Ying D, Lau YL, et al. (2013) PriVar: a toolkit for prioritizing SNVs and indels from next-generation sequencing data. Bioinformatics 29: 124-125.

6. Appendices

6.1. Subset of CNEs used in variant analyses

Saved in text file. Coordinates are from the Hg19 GRCh37 assembly.

6.2. CNEs in larger dataset used in phylogenetic analyses

Saved in text file. Coordinates are from the Hg19 GRCh37 assembly.

6.3. Perl Scripts used in analyses

6.3.1.Extract branch length from Newick trees

```
#!/usr/bin/perl -w

## Script to parse Newick trees from RAxML output and calculate
## branch length - Written with help from Pavlos Pavlidis, Heidelberg
## Institute for Theoretical Studies, Germany

use Bio::TreeIO;
use IO::String;
use strict;
use Bio::Tree::TreeFunctionsI;
use Bio::Tree::TreeI;


# Extract Branch Lengths for the following:
my @pairwise = qw(
        human/macaque
        human/mouse
        human/chicken
        human/frog
        human/zfish
        human/fugu
        macaque/mouse
        macaque/chicken
        macaque/frog
        macaque/zfish
        macaque/fugu
        mouse/chicken
        mouse/frog
        mouse/zfish
        mouse/fugu
        chicken/frog
        chicken/zfish
        chicken/fugu
        frog/zfish
        frog/fugu
        fugu/zfish);

my $filepath = $ARGV[0];
my $list = $ARGV[1]; #file containing list of filenames (one per line) for the newick trees to be used
my $filename= $ARGV[2]; #output filename

if($ARGV[0] eq "-h"){print "filepath file_w_list_of_filenames output_filename\n";}else{
        open OUT, "> $filename.txt";
        print OUT "Partition\t";
        foreach my $p (@pairwise){
           print OUT "$p\t";
        }
        print OUT "\n";

        # open the files - modify for the diff categories.
        open FILES, "< $list";
        my @myfiles = <FILES>;
        chomp @myfiles;
        close FILES;

        # Variable to hold the species that will become the root node
        my $root_species = "fugu";
```

```
my @groups = qw(50,100,150,200,250,300,350,400,450,500);

foreach my $group (@groups){
  my $i =0;

  while($i<100000){
        $i++;
        open IN, "< <Folder>$group/file$group_\$i."; # Insert the folder names and filename
        # parse in newick/new hampshire format
        my $input = new Bio::TreeIO(-fh   => \*IN,
                -format => "newick");
        my $tree = $input->next_tree;

        # root the tree on fugu
        $tree->set_root_node($root_species);

        my @leaves = sort($tree->get_leaf_nodes); #Returns the leaves (tips) of the tree
        my @nodes = sort($tree->get_nodes); # Returns array of Tree::Node objects
        my $size = $tree->number_nodes; #Returns the number of nodes


        for my $nd (@nodes) # foreach node in the tree (both leaf nodes and internal un-named nodes)
        {
                my @desc = $nd->get_all_Descendents;
                my @comp_name = (); # store names of all the leaf nodes that will describe the un-named
parent (internal) node

                my $name = "";

                foreach my $d (@desc)
                {
                        if( $d->is_Leaf )
                        {
                                push @comp_name, $d->id; # add the leaf node id to the array
                        }

                }

                $name = join(".", sort(@comp_name) ); # give the un-named internal node the name of it's
descendent leaves

                if(!$nd->is_Leaf)
                {
                        $nd->id( $name ); # Assign the name of all the descendents of that node as the
node id;
                }


                for( my $i=0; $i<@nodes; ++$i){
                        my $node = $nodes[$i];
                        print LOG "$i\t$node->id\n";
                }
        }

        my %IDHash = ();

        for(my $i=0; $i<@nodes; ++$i){
                if( length($nodes[$i]->id) > 0){ # if the nodes have been assigned names
                        $IDHash{$nodes[$i]->id} = $i;
                }
        }
```

```
                    foreach my $key (keys %IDHash)
                    {
                            print LOG "$key\t$IDHash{$key}\n";
                    }

                    #Get the distance between two nodes by adding up the branch lengths of all the connecting edges
between two nodes.
                    foreach my $pair (@pairwise){

                            my @species =();

                            @species =split("/", $pair);

                            my $first = $species[0];
                            my $second = $species[1];

                            my $distances = $tree->distance(-nodes => [$nodes[$IDHash{$first}],
$nodes[$IDHash{$second}]]);

                            print OUT "$distances\t";

                    }
                    print OUT "\n";
                    $partition++;
            }

            close OUT;
        }
```

6.3.2.Define NVRs and RVRs in CNEs

```
#!/usr/bin/perl -w

## Script to define nonvariable and restricted variable regions in CNEs using
## an alignment of seven species

use strict;

my(%filter,%sequences) =();

#filter 1809 CNEs

open FIL, "< 1809CNEs.txt";
while(<FIL>){
        chomp;
        $filter{$_} =1;
}

close FIL;

my($filename,$cne, $line, $state)="";
my($cnestart,$snploc, $pos, $position) =0;
my($hbase,$macbase,$mousebase, $chbase, $frogbase,$zbase, $fbase) = "";
my($humanseq, $macseq, $mouseseq, $chseq, $frogseq, $zseq, $fseq) =  "";

open OUT, "> 1KG.NVR.RVR.snps.txt";


open AF, "< 1KG.AF.GeneCluster.DAF.coordinates.txt";
while(<AF>){
        chomp;
        $line =$_;
        @_=split("\t",$_);
        $cne=$_[8];
        if($filter{$cne}){
          $cnestart=$snploc=0;
          $hbase=$macbase=$mousebase= $chbase= $frogbase=$zbase= $fbase= "";
          $humanseq= $macseq= $mouseseq= $chseq= $frogseq= $zseq= $fseq = "";
          %sequences=();
          $filename="RAxMLinputseq/$cne.aln";
          $cnestart = $_[12]-1;
          $snploc = $_[1];
          $state ="";
          $position = 0;
          $pos=0;
          open ALN, "< $filename";

          while(<ALN>){
                chomp;
                if(!/^CLUSTAL/){
                        if(/human/){
                                $humanseq .= substr($_, 16);
                        }
                        if(/macaque/){
                                $macseq .= substr($_, 16);
                        }
                        if(/mouse/){
                                $mouseseq .= substr($_, 16);
                        }
                        if(/chicken/){
                                $chseq .= substr($_, 16);
                        }
```

```
                        if(/frog/){
                                $frogseq .= substr($_, 16);
                        }
                        if(/zfish/){
                                $zseq .= substr($_, 16);
                        }
                        if(/fugu/){
                                $fseq .= substr($_, 16);
                        }
                }
                if(eof){
                        %sequences = ("human", $humanseq,
                                "macaque",$macseq,
                                "mouse",$mouseseq,
                                "chicken",$chseq,
                                "frog", $frogseq,
                                "zfish", $zseq,
                                "fugu", $fseq);

                        while($humanseq =~ /(A|T|G|C)/g){

                                $cnestart++;

                                if($cnestart == $snploc){
                                        $pos=pos($humanseq);

                                }
                        }
                        $position = $pos-1;
                        #get the base at the snp pos for each species in the alignment

                        $hbase = substr($sequences{human},$position,1);
                        $macbase = substr($sequences{macaque},$position,1);
                        $mousebase = substr($sequences{mouse},$position,1);
                        $chbase = substr($sequences{chicken},$position,1);
                        $frogbase = substr($sequences{frog},$position,1);
                        $zbase = substr($sequences{zfish},$position,1);
                        $fbase = substr($sequences{fugu},$position,1);

                        if(($hbase eq $macbase) && ($hbase eq $mousebase) && ($hbase eq $chbase) &&
($hbase eq $frogbase) &
                                & ($hbase eq $zbase) && ($hbase eq $fbase)){

                        $state = "NVR";

                                }else{
                                        $state = "RVR";
                                }

                                print OUT
"$line\t$hbase\t$macbase\t$mousebase\t$chbase\t$frogbase\t$zbase\t$fbase\t$state\t$positi
                                on\n";

                }
            }
            close ALN;
          }
}
close AF;
close OUT;
```

6.3.3.Extract consequences of coding variants using Biomart

```
#!/usr/bin/perl -w

## Script to retrieve synonymous/nonsynonymous state of coding variants from
## Ensembl biomart

use strict;
my @chrom = (1..22,"X");
foreach my $chr (@chrom){

        open F, "< LongestTrans_FullCDS_Chr/LongestTrans_FullCDS_Chr$chr.txt";

        # transcript IDs as input to get consequence of variants
        ##multiple chromosomal regions with exon start-stop to retrieve variant effects

        my $i=0;
        my $ids ="";
        my @array =();
        my %trans =();

        while(<F>){
          chomp;
          @_=split("\t",$_);
          $trans{$_[1]} =1;
        }close F;

        for my $key (keys %trans){
          $i++;

          $ids .= "$key,";

      # retrieve 100 IDs at a time

          if($i == 100){
                  chop $ids;
                  push(@array, $ids);
                  $i=0; $ids="";
          }
        }

        push(@array, $ids);

        my $j=0;

        foreach my $run (@array){

          my  $query = '<?xml version="1.0" encoding="UTF-8"?><!DOCTYPE Query><Query  virtualSchemaName =
"default" formatter = "TSV" header = "0" uniqueRows = "0" count = "" datasetConfigVersion = "0.6" ><Dataset name =
"hsapiens_gene_ensembl" interface = "default" ><Filter name = "with_validated_snp" excluded = "0"/><Filter name =
"ensembl_transcript_id" value = "'."$run".'"/><Filter name = "so_parent_name" value =
"frameshift_variant,missense_variant,synonymous_variant"/><Attribute name = "ensembl_transcript_id" /><Attribute
name = "external_id" /><Attribute name = "chromosome_name" /><Attribute name = "chromosome_location"
/><Attribute name = "synonymous_status" /></Dataset></Query>'."";

          $j++;

          system("wget -O BiomartConsequences/Trans_Variants_Chr$chr\_$j.txt
'http://www.ensembl.org/biomart/martservice?query=$query'");

          sleep 3;
        }
}
```

6.3.4.Subset 1000 Genome variants by population

```
#!/usr/bin/perl -w

## Script to parse variants from the 1000 Genomes vcf file and separate them by population with recalculated allele frequencies.

use strict;
my @folders= qw(CNENVR CNERVR NONSYN SYN Noncoding);
my @pops = qw(YRI CHB CEU CHS JPT CLM MXL FIN GBR IBS LWK PUR TSI ASW);
my @chrom = (1..22,"X");
my $path = "../../Interim_phase1_lowcov/";

foreach my $chr (@chrom){
        my $datafile= "$path"."ALL.wgs.project_consensus_vqsr2b.20101123.snps.low_coverage.sites.vcf.gz" ;

        foreach my $folder (@folders){
          my $outfolder = "GENO$folder";
          my $filename = "$path"."$folder/"."$folder.Chr$chr.bed";
          if(-e $filename){

                  my $outvcf = "$path$outfolder/$folder.genotypes.Chr$chr.vcf" ;
                  open F, "< $filename";

                  #use tabix to retrieve variants
                  #print header only
                  system("../../tabix-0.2.6/./tabix -fh $datafile $chr:1-100 > $outvcf" );
                  #retrieve the rest of the positions
                  while(<F>){
                          chomp;
                          s/chr//;
                          @_=split("\t",$_);
                          my $pos = "$_[0]:$_[1]-$_[2]";
                          system("../../tabix-0.2.6/./tabix -f $datafile $pos >> $outvcf " );
                  }
                  close F;

                  # subset the outvcf files into the different populations here
                  foreach my $pop (@pops){
                          my $c = "$path$pop.panel.txt"; # files containing list of individual IDs per population
                          my $subpath = "$path"."SUBSET/"."$pop";
                          my $infile = $outvcf;
                          my $outfile = "$pop.Chr$chr.genotypes.vcf";

                          print "Working on\n ./vcf-subset -e -f -c $c $infile | ./fill-an-ac > $subpath/$outfile\n" ;
                          system("./vcf-subset -e -f -c $c $infile | ./fill-an-ac > $subpath/$outfile");
                          print "file $outfile done....";
                          sleep 1;

                  }
          }
        }
}
```

6.3.5.Create dataframe of genotypes for F_{ST} calculations (Theta)

```
#!/usr/bin/perl -w

## Script to obtain genotype information from the 1000 Genomes variant data in the different
## categories to calculate Weir and Cockerham's statistic Theta in R

use strict;

#store SNPs in hash with Type as value
my %snps=();

open IN, "< Dilrini_1000Genomes_Lowcoverage_best.txt";
while(<IN>){
        if(/^\d/){

          @_=split("\t",$_);
          my $type = $_[1];
          my $loc = $_[0];
          $loc =~ s/\_/\t/;

          $snps{$loc} = $type;
        }
}

close IN;

my @type=qw(CNE_Nonvariable CNE_Restricted NONSYN2 SYN2 Noncoding2);
my @pops = qw(ASW CHB CEU CHS JPT CLM MXL FIN GBR IBS LWK PUR TSI YRI);
my $path = "../Interim_phase1_lowcov/POPSUBSET";
my @chrom = (1..22, "X");

#ASW.Restricted.Chr21.genotypes.vcf (example file name)

foreach my $group (@type){
        print "$group\n";
        open OUT, "> FST_input_$group.txt";
        print OUT "Pop\tType\tLocus\tNaa\tNab\tNbb\n";

        my $region = $group;
        $region =~ s/2//;

        foreach my $pop (@pops){
          foreach my $chr (@chrom){
                my $filename = "$path/$group/$pop/$pop.$region.Chr$chr.genotypes.vcf";

                if(-e $filename){
                        open VCF, "< $filename";
                        print "working on $filename\n";

                        while(<VCF>){
                                if(/^\d/){
                                        @_=split("\t",$_);
                                        my $id = "$_[0]\t$_[1]";
                                        if($snps{$id}){

                                                # extract the number of homozygotes and heterozygotes in
the sample

                                                my $aa =0;
                                                my $ab =0;
                                                my $bb =0;
```

```
                                                while(/\b0\|0\b/g){

                                                        $aa++;
                                                }
                                                while(/\b0\|1\b/g){

                                                        $ab++;
                                                }
                                                while(/\b1\|0\b/g){
                                                        $ab++;
                                                }
                                                while(/\b1\|1\b/g){
                                                        $bb++;
                                                }
                                                my $pos = $id;
                                                $pos =~ s/\t/\_/;

                                                print OUT "$pop\t$snps{$id}\t$pos\t$aa\t$ab\t$bb\n" ;
                                        }
                                                }
                                        }
                                        close VCF;
                                }
                          }
                        }
                        close OUT;
}
```

6.3.6.Create dataframe for F_{ST} calculations (statistical models)

```
#!/usr/bin/perl -w

## Script to create a dataframe with the derived allele count and allele number
## in each population from the 1000 Genomes project for all the variants used in the
## analysis

use strict;

my @type=qw(CNE_Nonvariable CNE_Restricted NONSYN SYN Noncoding);
my @pops = qw(ASW CHB CEU CHS JPT CLM MXL FIN GBR IBS LWK PUR TSI YRI);
my $path = "POPSUBSET";

open OUT, "> dataframe_1000GenomesLowcov.txt";

print OUT "SNP\tType\t";
my @hashesPop=();
my $num = -1;
while($num < 13){
        $num++;
        push @hashesPop, {};
}
$num=-1;

foreach my $name (@pops){
        print OUT "$name"."_DAC"."\t$name"."_ChrNo";
        if($name eq "YRI"){
          print OUT "\n";
   }else{print OUT "\t";}

   #get the derived allele from the different categories
   foreach my $cat (@type){
            print "processing $cat\n";
            my $file = "$cat"."_DerivedAllele.txt";
            open F, "< $file"; # derived alleles from the global file
            my %derived = ();
            while(<F>){
                    if(/^\d/){
                            chomp;
                            @_=split("\t",$_);
                            my $chr = $_[0];
                            my $snp= "$_[0]\_$_[1]";
                            my $ref = $_[2];
                            my $alt = $_[3];
                            my $da = $_[4];
                            my $dac = 0;

                            my $change =0; # derived allele is alternate allele by default;
                            if($da eq $ref){
                                    $change =1;
                            }
                            if($da ne $ref && $da ne $alt){
                                    $da ="NA";
                            }
                            if($da ne "NA"){

                                    print OUT "$snp\t$cat\t";

                                    # get the allele counts from the different populations
                                    # not all snps are in all populations, to deal with missing data
                                    # check against a corresponding file of available SNPs within
                                    # each population. Variants that are not polymorphic in a population
```

```
                                        # are included with AC=0

                                        $num =-1;
                                        foreach my $pop (@pops){
                                                $num++;
                                                my $folder = "$path/"."$cat/"."$pop/";
                                                my $filename = "$pop."."$cat."."Chr$chr".".genotypes.vcf";

                                                if(-e $folder.$filename){
                                                        my $infile = "$folder"."$filename";

                                                        open FILE, "< $infile";

                                                        while(<FILE>){
                                                                chomp;
                                                                if(/^\d/){
                                                                        my @line =split("\t",$_);
                                                                        my $id = "$line[0]\_$line[1]";
                                                                        if($snp eq $id){
                                                                                #e.g. AC=26;AN=122
                                                                                my @info = split(";",
$line[7]);

                                                                                $info[0] =~ s/AC\=//;
                                                                                $info[1] =~ s/AN\=//;
                                                                                my $altac = $info[0];
                                                                                my $an = $info[1];

                                                                                # i.e. derived allele
                                                                                # is the reference
                                                                                if($change == 1){
                                                                                        $dac = $an-$altac;
                                                                                }else{
                                                                                        $dac = $altac;
                                                                                }

                                                                                if($dac ne "NA"){
                                                                                        print OUT
"$dac\t$an";

                                                                                        if($pop ne "YRI"){
                                                                                                print
OUT "\t";
                                                }else{ print OUT "\n";}
                                                                        }
                                                                }
                                                        }
                                                }
                                        }
                                }
                        }
                }
        }
        close F;
}

close OUT;
sleep 2;
```

6.4. Alignments of 7 fast-evolving CNEs in Chapter 2: Table 1

6.4.1.Across seven vertebrate species

```
CRCNE00003541_CLUSTAL 2.1 multiple sequence alignment


mouse           -GCTGTCTACGTGGCATTCTTACATGATTTCATGCTTTGCCAATGGAAAATTAGACCCTT
frog            ------------------TTACATGATTTCATGCTTTGCTGATGGAAAATTAGACCCCT
human           -GCTGGTTATATGGCATTCTTACATGATTTCGTGCTCTGCTGATGGAAAATTAGACCCTT
chicken         -GCTGGTTATGTGGCATTCTTACATGATTTCATGCTTTGCTGATGGAAAATTAGACCCTT
macaque         -GCTGGTTATGTGGCATTCTTACATGATTTCATGCTTTGCTGATGGAAAATTAGACCCTT
fugu            CGCTTGTTGTGTGGCATTCTCTCTTGATTTCACGCTTTGCTGATGGAAAATTAGACCCTG
zfish           --CTTGTTGTGTGGCATTCTCTGCCGATTTCATGCTTTGCTGATGGAAAATTAGGCCCCT
                             *     ******  *** ***  ************ ***

mouse           GTGTATTCTAAAAGAATAGAAG-
frog            GTGTATTAAAAAA----------
human           GTGTATTATAAAAGCATAGAAG-
chicken         GTGTATTATAAAAGAATAGAAG-
macaque         GTGTATTATAAAAGAACAGAAG-
fugu            GTGTATCTCTGCATGAAGAGGAG
zfish           GCGTATC----------------
                * ****

CRCNE00001579_CLUSTAL 2.1 multiple sequence alignment


human           AAAAGCTGACCTTGTTTTATTTCCTGGGC-GGGTGTGGTATTGTCCCGTCAGAACCCCGA
macaque         AAAAGCTGACCTTGTTTTATTTCCTGGGC-GGGTGCAGGATTGTCCCATCAGAACCCCGA
mouse           AAAAGCTGACCTTGTTATATTTCCTGGGC-GGGTGTAGGATTGTCCCATCAGAGGCCCAA
chicken         AAATGCTGACCTTGTTATATTTACTGGGT-GGGTGCAGAATTGTTCCATCACTACCCCGA
frog            ----GCTGACCTTGTCATATTTAGAGGGT-GGGTGCAGAATTGTTCCATCACTATCCTGG
fugu            AAACGCTGACCTTGTCATATTTACAGGGT-GGCTCCAGAATTGTTCCATCAGTAGCCAGA
zfish           AAATGCTGACCTTGTCATATTTACGGGGCCGGCTGCAGAATTGTTCCATCAGTATCCAGA
                    ***********  *****   ***  ** *   * ***** ** ***    **

human           GGGCTGAACGATGGAGCTTGTCAAAGAGAGGCCCCATGTAATGAATCATCTTATCACAGG
macaque         GGGCTGAACGATGGAGCTTGTCAAAGAGAGGCCCCATGTAATGAATCATCTTATCACAGG
mouse           GGGCTGAACGATGGAGCTTGTCAAAGAGAGGCCTCGTGTAATGAATCATCTTATCACAGG
chicken         GGGTTGAACGATGGAGCTTGTCAAAGTGAGGCCTTATCTAATGAATCATCTTATCGCAGG
frog            GTGTTGAACGATGGAGCTTGTCAAAGTGAGACCATATCTAATGAATCATCTTATCGCAGG
fugu            GGCTTGAACGATGATACTTGTCATTGTGGGACTCTATCTAATGAATCATCTTATCGCAGG
zfish           GGCTTGAACGATGAGGCTTGTCAGACCAGAGCCCTATCTAATGAATCATCTTATCGCAGG
                *   *********   *******        *    * ***************** ****

human           TGCACACCGCAC--ATATTAAAGG-AGGGTTGTATGAACACTGTGCCCT
macaque         TGCACACCACAC--ATATTAAAGG-AGGGTTGTATGAACACTGTGCCCT
mouse           TGCACACCACAC--ATATTAAAGG-AGGGTTGTATGAACACCGTGCCCT
chicken         TGCACACCACAC--ATATTAAAGG-AGGATTGTATGAACAATGTGCCCT
frog            TGCACACCACAT--GTATTAAAAG-AGGATTGTATGAACAATGTGCCCT
fugu            TGCACACAGCTC--ATATTAGAGGGAAAAT-ATATGGACAACGTGCCCT
zfish           TGCGCACCGCTCCCATATTAAAGGAAAGATTGTATAGCCAATGTGCCCT
                *** ***  *     ***** * * *   *  ***   **  *******


CRCNE00005971_CLUSTAL 2.1 multiple sequence alignment


mouse           ---------------------ATATATCTTCATGACAT-GCTTTAAATTATTTATAACTT
chicken         -----ATCTGATGAATTATTAATACATCTTCATGACAT-GCTTTAAATTATTTATAACTT
macaque         -----ATCTGATGAATTATTAATACATCTTCATGACAT-CCTTTAAATTATTTATAACTT
human           TATTTATCTGATGAATTATTAATACATCTTCACGACAT-CCTTTAAATTATTTATAAGTT
fugu            TATTTATCTATTGAATTATTAATTCTTCTTCATGACTTTCCCTTTCATTATTCATGACTT
zfish           ---TAATCCGATGAGTTATTTATGGCGCATCNNNNNNN-NNNNNNNNNNNNNNNNNNNNN
                                  **    * **

mouse           AATTTTGAGTTTATTGCTTTTAGACGCTGCATTGTT
chicken         AATTTTGAGTTTATTGCTTTTCGACGCTGCATTGTT
macaque         AATTTTGAGTTTATTGCTTTTAGACGCTGCATTGTT
human           AATTTTGAGTTTATTGCTTTTAGACGCTGCATTGTT
fugu            AATTT-GAGATTATTGCAT----------CATTGTT
zfish           NNNNNNNNNNNNNNNNNNNNNNNNCCGTCCTCATG
                                       * *  *
```

```
CRCNE00008385_CLUSTAL 2.1 multiple sequence alignment


human           AGTAGGCATAAGACGCTTGAAGTAATTACTTAAGGACTCCCTTTCAAACAATC----GCT
macaque         AGTAGGCATAAGACACTTGAAGTAATTACTTAAGGACTCCCTTTCAAAGAATC----TCT
chicken         AGTAGGCCTAAGACACTTGAAGTAATTACTTAAAGGCTTCCTTTCAAAGAATC----TCC
mouse           AGCAGGCCTAAGATACTTGAAGCAATTACTTAAGGGCTCCCTTTCAAAGAATC----TCT
zfish           -----------GACACTTGGAGTAATTACTTAAAG-CTTTCCCTCAGATCTC-----TCC
fugu            -CAGGAAGTGCGAGACCAGGAGTAATTACTTAAGG-CTTTCTGTCACATCTCCCAACTCT
                           **  *  * ** ********** * **  *  *** *          *

human           CCCTTTTTTTAT----TAAATGCCTTTTAACATTAGCACTTTCTTCCTGACAAATTAAAT
macaque         CCCTTTTTT-AT----TAAATGCCTTTTAACATTAGCACTTTCTTCCTGACAAATTAAAT
chicken         CCCTTTTT--AT----TAAATGCCTTTTAACATTAGCACTTTCTTCCTGACAAATTAAAT
mouse           TCCCTTTTTTAT----TAAATGCCTTTTAACATTAGCACTTTCTTCCTGACAAATTAAAT
zfish           CGACTTTTTTATTATCCAAAAGCTTTTTAACATTAGCACTGTCTTCCTGACAAATTAAAT
fugu            CTTTTTTTTTATAATCCAAAAGCTTTTTAACATTAGCACTCTCTTTCTGACAAATTAAAT
                    ****  **     *** ** ********************* ********************

human           GAAATTTCATTACTCTCCCTGACTTAATGGCATTAGGCT
macaque         GAAATTTCATTACTCTCCCTGACTTAATGGCATTAGG--
chicken         GAAATTTCATTACTCTCCCTGACTTAATGGCATGACGCT
mouse           GAAATTTCATTACTCTCCCTGACTTAATGGCATGAGACT
zfish           GAAATTTCATTACTCCAGCCATCCTAATGGCATGAAACT
fugu            GAAATTTCATTACTCCGGCCGCCCTAATGGCATGAAACT
                ***************   *   * ********* *

CRCNE00005870_CLUSTAL 2.1 multiple sequence alignment


macaque         CTGAAAAGGCTTGGGTTGATTAATATTTTCCGCTGTCCGGTAACCATGGCAATAGGGTCA
mouse           CTGAAAAGACTTAGGTTGATTAATATTTTCCGCTGTCCGGTAACCATGGCAATAGGGTCA
human           CTGAAAAGGCTCGGGTTGATGAATATTTTCCGCTGTCCGGTAACCATGGCAATAGGGTCA
chicken         CTGTAAAGGCTTGGCTTGATTAATATTTTCCGCTGTCCGGTAACCATGGCAATGGGGTCA
zfish           -----AAGTGCAGACTTGATTAATATTTTCTTCTGTTTGGTAACCATGGCAATGAGGTCA
fugu            CTTCCAAGTGCAGGCTTGATTAATATTTTCCACTGTTCCGTAACCATGGCAATAAGGTCA
                     ***       ***** *********  ****   **************  *****

macaque         GCAGATATAATACTTAATTTAATCTTCCAGCTGCCTCAGTGTGCTTCACCAGTGTCCCGA
mouse           GCAGATATAATATTTAATTTAATCTTCCAGCTGCCTCAGTGAGCTTCAC-----------
human           GCAGATATGATACTTAATTTAATCTTCCAGCTGCCTCAGTGTGCTTCACCAGTGCCC---
chicken         GCAGATATAATACTTAATTTAATCTTCCAACTGCCTCAGTGTGCATCACCAGTGTCC---
zfish           GTGGATATGATGCTTAATTGAATCTTTCCCGAGCTTTTGTGTTCACCTCGTGTGTCCAAA
fugu            CTGCATATGATACTTAATTGAATCTTCCCAGTGCTTTTGTGAGCATCTCCTGTGTCCGGA
                    **** **  ****** ****** *    ** *  ***  *  * *

macaque         ATC------
mouse           ---------
human           ---------
chicken         ---------
zfish           AACTACAAA
fugu            CAATAAAAA


CRCNE00008599_CLUSTAL 2.1 multiple sequence alignment


human           AGGGCTCACATTGTGCTTTTCTAGTGGCATTGATTTGTTAAAGCGCTGAATCAGCATCTC
macaque         AGGGCTCGCATTGTGCTTTTCAAGTGGCATTGATTTGTTAAAGCGCTGAATTAGCATCTC
chicken         AGGGCTCGCATTGTACTTTTCAAGTGGCATTGATTTGTTAAAGCGCTGAATTAGCATCTC
mouse           AGGGTTCTCATTGTACTTTTCAAGTGGCATTGATTTGTTAAAGCGCTGAATTAGCATCTC
zfish           ------------GTACTTGTCAAGTGGCATTGATTTGTTAAAGTGCTGAATTAGCAC--A
fugu            AGGGTGCAGATTGTACTCTTCAAGTGGCATTGATTTGTTAAAGTGCTGAATTAGCAC--A
                          ** **  ** ******************** ******* ****

human           AGGCCAGCTGC
macaque         AGGCCAGCTGC
chicken         AGGCCAGCTGC
mouse           CGGCCAGCTGC
zfish           -----------
fugu            AGTCCAGCTGC
```

```
CRCNE00003046_CLUSTAL 2.1 multiple sequence alignment

human           GGGCCCGAGCGCCGAGGGTCCCATTTCGCACACAGCATGGAGTGGGTGAGCTCCTCAACT
macaque         GGACCCGAGTGCCGAGGGTCCCATTTCTCACACAGCATGGAGTGGGTGAGCCCCTCAACT
mouse           GGACCCGAGTGACGAGGGTCCCATTTCCAGCACAGCATGGAATGGGTGATCTCCTCAACT
chicken         ------GAGTGACGAGGGTCCCATTTCCAACACAGCATGGAGTGGGTGATCTTGTCAACT
zfish           --------GTAACTGGGGTCCTATTTCAGCCGCAGCGTGAAGCCACTGAGTTCCTCAGCT
fugu            GGGCCACTGTGACTGGGGTCCCATTTCAGGCCCAGCGTGGAGCCGGAGAGCTCCTCAGCT
                        *   *  ****** *****   * **** ** *      **     *** **

human           GTAAAATGAAGCCGTCAGAGCACAGCGTGAGAATGTATTAGAGACAGTTACACGAGG---
macaque         GTAAAATGAAGCCGTCAGAGCACAGCATGAAAATGTATTAGAGACAGTTACACGAGG---
mouse           GTAAAATGAAGCCGTAAGAGCTCAGCATGAAAATGTATTAGCAACAGTTAGATAAGGG-A
chicken         GTAAAATGAAGCCATAAGAGCACAGCATGAAAATGTATTAGCAACAGTTAGATAAGGA-A
zfish           GTAAAATGAAAGCAGCGGAGCACAGTGAGAAAATGTATTAGCCTCGCTGTGATAAGGAGA
fugu            GTAAAATGAGAGCGTCGGAGCGCAGTGTGAAAATGGATTAGCCTCGCTGAGATAAGGGGA
                *********   *    **** ***   ** **** *****   *  *   *  ***

human           GGCCCCAGCACGGCGCTCAC-ACTTTCAATTGGT-AAAATAATGTCTGTACAAATTGTTT
macaque         GGCCCCAGCACGGCGCTCAC-ACTTTCAATTGGT-AAAATAATGTCTGTACAAATTGTTT
mouse           GGCCTCAGCAAGGCGTTCAC-ACTAGCAATTGGT-AAAATAATGTCTGTACAAATTGTTT
chicken         GACCTCAGCAAGGCGTACAC-ACTTTCAATTGGT-GAAATAATGCTTGTACAAATTGTTT
zfish           GCCCTTAGCGAGTGACGCAC-GCCTTCAGTGGGG-AAAATAACCACTGGACAAATTGTTT
fugu            GCCCTCAGCGAGCCATGCACAGCCCTGAATGGGGGAAAAGAACTGCTGTACAAATTGTTT
                * **  ***  *     ***  *    * * **   *** **    ** ***********

human           TTGGTACATTAAGTACCTTTCACAGCCAAAATTAATAAAAGAAGAA
macaque         TTGGTACATTAAGTACCTTTCACAGCCAAAATTAAAAAAAGAAGAA
mouse           TTGGTACATTAAGTACCTTTCACAGCCAAAATTAAAAAAAAAAAAA
chicken         TTGGTACATTAAGTACCTTTCACAACCAAAATTAAAAAAAGAAGAA
zfish           TTGCTACATTAAGTACCTGTTACTTT---ATTTCACAGATGGAAAA
fugu            TTGGTACATTAAGTACCTGTTACTTC---ATTTCACAGATGGAAAA
                *** ************** * **      * ** * * *   * **
```

6.4.2.Extended alignments including all vertebrate species available

CRCNE00003541

```
frog            -------------------TTACATGATTTCATGCTTTGCTGATGGAAAATTAGACCCCT 41
macaque         -GCTGGTTATGTGGCATTCTTACATGATTTCATGCTTTGCTGATGGAAAATTAGACCCTT 59
rat             -GCTGTCTACGTGGCATTCTTACATGATTTCATGCTTTGTCGATGGAAAATTAGACCCTT 59
mouse           -GCTGTCTACGTGGCATTCTTACATGATTTCATGCTTTGCCAATGGAAAATTAGACCCTT 59
opossum         -------TATGTGGCATTCTTACATGATTTCATGCTTTGCTGATGGAAAATTAGACCCTT 53
bushbaby        -GCTGGTTATGTGGCATTCTTACATGATTTCATGCTTTGCTGATGGAAAATTAGACCCTT 59
dog             -GCTGGTTATGTGGCATTCTTACATGATTTCATGCTTTGCTGATGGAAAATTAGACCCTT 59
armadillo       -GCTGGTTATGTGGCATTCTTACATGATTTCATGCTTTGCTGATGGAAAATTAGACCCTT 59
rabbit          -GCTGGTTATGTGGCATTCTTACATGATTTCATGCTTTGCTGATGGAAAATTAGACCCTT 59
horse           -GCTGGTTATGTGGCATTCTTACATGATTTCATGCTTTGCTGATGGAAAATTAGACCCTT 59
chicken         -GCTGGTTATGTGGCATTCTTACATGATTTCATGCTTTGCTGATGGAAAATTAGACCCTT 59
squirrel        -GCTGGTTATGTGGCATTCTTACATGATTTCATGCTTTGCTGATGGAAAATTAGACCCTT 59
cow             -GCTGGTTATGTGGCATTCTTACATGATTTCATGCTTTGCTGATGGAAAATTAGACCCTT 59
human           -GCTGGTTATATGGCATTCTTACATGATTTCGTGCTCTGCTGATGGAAAATTAGACCCTT 59
chimp           -GCTGGTTATATGGCATTCTTACAGGATTTCATGCTTTGCTGATGGAAAATTAGACCCTT 59
bat             -GCTGGTTATGTGGCATTCTTACATGATTTCCTGCTTTGCTGATGGAAAATTAGACTCTT 59
orangutan       -GCCGGTTATGTGGCATTCTTACATGATTTCGTGCTTTGCTGATGGAAAATTAGACCCTT 59
elephant        -GCTGGTTATGTGGCATTCTGACATGATTTCATGCTTTGCTGATGGAAAATTAGACCCTT 59
cat             -GCCGGTTATGTGGCATTCTTACATGATTTCATGCTGTGCTGATGGAAAATTAGATCCTT 59
medaka          --CTTGTTGTGTAGCATTCTCCCTTGATTTCATGCTTTGCTGATGGAAAATTAGGCC-TG 57
fugu            CGCTTGTTGTGTGGCATTCTCTCTTGATTTCACGCTTTGCTGATGGAAAATTAGACCCTG 60
tetraodon       CGCTTGTTGTGTGGCATTCTCTCC-GATTTCATGCTTTGCTGATGGAAAATTAGACCCTG 59
stickleback     -GCTTGTTGTGTGGCATTTTCTCTTGATTTCATGCTTTGCTGATGGAAAATTAGCCCCTG 59
zfish           --CTTGTTGTGTGGCATTCTCTGCCGATTTCATGCTTTGCTGATGGAAAATTAGGCCCCT 58
                                 *     ******  *** **   ************

frog            GTGTAT--T-AAAAAA--------- 54
macaque         GTGTAT--T-ATAAAAGAACAGAAG 81
rat             GTGTAT--T-CTAAAAGAATAGAAG 81
mouse           GTGTAT--T-CTAAAAGAATAGAAG 81
opossum         GTGTAT--T-ATAAAAGAATAGAAG 75
bushbaby        GTGTAT--T-ATAAAAGAATAGAAG 81
dog             GTGTAT--T-ATAAAAGAATAGAAG 81
armadillo       GTGTAT--T-ATAAAAGAATAGAAG 81
rabbit          GTGTAT--T-ATAAAAGAATAGAAG 81
horse           GTGTAT--T-ATAAAAGAATAGAAG 81
chicken         GTGTAT--T-ATAAAAGAATAGAAG 81
squirrel        GTGTAT--T-ATAAAAGAATAGAAG 81
cow             GTGTAT--T-ATAAAAGAATAGAAG 81
human           GTGTAT--T-ATAAAAGCATAGAAG 81
chimp           GTGTAT--T-ATAAAAGAATAGAAG 81
bat             GTGTAT--T-ATAAAAGAATAGAAG 81
orangutan       GTGTAT--T-ATAAAAGAATAGAAG 81
elephant        GTGTAT--T-ATAAAAGAATAGGAG 81
cat             GTGTAT--TTATAAAAGGAT----- 77
medaka          GCGCAT--CTCTGCGTGAAGAGGAG 80
fugu            GTGTAT--CTCTGCATGAAGAGGAG 83
tetraodon       GTGTATATCTCTGCATGAAGAGGAG 84
stickleback     ACGTAT--CCCTCCATGAAAAGG-- 80
zfish           GCGTAT--C---------------- 65
                  * **
```

CRCNE00001579

```
rat             AAAAGCTGACCTTGTTATATTTCCTGGGC-GGGTGTAGGTTTGTCCCATCACAGGCCCGA 59
mouse           AAAAGCTGACCTTGTTATATTTCCTGGGC-GGGTGTAGGATTGTCCCATCAGAGGCCCAA 59
chimp           AAAAGCTGACCTTGTTTTATTTCCTGGGC-GGGTGCAGGATTGTCCCGTCAGAACCCCGA 59
orangutan       AAAAGCTGACCTTGTTTTATTTCCTGGGC-GGGTGCAGGATTGTCCCGTCAGAACCCCGA 59
human           AAAAGCTGACCTTGTTTTATTTCCTGGGC-GGGTGTGGTATTGTCCCGTCAGAACCCCGA 59
macaque         AAAAGCTGACCTTGTTTTATTTCCTGGGC-GGGTGCAGGATTGTCCCATCAGAACCCCGA 59
rabbit          AAAAGCTGACCTTGTTATATTTCCTGGGC-GGGCGCAGGATTGTCCCGTCAGAACCCCAA 59
bushbaby        ----------------------------GC-GGGTGCAGGATTGTCC-GTCAGA-CCCCAA 30
squirrel        AAAAGCTGACCTTGTTATATTTCCTGGGC-GGGTACAGGATTGTCCCGTCAGAACCCGGA 59
horse           AAAAGCTGACCTTGTTATATTTCCTGGGC-GGGTGCAGGATTGTCCCGTCAGAACCCCGA 59
dog             AAAAGCTGACCTTGTTATATTTCCTGGGC-GGGTGCAGGATTGTCCCGTCAGAACCCCGA 59
cat             AAAAGCTGACCTTGTTATATTTCCTGGGC-GGGTGCAGGATTGTCCCGTCAGAACCCCGA 59
cow             -AAAGCTGACCTTGTTATATTTCCTGGGC-GGGTGCAGGATTGTCCCGTCAGAACCCCGA 58
bat             AAAAGCTGACCTTGTTATATTTCCCGGGC-GGGTGCAGGATTGTCCCGTCAGAACCCCGA 59
elephant        AAAAGCTGACCTTGTTATATTTCCTGGGC-GGGTGCAGGATTGCCCCTTCAGAACCCTGA 59
armadillo       AAAAGCTGACCTTGTGATATTTACTGGGT-GGGTGCAGGATTGTCCCGTCAGAACTCTGA 59
opossum         AAATGCTGACCTTGTGCTATTTCCTGGGT-GGGTGCAGAATTGTTCTATCAGTACCCAGA 59
chicken         AAATGCTGACCTTGTTATATTTACTGGGT-GGGTGCAGAATTGTTCCATCACTACCCCGA 59
frog            ----GCTGACCTTGTCATATTTAGAGGGT-GGGTGCAGAATTGTTCCATCACTATCCTGG 55
medaka          AAATGCTGACCTTGTCGTATTTA-AGGGT-GGCTGCAGAATTGTTACATCAGTAGCCAGA 58
fugu            AAACGCTGACCTTGTCATATTTACAGGGT-GGCTCCAGAATTGTTCCATCAGTAGCCAGA 59
tetraodon       AAATGCTGACCTTGTCTTATTTACAGGGG-GGCTGTAGAATTGTTCCATCAGTAGCCAGA 59
stickleback     AAACGCTGACCTTGTCATATTTACAGGGT-GGCTACAGAATTGTTCCATCAGTAGCCAGA 59
zfish           AAATGCTGACCTTGTCATATTTACGGGGCCGGCTGCAGAATTGTTCCATCAGTATCCAGA 60
shark           ----GCTGACCTTGTCATCTTTACTGGGT-GGGTACAGAATTGCTCCATCAGTACTCAGA 55
                                         *  **     *  ***     ***     *

rat             GGGCTGAACGATGGAGCTTGTCAAAGAGAGGCCCCGTGTAATGAATCATCTTATCACAGG 119
mouse           GGGCTGAACGATGGAGCTTGTCAAAGAGAGGCCTCGTGTAATGAATCATCTTATCACAGG 119
chimp           GGGCTGAACGATGGAGCTTGTCAAAGAGAGGCCCCATGTAATGAATCATCTTATCACAGG 119
orangutan       GGGCTGAACGATGGAGCTTGTCAAAGAGAGGCCCCATGTAATGAATCATCTTATCACAGG 119
human           GGGCTGAACGATGGAGCTTGTCAAAGAGAGGCCCCATGTAATGAATCATCTTATCACAGG 119
macaque         GGGCTGAACGATGGAGCTTGTCAAAGAGAGGCCCCATGTAATGAATCATCTTATCACAGG 119
rabbit          GGGCTGAACGATGGAGCTTGTCAAAGAGAGGCCCCATGTAATGAATCATCTGATCACAGG 119
bushbaby        GGACTGAACGATGGAGCT-GTCAAAGACAGGCCCCATGTAATGAATCATCT-ATCACAGG 88
squirrel        GGGCTGAACGATGGAGCTTGTCAAAGACAGGCCCCATGTAATGAATCATCTTATCACAGG 119
horse           GGGCTGAACGATGGAGCTTGTCAGAGAGCAGCCCCAGGTAATGAATCATCTTATCACAGG 119
dog             GGGCTGAACGATGGAGCTTGTCAAAGAGCAGCCCCAGGTAATGAATCATCTTATCACAGG 119
cat             GGGCTGAACGATGGAGCTTGTCAAAGAGCAGCCCCAGGTAATGAATCATCTTATCACAGG 119
cow             GGCCTGAACGATGGAGCTTGTCAAAGAGCATTCCCAGGTAATGAATCATCTTATCACAGG 118
bat             GGGCTGAACGATAGAGCTTGTCAAAGAGCGGCCCCAGGTAATGAATCATCTTATCACAGG 119
elephant        GGGCTGAACGATGGAGCTTGTCAAAGAGAGGCCCTGTGTAATGAATCATCTTATCACAGG 119
armadillo       GGGCTGAACGATGGAGCTTGTCAGAGAGAGGCCCTATGTAATGAATCATCT-ATCACAGG 118
opossum         GGGTTGAACGATGGAGCTTGTCAAAGGGAGCCCTTATCTAATGAATCATCTTATCGCAGG 119
chicken         GGGTTGAACGATGGAGCTTGTCAAAGTGAGGCCTTATCTAATGAATCATCTTATCGCAGG 119
frog            GTGTTGAACGATGGAGCTTGTCAAAGTGAGACCATATCTAATGAATCATCTTATCGCAGG 115
medaka          GGCTTGAACGATGATGCTTGTCATGGTGGGACTCTATCTAATGAATCATCTTATCTCAGG 118
fugu            GGCTTGAACGATGATACTTGTCATTGTGGGACTCTATCTAATGAATCATCTTATCGCAGG 119
tetraodon       GGCTTGAACGATGGTGCTTGTCGTTGTGGGACCCTATCAAATGAATCATCTTATCGCAGG 119
stickleback     GGCTTGAACGATGATGCTTGTCATTGTGGGTCCCTATCTAATGAATCATCTTATCGCAGG 119
zfish           GGCTTGAACGATGAGGCTTGTCAGACCAGAGCCCTATCTAATGAATCATCTTATCGCAGG 120
shark           GGGTTGAACGATGAACCTTGTCACAGTGGAGCATGATCCAATGAATCATCTTATCTCAGG 115
                *   ********    ** ***                ************ *** ****

rat             TGCACACCACAC--ATATTAAAGGAGG-GTTGTATGAACACTGTGCCCT 165
mouse           TGCACACCACAC--ATATTAAAGGAGG-GTTGTATGAACACCGTGCCCT 165
chimp           TGCACACCGCAC--ATATTAAAGGAGG-GTTGTATGAACACTGTGCCCT 165
orangutan       TGCACACCGCAC--ATATTAAAGGAGG-GTTGTATGAACACTGTGCCCT 165
human           TGCACACCGCAC--ATATTAAAGGAGG-GTTGTATGAACACTGTGCCCT 165
macaque         TGCACACCACAC--ATATTAAAGGAGG-GTTGTATGAACACTGTGCCCT 165
rabbit          TGCACACCACAC--ATATTAAAGGAGG-GTTCTGTGAACACTGTGCCCT 165
bushbaby        TGCACACCACAC--ATAT-AAAGGAGG-GCTGTATGAACACTGTGCCCT 133
squirrel        TGCACACCACAC--ATATTAAAGGAGG-GTTGTATGAACACTGTGCCCT 165
horse           TGCACACCACAC--ATATTAAAGGAGG-GTTGTATGAACACCGTGCCCT 165
dog             TGCACACCACAC--ATATTAAAGGACG-GTTGTATGAACACCGTGCCCT 165
cat             TGCACACCACAC--ATATTAAAGGAGG-GTTGTATGAACACGGTGCCCT 165
cow             TGCACACCACAC--ATATTAAAGGAGG-GATGTATGAACACCGTGCCCT 164
bat             TGCACACCGCAC--ATATTAAAGGAGG-ATCGTATGAACACCGTGCCCT 165
elephant        TGCACACCGCAC--ATATTAAAGGGGG-AATGTATGAACACTGTGCCCT 165
armadillo       TGCACACCACAC--ATATTAAAGGAGG-GTGGTACGAACGCTCTGCCCT 164
opossum         GGCGCACCACAC--ATATTAAAGGGGG-ATTGTATGAACAGTGTGCCCT 165
chicken         TGCACACCACAC--ATATTAAAGGAGG-ATTGTATGAACAATGTGCCCT 165
frog            TGCACACCACAT--GTATTAAAAGAGG-ATTGTATGAACAATGTGCCCT 161
medaka          TGTACACAGC----------------------------------------- 128
fugu            TGCACACAGCTC--ATATTAGAGGGAA-AATATATGGACAACGTGCCCT 165
tetraodon       TGCACATAGCTC--GTATTAGAGAGAA-AATGTATGGACAATGTGCCCT 165
stickleback     TGCACACAGCAT--ATATTAAAGGAGA-ATCGTGTGGACAATGTGCCCT 165
zfish           TGCGCACCGCTCCCATATTAAAGGAAAGATTGTATAGCCAATGTGCCCT 169
shark           TGCACACAGCGCC--CATTAAAGGAAG-ATTGTCTGAACAATGTGCC-- 159
                 *  **   *
```

```
CRCNE00005971

cow              --------------------ATA---CATCTTCAT----------GAC-ATGCTTTAAAT 26
rabbit           ----ATCTGTTGAATTATTAATA---CATCTTCAT----------GAC-ATGCTTTAAAT 42
mouse            --------------------ATA---TATCTTCAT----------GAC-ATGCTTTAAAT 26
rat              ----ATCTGATGAATTATTAATA---TATCTTCAT----------GAC-ATGCTTTAAAT 42
squirrel         ----ATCTGATGAATTATTAATA---TATCTTCAT----------GAC-ATGCTTTAAAT 42
opossum          ----ATCTGATGAATTATTAATA---CATCTTCAT----------GAC-ATGCTTTAAAT 42
chicken          ----ATCTGATGAATTATTAATA---CATCTTCAT----------GAC-ATGCTTTAAAT 42
elephant         ----ATCTGATGAATTATTAATA---CATCTTCAT----------GAC-ATGCTTTCAAT 42
chimp            ----ATCTGATGAATTATTAATA---CATCTTCAT----------GAC-ATCCTTTAAAT 42
human            ATTTATCTGATGAATTATTAATA---CATCTTCAC----------GAC-ATCCTTTAAAT 46
macaque          ----ATCTGATGAATTATTAATA---CATCTTCAT----------GAC-ATCCTTTAAAT 42
armadillo        ----ATCTGATGAATTATTAATA---CATTTTCAT----------GAC-ATGCTTTAAAT 42
dog              ----ATCTGATGAATTATTAATA---CATCTTCAT----------GAC-ATGCTTTAAAT 42
frog             ----ATCTTATGAATTATTAATA---TGTTTTGAT----------GAT-GTGCCTTAAAT 42
medaka           -TTAATTTGATACTTTATTTATGACGCGTCATCATTTATTTATTTGATTATTTTGTAAAT 59
stickleback      -TTAATCTGATGCTTTATTTATGGCCTGTCATCATTTATTTATTTGATTATTTTGTAAAT 59
fugu             ATTAATCTGATGCTTTATTTATGGAGTGTCATTATTTATTTATTTGATTATTTTGTAAAT 60
                                    **      *  * *           **   *    * ***

cow              TATTTAT------AACT--TAATTTT-------GAGTTT-------ATTGCT-------- 56
rabbit           TATTTAT------AACT--TAATTTT-------GAGTTT-------ATTGCT-------- 72
mouse            TATTTAT------AACT--TAATTTT-------GAGTTT-------ATTGCT-------- 56
rat              TATTTAT------AACT--TAATTTT-------GAGTTT-------ATTGCT-------- 72
squirrel         TATTTAT------AACT--TAATTTT-------GAGTTT-------ATTGCT-------- 72
opossum          TATTTAT------AACT--TAATTTT-------GAGTTT-------ATTGCT-------- 72
chicken          TATTTAT------AACT--TAATTTT-------GAGTTT-------ATTGCT-------- 72
elephant         TATTTAT------AACT--TAATTTT-------GAGTTT-------ATTGCT-------- 72
chimp            TATTTAT------AAGT--TAATTTT-------GAGTTT-------ATTGCT-------- 72
human            TATTTAT------AAGT--TAATTTT-------GAGTTT-------ATTGCT-------- 76
macaque          TATTTAT------AACT--TAATTTT-------GAGTTT-------ATTGCT-------- 72
armadillo        TATTTAT------AACT--TAATTTT-------GAGTTT-------ATTGCT-------- 72
dog              TATTTAT------AACT--TAATTTT-------GAGTTT-------ATTGCT-------- 72
frog             TATTTAT------ACCT--TAATTTT-------GAGTTT-------ATTGCT-------- 72
medaka           TATTTATCTATTGAATTATTAATTCTTCTTCATGACTTTTCGCTGCATTATTTATGACTT 119
stickleback      AATTTATCTATTGAATTATTAATTCTTCTTCATGACTTTCCCTTACATTATTCATGACTT 119
fugu             TATTTATCTATTGAATTATTAATTCTTCTTCATGACTTTCCCTTTCATTATTCATGACTT 120
                  ******      *  *  ***** *       ** ***       ***  *

cow              ---TTTAGACG-CTGCATTGTT--- 74
rabbit           ---TTTAGACG-CTGCATTGTT--- 90
mouse            ---TTTAGACG-CTGCATTGTT--- 74
rat              ---TTTAGACG-CTGCATTGTT--- 90
squirrel         ---TTTAGACG-CTGCATTGTT--- 90
opossum          ---TTTAGACG-CTGCATTGTT--- 90
chicken          ---TTTCGACG-CTGCATTGTT--- 90
elephant         ---TTTAGACG-CTGCATTGTT--- 90
chimp            ---TTTAGACG-CTGCATTGTT--- 90
human            ---TTTAGACG-CTGCATTGTT--- 94
macaque          ---TTTAGACG-CTGCATTGTT--- 90
armadillo        ---TTTAGACG-CTGCATTGTT--- 90
dog              ---TTTAGAGC-CTGCATTGTT--- 90
frog             ---TTTAGATG-TTGCTTTG----- 88
medaka           AATTTTAAATGATTGCATCA----- 139
stickleback      AATTTGAGATTATTGCATCA----- 139
fugu             AATTTGAGATTATTGCATCATTGTT 145
                    **   *    *** *

CRCNE00008385

frog             ---AGGCCGAAGACACTTGAAGTAATTACTTAAAGGCTTCCTTTCAAA-AATC--CCCC- 53
chicken          AGTAGGCCTAAGACACTTGAAGTAATTACTTAAAGGCTTCCTTTCAAAGAATC--TCCC- 57
opossum          AGTAGGCCTAAGACACTTGAAGTAATTACTTAAGGGCTTCCTTTCAAAGAATCTCCCCC- 59
platypus         AGTAGGCCTAAGACACTTGAAGTAATTACTTAAGGGCTCCCTTTCAAAGAATC--TTCC- 57
rat              AGCAGGCCTAAGATACTTGAAGCAATTACTTAAGGGCTCCCTTTCAAAGAATCTCTTCC- 59
mouse            AGCAGGCCTAAGATACTTGAAGCAATTACTTAAGGGCTCCCTTTCAAAGAATCTCTTCC- 59
squirrel         AGCAGGCCTAAGACACTTGAAGTAATTACTTAAGGGCTCCCTTTCAAAGAATCTCTTCC- 59
dog              AGTAGGCCTAAGACACTTGAAGTAATTACTTAAGGCCTCCCTTTCAAAGAATCT-CTCC- 58
horse            AGTAGGCCTAAGACACTTGAAGCAATTACTTAAGGACTCCCTTTCAAAGAATCTCTCCC- 59
bushbaby         AGTAGGCCTAAGACACTTGAAGTAATTACTTAAGGACTCCCTTTCAAAGAATCTCTCCC- 59
human            AGTAGGCATAAGACGCTTGAAGTAATTACTTAAGGACTCCCTTTCAAACAATCGCTCCC- 59
chimp            AGTAGGCATAAGACACTTGAAGTAATTACTTAAGGACTCCCTTTCAAACAATCGCTCCC- 59
macaque          AGTAGGCATAAGACACTTGAAGTAATTACTTAAGGACTCCCTTTCAAAGAATCTCTCCC- 59
orangutan        AGTAGGCATAAGACACTTGAAGTAATTACTTAAGGACTCCCTTTCAAAGAATCTCTCCC- 59
cat              ---AGGCCTAAGACACTTGAAGTAATTACTTAAGGACTCCCTTTCAAAGAATCTCTCCC- 56
cow              AGGAGGCCTAAGACACTTGAAGTAATTACTTAAGGACTCCCTTTCAAAGAATCTCTCCC- 59
armadillo        AGTAGGCCTAAGACACTTGAAGTAATTACTTAAGGACTCCCTTTCAAAGAATCTCTCCC- 59
bat              AGTAGGGCTAAGACACTTGAAGTAATTACTTAAGGACTCCCTTTCAAAGAATCTCTCCC- 59
rabbit           AGTAGGCCTAAGACACTTGAAGTAATTACTTAAGGACT-CCTTTCAAAGAATCTCTCTCG 59
medaka           --------------------GAGTAATTACTTAAGG-CTTTCCTTCACA---TCTCTCAGG 37
stickleback      --------------------GAGTAATTACTTAAGG-CTTTCCTTCACA---TCTCTCAAG 37
fugu             -CAGGAAGTGCGAGACCAGGAGTAATTACTTAAGG-CTTTCTGTCACA---TCTCCCAA- 54
tetraodon        ---------------ACCAGGAGTAATTACTTAAGG-CTTTCTGTCACA---TTTCCCAAG 42
zfish            ------------GACACTTGGAGTAATTACTTAAAG-CTTTCCCTCAGA---TCTCTCCCG 45
                             ** ********** * **  *  *** *   *
```

```
frog             -----------CGTTTTTAT----TAAATGCCTTTTAACATTAGCACTTTCCTCCTGACA 98
chicken          -----------CCTTTTTAT----TAAATGCCTTTTAACATTAGCACTTTCTTCCTGACA 102
opossum          -----------CTTTTTTAT----TAAATGCCTTTTAACATTAGCGCTTTCTTCCTGACA 104
platypus         -----------CCTTTTTAT----TAAATGCCTTTTAACATTAGCACTTTCTTCCTGACA 102
rat              -----------CTTTTT-AT----TAAATGCCTTTTAACATTAGCACTTTCTTCCTGACA 103
mouse            -----------CTTTTTTAT----TAAATGCCTTTTAACATTAGCACTTTCTTCCTGACA 104
squirrel         -----------TTTTTT-AT----TAAATGCCTTTTAACATTAGCACTTTCTTCCTGACA 103
dog              -----------CTTTTTTAT----TAAATGCCTTTTAACATTAGCACTTTCTTCCTGACA 103
horse            -----------TTTTTT-AT----TAAATGCCTTTTAACATTAGCACTTTCTTCCTGACA 103
bushbaby         -----------TTTTTT-AT----TAAATGCCTTTTAACATTAGCACTTTCTTCCTGACA 103
human            -----------TTTTTTTAT----TAAATGCCTTTTAACATTAGCACTTTCTTCCTGACA 104
chimp            -----------TTTTTTTAT----TAAATGCCTTTTAACATTAGCACTTTCTTCCTGACA 104
macaque          -----------TTTTTT-AT----TAAATGCCTTTTAACATTAGCACTTTCTTCCTGACA 103
orangutan        -----------TTTTTT-AT----TAAATGCCTTTTAACATTAGCACTTTCTTCCTGACA 103
cat              -----------TTTTTT-AT----TAAATGCCTTTTAACATTAGCACTTTCTTCCTGACA 100
cow              -----------TTTTTT-AT----TAAATGCCTTTTAACATTAGCACTTTCTTCCTGACA 103
armadillo        -----------TTTTTT-AT----TAAATGCCTTTTAACATTAGCACTTTCTTCCTGACA 103
bat              -----------TGTTTT-AT----TAAATGCCTTTTAACATTAGCACTTTCTTCCTGACA 103
rabbit           -----------TTTTTTTAT----TAAATGCCTTTTAACATTAGCACTTTCTCCCTGACA 104
medaka           AGT--------TTTTTTTATTACCCAAAAGCTTTTTAACATTATCACTCTCTTCCTGACA 89
stickleback      ACTCTT----TTTTTTTTATTACCCAAAAGCTTTTTAACATTAGCACTCTCTTCCTGACA 93
fugu             ----CTCTCTTTTTTTTTATAATCCAAAAGCTTTTTAACATTAGCACTCTCTTTCTGACA 110
tetraodon        TCTTTTCTCCTTTTTTTTATAATCCAAAAGCTTTTTAACATTAGCACTCTCTTTCTGACA 102
zfish            A----------CTTTTTTATTATCCAAAAGCTTTTTAACATTAGCACTGTCTTCCTGACA 95
                               **** **     *** ** *********** * ** **   ******

frog             AATTAAA----------------------------------------- 105
chicken          AATTAAATGAAATTTCATTACTCT-CCCTGACTTAATGGCATGACGCT 149
opossum          AATTAAATGAAATTTCATTACTCT-CCTTGACTTAATGGCATGAG--- 148
platypus         AATTAAATGAAATTTCATTACTCT-CCCTGACTTGGTGG--------- 140
rat              AATTAAATGAAATTTCATTACTCT-CCCTGACTTAATGGCATGAGACT 150
mouse            AATTAAATGAAATTTCATTACTCT-CCCTGACTTAATGGCATGAGACT 151
squirrel         AATTAAATGAAATTTCATTACTCT-CCCTGACTTAATGGCATGAGGCT 150
dog              AATTAAATGAAATTTCATTACTCT-CCCTGACTTAATGGCACGAGGCT 150
horse            AATTAAATGAAATTTCATTACTCT-CCCTGACTTAATGGCGTGAGGCT 150
bushbaby         AATTAAATGAAATTTCATTACTCT-CCCGGACTTAATGGCCTGAGGCT 150
human            AATTAAATGAAATTTCATTACTCT-CCCTGACTTAATGGCATTAGGCT 151
chimp            AATTAAATGAAATTTCATTACTCT-CCCTGATTTAATGGCATTAGGCT 151
macaque          AATTAAATGAAATTTCATTACTCT-CCCTGACTTAATGGCATTAGG-- 148
orangutan        AATTAAATGAAATTTCATTACTCT-CCCTGACTTAATGGCATTAGGCT 150
cat              AATTAAATGAAATTTCATTACTCT-CCCTGACTTAATGGCATGAGGCT 147
cow              AATTAAATGAAATTTCATTACTCT-CCCTGACTTAATGGCATGGGGCT 150
armadillo        AATTAAATGAAATTTCATTACTCT-CCCTGACTTAATGGCATGAGGCT 150
bat              AATTAAATGAAATTTCATTACTCT-CCCTGGCTTAATGGCATGAGACT 150
rabbit           AATTAAATGAAATTTCATTACTCT-CGCGGACTTAATGGCATGAGACT 151
medaka           AATTAAATGAAATTTCATTACTCT-AGCCATCCTAATGGCATGAAACT 136
stickleback      AATTAAATGAAATTTCATTACTCT-AGCCATCCTAATGGCATGAGACT 140
fugu             AATTAAATGAAATTTCATTACTCC-GGCCGCCCTAATGGCATGAAACT 157
tetraodon        AATTAAATGAAATTTCATTACTCCCGGCTGCTCTAATGGCGTGGGACT 150
zfish            AATTAAATGAAATTTCATTACTCC-AGCCATCCTAATGGCATGAAACT 142
                 *******

CRCNE00005870

dog              CTGAAAAGGCTTGGGTTGATTAATATTTTCTGCTGTCCGGTAACCATGGCAATAGGGTCA 60
bushbaby         CTGAAAAGGCTAGGGTTGATTAATATTTTCTGCTGTCCGGTAACCATGGCAATAGGGTCA 60
human            CTGAAAAGGCTCGGGTTGATGAATATTTTCCGCTGTCCGGTAACCATGGCAATAGGGTCA 60
chimp            CTGAAAAGGCTTGGGTTGATTAATATTTTCCGCTGTCCGGTAACCATGGCAATAGGGTCA 60
orangutan        CTGAAAAGGCTTGGGTTGATTAATATTTTCCGCTGTCCGGTAACCATGGCAATAGGGTCA 60
elephant         CTGAAAAGGCTTGGGTTGATTAATATTTTCCGCTGTCCGGTAACCATGGCAATAGGGTCA 60
horse            CTGAAAAGGCTTGGGTTGATTAATATTTTCCGCTGTCCGGTAACCATGGCAATAGGGTCA 60
cow              CTGAAAAGGCTTGGGTTGATTAATATTTTCCGCTGTCCGGTAACCATGGCAATAGGGTCA 60
macaque          CTGAAAAGGCTTGGGTTGATTAATATTTTCCGCTGTCCGGTAACCATGGCAATAGGGTCA 60
rat              CTGAAAAGACTTAGGTTGATTAATATTTTCCGCTGTCCGGTAACCATGGCAATAGGGTCA 60
mouse            CTGAAAAGACTTAGGTTGATTAATATTTTCCGCTGTCCGGTAACCATGGCAATAGGGTCA 60
bat              CTGAAAAGGCTTGGGTTGATTAATATTTTCCACTGTCCGGTAACCATGGCAATAGGGTCA 60
rabbit           CTGAAAAGGCTTGGGTTGATTAATATTTTCCACTGTCCGGTAACCATGGCAATGGGGTCA 60
armadillo        CTGAAAAGGCTTGGGTTGATTAATATTTTCCGCTGTCCGGTAACCATGGCAATAGGGTCA 60
opossum          CTCAAAAGGCTTGGGTTGATTAATATTTTCCGCTGTCCGGTAACCATGGCAATAGGGTCA 60
platypus         CTGAAAAGGCTTGGGTTGATTAATATTTTCCGCTGTCCG-TAACCATGGCAATAGGGTCA 59
chicken          CTGTAAAGGCTTGGCTTGATTAATATTTTCCGCTGTCCGGTAACCATGGCAATGGGGTCA 60
frog             CTGAAAAGGCTCGGCTTGATTAATATTTTCTGCTGTCCGGTTACCATGGCAATGGGGTCA 60
shark            CTGAAAAGTCGGGAGTTGATTAATATTTTCCATTGTCCTGTAACCATAGCAATAGGGTCA 60
fugu             CTTCCAAGTGCAGGCTTGATTAATATTTTCCACTGTTCCGTAACCATGGCAATAAGGTCA 60
stickleback      -------GTGCAGGCTTGATTAATATTTTCCACTGTTCCGTAACCATGGCAATAAGGTCA 53
tetraodon        -TTGCAAGTGCAGGCTTGATTAATATTTTCCACTGTTCCGTAACCATGGCAATAAGGTCA 59
medaka           -TTGCAAGTGCAGGCCTGATTAATATTTCCCACTGCTCCGTAACCATGGCAATAAGGTCA 59
zfish            -----AAGTGCAGACTTGATTAATATTTTCTTCTGTTTGGTAACCATGGCAATGAGGTCA 55
                       *        **** ******* *   **     * ***** *****  *****
```

```
dog               GCAGATATAATACTTAATTTAATCTTCCAGCTGCCTCAGTGTGCTTCACCGGTGTCC--- 117
bushbaby          GCAGATATAATACTTAATTTAATCTTCCAGCTGCCTCAGTGTGCTTCACCAGTGTCC--- 117
human             GCAGATATGATACTTAATTTAATCTTCCAGCTGCCTCAGTGTGCTTCACCAGTGCCCCAA 120
chimp             GCAGATATAATACTTAATTTAATCTTCCAGCTGCCTCAGTGTGCTTCACCAGTGCCC--- 117
orangutan         GCAGATATAATACTTAATTTAATCTTCCAGCTGCCTCAGTGTGCTTCACCAGTGCCC--- 117
elephant          GCAGATATAATACTTAATTTAATCTTCCAGCTGCCTCAGTGTGCTT-------------- 106
horse             GCAGATATAATACTTAATTTAATCTTCCAGCTGCCTCAGTGTGCTTCAC----------- 109
cow               GCAGATATAATACTTAATTTAATCTTCCAGCTGCCTCAGTGTGCTTCACCAGTGTCC--- 117
macaque           GCAGATATAATACTTAATTTAATCTTCCAGCTGCCTCAGTGTGCTTCACCAGTGTCCCGA 120
rat               GCAGATATAATATTTAATTTAATCTCCCAGCTGCCTCAGTGAGCTTCACC---------- 110
mouse             GCAGATATAATATTTAATTTAATCTTCCAGCTGCCTCAGTGAGCTTCAC----------- 109
bat               GCAGATATAATACTTAATTTAATCTTCCAGTTGCCTCAGTGTGCTTCACCGGTGTCC--- 117
rabbit            GCAGATATAATACTTAATTTAATCTTCCAGCTGCCTCAGTGTGCTTCACCAGTGTCC--- 117
armadillo         GCAGATATAATACTTAATTTAATC------------------------------------ 84
opossum           GCAGATATAATACTTAATTTAATCTTCCAACTGCCTCAGTGTGCATCACCGGTGTCC--- 117
platypus          GCAGATATAATACTTAATTTAATCTTCCAACTGACTCAGTGTGCATCACCGGTGTCC--- 116
chicken           GCAGATATAATACTTAATTTAATCTTCCAACTGCCTCAGTGTGCATCACCAGTGTCC--- 117
frog              GCAGATATAATACTTAATTTAATCTTCTGGCTGGCTCTGTGTGCATCACCAATGTCC--- 117
shark             GCACATATAATACTTAATTTAATCTTGCCCCTGCTT--GTGTGCATCGCCAGTGTCC--- 115
fugu              CTGCATATGATACTTAATTGAATCTTCCCAGTGCTTTTGTGAGCATCTCCTGTGTCCGGA 120
stickleback       CTGCATATGATACTTAATTGAATCTTCTCAGTGCTTTTGTGCGCATCTCCTGTGTCTGAA 113
tetraodon         CTGCATATGATACTTAATTGAATCTTCCCAGTGCTTTTGTGTGTATCTCCTGTGTCCGGA 119
medaka            CTGCATATGATACTTAATTGAATTTTCCCTGTGCTTTTGTCTGTGCCTCCACTGTCCGGA 119
zfish             GTGGATATGATGCTTAATTGAATCTTTCCCGAGCTTTTGTGTTCACCTCGTGTGTCCAAA 115
                      **** **  ****** ***
```

CRCNE00008599

```
cat               AGGGCTCGCATTGTACTTTTCAAGTGGCATTGATTTGTTAAAGCGCTGAATTAGCATCTC 60
elephant          -GGGCTCGCATTGTACTTTTCAAGTGGCATTGATTTGTTAAAGCGCTGAATTAGCATCTC 59
cow               AGGGCTCGCATTGTACTTTTCAAGTGGCATTGATTTGTTAAAGCGCTGAATTAGCATCTC 60
squirrel          AGGGCTCGCATTGTACTTTTCAAGTGGCATTGATTTGTTAAAGCGCTGAATTAGCATCTC 60
shark             AGGGCTCGCATTGTACTCTTCAAGTGGCATTGATTTGCTAAAGCGCTGAATGAGCATCTC 60
frog              ---------------TTTCCAAGTGGCATTGATTTGTTAAAGCGCTGAATTAGCATGTC 44
rabbit            AGCGCTAGCATTGTGCTTTTCAAGTGGCATTGATTTGTTAAAGCGCTGAATTAGCATCTC 60
human             AGGGCTCACATTGTGCTTTTCTAGTGGCATTGATTTGTTAAAGCGCTGAATCAGCATCTC 60
macaque           AGGGCTCGCATTGTGCTTTTCAAGTGGCATTGATTTGTTAAAGCGCTGAATTAGCATCTC 60
chimp             AGGGCTCACATTGTACTTTTCAAGTGGCATTGATTTGTTAAAGCGCTGAATTAGCATCTC 60
orangutan         AGGGCTCACATTGTACTTTTCAAGTGGCATTGATTTGTTAAAGCGCTGAATTAGCATCTC 60
chicken           AGGGCTCGCATTGTACTTTTCAAGTGGCATTGATTTGTTAAAGCGCTGAATTAGCATCTC 60
horse             AGGGCTCGCATTGTACTTTTCAAGTGGCATTGATTTGTTAAAGCGCTGAATTAGCATCTC 60
platypus          AGGGCTCGCATTGTACTTTTCAAGTGGCATTGATTTGTTAAAGCGCTGAATTAGCATCTC 60
opossum           AGGGCTCGCATTGTACTTTTCAAGTGGCATTGATTTGTTAAAGCGCTGAATTAGCATCTC 60
armadillo         AGAGCTCGCATTGTACTTTTCAAGTGGCATTGATTTGTTAAAGCGCTGAATTAGCATCTC 60
dog               AGGGCTCGCATTGTACTTTTCAAGTGGCATTGATTTGTTAAAGCGCTGAATTAGCATCTC 60
bushbaby          AGGGCTCGCATTGTACTTTTCAAGTGGCATTGATTTGTTAAAGCGCTGAATTAGCATCTC 60
rat               AGGGCTCTCATTGTACTTTTCAAGTGGCATTGATTTGTTAAAGCGCTGAATTAGCATCTC 60
mouse             AGGGTTCTCATTGTACTTTTCAAGTGGCATTGATTTGTTAAAGCGCTGAATTAGCATCTC 60
bat               AGGGCTCACATTGTACTTTTCAAGTGGCATTGATTTGTTAAAGCGCTGAATTAGCATCTC 60
medaka            AGGATGTGGATTGCTCTCCTCAAGTGGCATTGATTTGTCAAAGCGCTGAATTAGCAGA-- 58
stickleback       AGGGTGCGGATTGTGCTCCTCAAGTGGCATTGATTTGTTAAAGTGCTGAATTAGCAGA-- 58
tetraodon         AGGGTGCAGACTGTGCTCTTCAAGTGGCATTGATTTGTTAAAGTGCTGAATTAGCACA-- 58
fugu              AGGGTGCAGATTGTACTCTTCAAGTGGCATTGATTTGTTAAAGTGCTGAATTAGCACA-- 58
zfish             ------------GTACTTGTCAAGTGGCATTGATTTGTTAAAGTGCTGAATTAGCACA-- 46
                              *   * ***************  **** ******* ****

cat               TGGCCAGCTGC 71
elephant          TGGCCAGCTGC 70
cow               TGGCCAGCTGC 71
squirrel          TGGCCAGCTGC 71
shark             GGGCCAGCTGC 71
frog              AGGCCAGCTGC 55
rabbit            -GGCCAGCTGC 70
human             AGGCCAGCTGC 71
macaque           AGGCCAGCTGC 71
chimp             AGGCCAGCTGC 71
orangutan         AGGCCAGCTGC 71
chicken           AGGCCAGCTGC 71
horse             AGGCCAGCTGC 71
platypus          AGGCCAGCTGC 71
opossum           AGGCCAGCTGC 71
armadillo         CGGCC------ 65
dog               CGGCCAGCTGC 71
bushbaby          CGGCCAGCTGC 71
rat               CGGCCAGCTGC 71
mouse             CGGCCAGCTGC 71
bat               CGTCCAGCTGC 71
medaka            A---------- 59
stickleback       AGTCCAGCTGC 69
tetraodon         AATCCACCTGC 69
fugu              AGTCCAGCTGC 69
zfish             -----------
```

CRCNE00003046

```
rat             GGACCTGAGTGACGAGGGTCCCATTTCCAGCACAGCATGGAATGGGTGATCTCCTCAACT 60
mouse           GGACCCGAGTGACGAGGGTCCCATTTCCAGCACAGCATGGAATGGGTGATCTCCTCAACT 60
human           GGGCCCGAGCGCCGAGGGTCCCATTTCGCACACAGCATGGAGTGGGTGAGCTCCTCAACT 60
chimp           GGGCCCGAGCGCCGAGGGTCCCATTTCACACACAGCATGGAGTGGGTGAGCTCCTCAACT 60
orangutan       GGGCCCCGGTGCCGAGGGTCCCATTTTGCACACAGCATGGAGTGGGTGAGCTCCTCAACT 60
macaque         GGACCCGAGTGCCGAGGGTCCCATTTCTCACACAGCATGGAGTGGGTGAGCCCCTCAACT 60
horse           GGACCCGAGTGGCGAGGGTCCCATTTCCAGCACAGCATGGAGTGGGTGAGCCCCTCAACT 60
elephant        GGACCCGAGTGACGCGGGTCCCATTTCCAACACAGCATGCAGTGGGTGAGCCCCTCAACT 60
bushbaby        GGACCCGAGCGGCGCGGGTCCCATTTGCCGCACAGCATGCAGTGGGTGAGCCC-TCAACT 59
cow             GGACCCGAGTGACGAGGGTCCCATTTCCAACACAGCATGGAGTGGGTGATCTCCTCAACT 60
dog             GGACCCGAGTGACGAGGGTCCCATTTCCAACACAGCATGGAGTGGGTGATCTCCTCAACT 60
chicken         ------GAGTGACGAGGGTCCCATTTCCAACACAGCATGGAGTGGGTGATCTTGTCAACT 54
armadillo       GGGCCGGAGTGACGAGGGTCCCATTTCCAACACAGCATGGAGTGGGTGATCTCCTCAACT 60
opossum         GGACCAGAGTGACGAGGGTCCCATTTCCAACACAGCATGGAGTGGGTGATCTCCTCAACT 60
frog            GGACCAAAGTGATGAGGGTCCCATTTCAAACACAGCATGGAGTTGGTGATCTCCTCAACT 60
squirrel        GGACCCGAGTGACGAGGGTCCCATTTCCAACCCAGCATGGAGTAGGTGATCTCCTCAACT 60
fugu            GGGCCACTGTGACTGGGGTCCCATTTCAGGCCCAGCGTGGAGCCGGAGAGCTCCTCAGCT 60
tetraodon       GGGCCACTGTGACTGGGGTCCCATTTCAGGCCCAGCGTGGAGCCGGAGAGCTCCTCAGCT 60
stickleback     GGGCCACAGTGACTAGGGTCCCATTTCAGGCCCAGCGTGGAGCCGGAGAGCTTCTCAGCT 60
zfish           --------GTAACTGGGGTCCTATTTCAGCCGCAGCGTGAAGCCACTGAGTTCCTCAGCT 52
                        *      ****** ****    * **** ** *      **     *** **

rat             GTAAAATGAAGCCGTAAGAGCTCAGCATGAAAATGTATTAGCAACAGTTAGATAAGG-GA 119
mouse           GTAAAATGAAGCCGTAAGAGCTCAGCATGAAAATGTATTAGCAACAGTTAGATAAGG-GA 119
human           GTAAAATGAAGCCGTCAGAGCACAGCGTGAGAATGTATTAGAGACAGTTACACGA-G-G- 117
chimp           GTAAAATGAAGCCGTCAGAGCACAGCGTGAAAACGTATTAGAGACAGTTACACG--G-G- 116
orangutan       GTAAAATGAAGCCGTCAGAGCACAGCTTGAAAATGTATTAGAGACAGTTACACGA-G-G- 117
macaque         GTAAAATGAAGCCGTCAGAGCACAGCATGAAAATGTATTAGAGACAGTTACACGA-G-G- 117
horse           GTAAAATGAAGCCGTAAGAGCACAGCATGAAAATGTATTAGCAACAGTTAGATAAGG-G- 118
elephant        GTAAAATGAAGCCATAAGAACTCAGCGTGAAAATGTATTAGCAACAGTTAGATAAGG-GA 119
bushbaby        GTAAAATGAAGCCGCGAGAGCACAGCGTGAAGATGTATTAGCAACAGTTAGATGAGG-G- 117
cow             GTAAAATGAAGCCGTAAGAGCACAGCGTGAAAATGTATTAGCAACAGTTAGATAAGG-GG 119
dog             GTAAAATGAAGCCGTAAGAGCACAGCATGAAAATGTATTAGCAACAGTTAGATAAGG-GG 119
chicken         GTAAAATGAAGCCATAAGAGCACAGCATGAAAATGTATTAGCAACAGTTAGATAAGG-AA 113
armadillo       GTAAAATGAAGCCATAAGAACCCAGCATGAAAATGTATTAGCAACAGTTAGATAAGG-GA 119
opossum         GTAAAATGAAGCCATAAGAGCACAGCATGAAAATGTATTAGCAACAGTTAGATAAGG-GA 119
frog            GTAAAATGAAGCCATCAGAGCACAGCATGAAAATGTATTAGCAACAGTTAGATAAGG-GA 119
squirrel        GTAAAATGAAGCCGTAAGAGCACAGCATGAAAATGTATTAGCAACAGTTAGACAAGG-G- 118
fugu            GTAAAATGAGAGCGTCGGAGCGCAGTGTGAAAATGGATTAGCCTCGCTGAGATAAGGGGA 120
tetraodon       GTAAAATGAGAGCGTCGGAGCGCAGTGTGAAAATGGATTAGCCTCGCTGAGATAAGGGGA 120
stickleback     GTAAAATGAGAGCGCCGGAGCGCAGTGTGAAAATGGATTAGCCTCGCTGAGATAAGGGGA 120
zfish           GTAAAATGAAAGCAGCGGAGCACAGTGAGAAAATGTATTAGCCTCGCTGTGATAAGGAGA 112
                *********   *    ** * ***   **  * * *****   *  *   *    *

rat             GGCCTCAGCAAGGCGTACAC-ACTTAGCAATTGGT-AAAATAATG-TCTGTACAAATTG- 175
mouse           GGCCTCAGCAAGGCGTTCAC-ACT-AGCAATTGGT-AAAATAATG-TCTGTACAAATTG- 174
human           GGCCCCAGCACGGCGCTCAC-ACTTT-CAATTGGT-AAAATAATG-TCTGTACAAATTG- 172
chimp           GGCCCCAGCACGGCGCTCAC-ACTTT-CAATTGGT-AAAATAATG-TCTGTACAAATTG- 171
orangutan       GGCCCCAGCACGGCGCTCAC-ACTTT-CAATTGGT-AAAATAATG-TCTGTACAAATTG- 172
macaque         GGCCCCAGCACGGCGCTCAC-ACTTT-CAATTGGT-AAAATAATG-TCTGTACAAATTG- 172
horse           GGCCTCAGCAAGGCGCGCAC-GCTTT-CAATTGGT-AAAATAATG-TCTGTACAAATTG- 173
elephant        GACCTCAGCAAGGCGTACAC-ACTTT-CAATGGGT-AAAATAATG-TCTGTACAAATTG- 174
bushbaby        GACCGCAGCAAGGCGTACGC-ACTTT-CAATTGGT-AAAATAATG-TCTGTACAAATTG- 172
cow             -ACCTCAGCAAGGCGTACAC-GCTTT-CAATTGGT-AAAATAATG-TCTGTACAAATTG- 173
dog             GACCTCAGCAAGGCGTACAC-ACTTT-CAATTGGT-AAAATAATG-TCTGTACAAATTG- 174
chicken         GACCTCAGCAAGGCGTACAC-ACTTT-CAATTGGT-GAAATAATG-CTTGTACAAATTG- 168
armadillo       GACCTCAGCAAGGCGTACAC-ACTTT-CAATTGGT-AAAATAATG-TCTGTACAAATTG- 174
opossum         GACCTCAGCAAGGCGTACAC-ACTTG-CAATTGGT-AAAATAATG-CCTGTACAAATTG- 174
frog            GACCTCAGCAAGGCGTATAC-ACTTT-AAATTGGT-AAAATAATG-CCTGTACAAATTG- 174
squirrel        GGCCTCAGCAAGGCGTACGC-ACTTT-CAATTGG--AAAATAATGGCCTGCACAAATTGG 174
fugu            GCCCTCAGCGAGCCATGCACAGCCCT-GAATGGGGGAAAAGAACT-GCTGTACAAATTG- 177
tetraodon       GCCCTCAGCGAGCCATGCACAGCCCT-GAATGGGGGAAAAGAACT-GCTGTACAAATTG- 177
stickleback     GCCCTCAGCAAGCCGTGCAC-GCCTT-CAATGGGGGAAAAGAACC-GCTGTACAAATTG- 176
zfish           GCCCTTAGCGAGTGACGCAC-GCCTT-CAGTGGGG-AAAATAACC-ACTGGACAAATTG- 167
                  **  ***  *       *  *     * * **   *** **     ** ********
```

```
rat             TTTTTGGTACATTAAGTACCTTT-CACAGCCAAAATTAAAAAAA------ 218
mouse           TTTTTGGTACATTAAGTACCTTT-CACAGCCAAAATTAAAAAAAAAAAAA 223
human           TTTTTGGTACATTAAGTACCTTT-CACAGCCAAAATTAATAAAAGAAGAA 221
chimp           TTTTTGGTACATTAAGTACCTTT-CACAGCCAAAATTAAAAAAAGAAGAA 220
orangutan       TTTTTGGTACATTAAGTACCTTT-CACAGCCAAAATTAAAAAAAGAAGAA 221
macaque         TTTTTGGTACATTAAGTACCTTT-CACAGCCAAAATTAAAAAAAGAAGAA 221
horse           TTTTTGGTACATTAAGTACCTTT-CACAGCCAAAATTAAAAAAA------ 216
elephant        TTTTTGGTACATTAAGTACCTTT-CACAGCCAAAATTAAAAAAAAAAAAA 223
bushbaby        TTTTTGGTACATTAAGTACCTTT-CACAGCCAAAATTAAAAAAAAAAAAA 221
cow             TTTTTGGTACATTAAGTACCTTT-CACAGCCAAAATTAAAAAAGAAAAAA 222
dog             TTTTTGGTACATTAAGTACCTTT-CACAGCCAAAATTAAAAAAAAAAAAA 223
chicken         TTTTTGGTACATTAAGTACCTTT-CACAACCAAAATTAAAAAAAGAAGAA 217
armadillo       TTTTTGGTACATTAAGTACCTTT-CACAGCCAAAATTAAAAAA-GAAGAA 222
opossum         TTTTTGGTACATTAAGTACCTTT-CACAGCCAAAATTAAAAAAAAAA--- 220
frog            TTTTTGGTACATTAAGTACTTTT-CACAGCCAATGTGACAAAAAAAAAAA 223
squirrel        TTTTTGGGACATTAAGTACCTTTTCCCAGCCAAAA--------------- 209
fugu            TTTTTGGTACATTAAGTACCTGT---TACTTCATTTCACAGATGGAAAA- 223
tetraodon       TTTTTGGTACATTAAGTACCTGT---TACTTCATTTCACAGATGGAAAA- 223
stickleback     TTTTTGGTACATTAAGTACCTGT---TACTTTATTTCACAGATGGAAAA- 222
zfish           TTTTTGCTACATTAAGTACCTGT---TACTTTATTTCACAGATGGAAAA- 213
                ******  ********** * *    *    *
```