Thesis submitted for the degree of Doctor of Philosophy at the University of Oxford

Analyses of functional sequence in mammalian and avian genomes

Chris M. D. Rands

HERTFORD COLLEGE

OXFORD

HILARY TERM

2014

Abstract

Analyses of functional sequence in mammalian and avian genomes

Thesis submitted for the degree of Doctor of Philosophy at the University of Oxford

Chris M. D. Rands, Hertford College, Hilary Term 2014

The first draft sequence of the human genome was published over a decade ago, yet interpreting the functional importance of nucleotides in genomes is still an ongoing challenge. I took a comparative genomic approach to identify functional sequence using signatures of natural selection in DNA sequences. Mutations that are purged or propagated by selection mark sequences of significance for biological fitness. I developed and refined methods for estimating the quantity of sequence constrained with respect to insertions and deletions (indels) between two genome sequences, a quantity I termed $\alpha_{selIndel}$. This sequence is evolving more slowly than surrounding neutral sequence due to the purging of deleterious indel variants, and thus this sequence is likely to be functional. I estimated $\alpha_{selIndel}$ between diverse mammalian and avian species pairs, and found a strong negative correlation between $\alpha_{selIndel}$ and the divergence between the species' genome sequences. This implies that functional sequence turns over rapidly as it is lost and gained over time. I quantified the variable levels of sequence constraint, and rates of sequence turnover, for different types of human biochemically annotated element. Furthermore, I found that similar rates of functional turnover have occurred across mammalian and avian evolution. Finally, I identified positively selected amino acid residues that may be important for Darwin's finch beak development, and found evidence of adaptively evolving reproductive proteins in the ancestral songbird lineage. Collectively these results demonstrate the wide-spread nature of lineage-specific functional sequence with implications for understanding species traits and the use of model organisms to inform human biology.

Acknowledgments

I am very grateful to Chris P Ponting for his tireless supervision, advice, guidance, and support throughout my DPhil.

I would like to show my gratitude to Gerton Lunter for many helpful suggestions and for developing computer code that was essential for my project. I thank Stephen Meader for useful discussions and for his previous research on the turnover of functional sequence. I thank Phil Green for his comments and ideas on the Neutral Indel Models.

I thank Matt Fujita, Lesheng Kong, and other co-authors for their collaboration on the Darwin's finch genome project. I thank Christoffer Nellåker for advice on computer programming, and Andreas Heger for building many useful computational scripts and pipelines. I thank Wilfried Haerty and Yang I Li for helpful discussions.

I thank Nick Ilott for jolly times, Jennifer Tan for amusing chats, Will Pembroke for refreshing tea breaks, and all my friends from outside work for helping me enjoy my time as a graduate, particularly Pritesh Chandarana and Ali Cavalla.

I thank my parents Mike and Gill Rands, and my sister Jessica Rands, for their ongoing encouragement and moral support.

Finally, I am grateful to the UK Medical Research Council for financial support, and Hertford College and The Society for Molecular Biology and Evolution for sponsorship to attend scientific conferences.

Table of Contents

Chapter 1: Introduction

1.1 Evolutionary biology

Before the advent of evolutionary theory, biology was largely a descriptive science, simply an attempt to catalogue the diversity of life on the planet. When Charles Darwin and Alfred Russel Wallace famously introduced the theory of evolution by natural selection and suggested a common descent of all life (Darwin and Wallace 1858; Darwin 1859), biology could be interpreted meaningfully for the first time. As Theodosius Dobzhansky famously later put it, "Nothing in biology makes sense except in the light of evolution" (Dobzhansky 1973).

Darwin was not the first to have conceptions of evolution. For example, Comte de Buffon (1780) introduced the idea that species could transform into different species, and Lamarck (1809) proposed a theory of species transmutation. There were even inclinations of evolutionary thinking as early as the ancient Greek and Chinese philosophers, such as Epicurus's work, which implied that although many forms of organism exist, only the functional forms survive, which encapsulates the essence of natural selection. These ideas, and those of many others, can be thought of as precursors to theories of evolution (Gould 2002). However, none of these early thinkers proposed a coherent system, and they all tended to invoke some logic we would now consider non-scientific, such as Lamarck's alchemical belief that species somehow became more complex due to an innate life force. This all changed when Darwin meticulously laid out the first comprehensive theory of evolution (Darwin 1859).

There were two fundamental conceptual breakthroughs of this early evolutionary biology led by Darwin. First, the idea of natural selection, namely that traits which are beneficial for an individual's biological fitness (that is traits that help an individual and its relatives to survive and reproduce) are more likely to spread through populations over time than traits

that are of no consequence or harmful for an individual's fitness. So, natural selection offers an explanation for how complex traits can be generated out of more simple biological systems. Even though a novel trait is unlikely to persist in any given small population, in large populations over long time periods such complexity can become established. By contrast, without natural selection, where the evolution of traits would be completely random, there is a negligible chance of any complexity becoming prevalent in populations and species.

The second major innovation of early evolutionary theory was the concept that all extant life, including humans, shares an origin from a common ancestor. This concept is crucial as it allows biology to be understood in a comparative framework, where shared attributes that are common between groups (for example, between different cells, individuals, populations, species, or phyla) can be interpreted as having a common origin from a shared ancestor, and different traits a more recent origin since the divergence of the groups. By comparing different groups, this allows the evolution of attributes to be traced back through time and gives an indication of which attributes are of fundamental importance for life.

Another very important scientific breakthrough that was published just a few years after Darwin's initial work was Gregor Mendel's study of pea plants that led him to establish some basic principles of genetic inheritance (Mendel 1865). Mendel found that pea plant characteristics, such as the flower colour or seed shape, were inherited through discrete hereditary units that we now describe as genes. His work initially seemed to contradict the principles of evolution, since selection appears to act on continuous traits. However, the rediscovery and promotion of Mendel's work by scientists such as William Bateson, and the advent of population genetics spearheaded by Sewall Wright, Ronald Fisher, and J.B.S. Haldane led to the fusion of genetics with evolutionary biology around the 1930s. This

reconciled the discrete principles of Mendelian inheritance with the previously seemingly continuous principles of evolution. The basic realisation was that evolution could proceed through the inheritance of discrete units if these units were sufficiently small that traits appeared to be continuous. This was a significant scientific breakthrough that has been branded the modern evolutionary synthesis (Gould 2002).

Subsequent to the modern evolutionary synthesis, population genetics and molecular biology continued to grow and expand as disciplines leading to the birth of molecular evolution, where biologists examine the evolution of genetic material, such as DNA, RNA and protein sequences. Importantly, this led to the recognition that natural selection is just one mechanism by which evolution occurs. Natural selection can be swamped by changes in the rates at which different genetic variants initially arise (mutation rate) and when traits are rare, stochastic forces (genetic drift) can lead to random changes in the abundance and distribution of genetic variants (Kimura 1983).

With the principles of the modern evolutionary synthesis and molecular evolution established, evolutionary biology could be understood for the first time in a quantitative molecular framework with traits that are propagated through populations deriving from some underlying genetic basis (with a degree of environmental interaction) that is favoured by natural selection. This has led to the ongoing attempts to link genetics to traits, or more properly genotypes to phenotypes. If we can comprehensively elucidate the genetic substrate of traits, then this opens up enormous possibilities to help solve some of the most important global issues, such as how to effectively treat human diseases and improve food production.

1.2 DNA sequence evolution

The most fundamental unit of molecular evolution is DNA, the underlying genetic material from which other genetic units, such as RNA transcripts and proteins, are derived. An eukaryotic genome consists of linear sequences of DNA nucleotide bases. There are four DNA nucleotide bases: adenine (A), thymine (T), cytosine (C), and guanine (G). These nucleotides can be grouped into two types based on their chemical structure: A and G are purines, which include two carbon-nitrogen rings, and C and T are pyrimidines, which consist of a single carbon-nitrogen ring. When an individual reproduces, a copy of the genome is transmitted to the individual's offspring in the next generation. In the case of diploid organisms, such as mammals and birds, a haploid gamete is produced by each parent and these fuse to form a diploid offspring that normally has two copies of each autosomal chromosome and a pair of sex chromosomes. However, the replication machinery makes 'mistakes', known as mutations, during the DNA replication copying process that occurs at meiosis, so a perfect copy of the genome is very unlikely to be made. These mutational events are unlikely to occur at any given site; typically mutation rates are estimated at 2.5×10^{-8} mutations per nucleotide site for humans. But given that human genomes are over 3 billion bases (gigabase; Gb) in size, many mutations accumulate in their genomes over time; an estimated 175 mutations per diploid genome per generation (Nachman and Crowell 2000).

Mutations can be broadly divided into five types: point mutations, insertions or deletions (indels), large structural changes, chromosomal abnormalities, and whole genome duplications. The most abundant type of mutation is point mutations, where one base is substituted for by another. Point mutations can be further sub-divided into transitions and transversions. In biological data transitions (that is purine to purine or pyrimidine to pyrimidine mutations; A ↔ G or C ↔ T mutations) are approximately 2–3 times more

likely to occur than transversions (that is purine to pyrimidine or pyrimidine to purine mutations; T ↔ A, T ↔ G, C ↔ A, or C ↔ G mutations). This is because mutations are more likely to occur between bases of similar chemical structure.

Secondly, there are indel events where sequences are duplicated or deleted. Indels can vary in size from just a single base pair to many million bases (megabases; Mb). Indel events occur about 8–14 times less frequently than point mutations (Lunter 2007). However, the effect size of indels is generally larger than for point mutations, in the sense that their occurrence is more likely to result in disruptive changes, and thus are more strongly selected against than point mutations (Montgomery et al. 2013). Indels are relatively understudied as a form of evolutionary change compared to point mutations, in part because the inference of indels is technically challenging (Montgomery et al. 2013).

The third category of mutations is large structural changes. These are essentially just long and complex indel events, but are often treated in a separate category from short indels due to their scarcity and complexity. These mutations tend to include those termed copy number variants (which are large duplication or deletion events), and often transpositions (which are cases where sequences are moved within the genome).

Fourthly, there are chromosomal abnormalities, where a large portion of a chromosome, such as a chromosomal arm, is duplicated, deleted, moved (translocated), or inverted. Sometimes an entire chromosome copy is either duplicated or deleted, so a diploid individual has either a single (monosomy) or three (trisomy) copies of a chromosome. The majority of such chromosomal changes affect the individual so severely that they are lethal prior to birth in humans. There are some exceptions where the genetic abnormalities are not lethal, but these individuals tend to have severely affected phenotypes. For example, trisomy 21 individuals have Down's syndrome that may entail severe learning difficulties,

and monosomy X results in Turner syndrome that manifests with physical symptoms such as short stature and a webbed neck.

Finally, the largest scale form of mutation is where the entire genome is duplicated, sometimes multiple times. This can lead to considerable innovation in a single generation in plants, where a genome duplication can have a dosage effect, such as increasing the size of the organism, as has happened several times during the domestication of the wheat crop (Dubcovsky and Dvorak 2007). However, whole genome duplications are rarely viable in amniotes, because their genomes tend to be too complex in structure to withstand such vast changes.

Since all extant organisms likely derive from a common ancestor, it is the accumulation of these mutations that underlies the genetic basis for the differences between species. So, the evolution of different species can be traced by examining similarities and differences between their genome sequences. Genomic sequences that are sufficiently similar between different species can be classified as shared directly by descent, meaning that these sequences were present in the common ancestor of the species. Such a classification process requires probabilistic modelling, and there are many complications to distinguish sequences in different species that are actually inherited from the same common ancestor from those that represent examples of recurrent mutation or convergent evolution. Even considering homologous sequences that are derived from a common ancestor, it can be important to distinguish single copy orthologous sequences, which share an identical evolutionary history, from paralogous sequence, where homologous sequence is a duplicated, and thus there may be many copies of sequences sharing a high sequence identity (Koonin 2005).

1.3 The C-value paradox

Genome sizes vary vastly across the tree of life, with the smallest cellular genome currently estimated at a size of just 160kilobases (Kb) (Nakabachi et al. 2006), while the largest eukaryotic genome is 149Gb in haploid size (Pellicer et al. 2010). Even restricting our consideration to vertebrate genomes, there is still considerable genome size variation, with mammalian haploid genomes typically about 3Gb, whereas avian genomes are approximately one third this size. Surprisingly, the variation in genome size does not appear to correlate well with any obvious conception of organismal complexity. For example, the onion, marbled lung fish and some amoeba genomes are respectively 5, 40 and 250 times larger than the human genome (Graur et al. 2013). This observation has been dubbed the C-value (Constant-value) or G-value (Genome-value) paradox or enigma (Thomas 1971).

There are some hypotheses to explain this paradox that suggest there is cryptic functional significance for much of the genome (for example, much sequence could have some structural role in the nucleus) (Eddy 2012). However, these explanations are generally dismissed since even if they could explain the C-value paradox, they fail to clearly resolve the related issue of mutational load (Eddy 2012). Mutational load is a feature predicted from population genetics: if all the human genome is functional in the sense that it is important for an individual's fitness, then there would be so many harmful (deleterious) mutations that humans would not be able to survive this burden. Whilst this feature can be partially resolved by invoking large scale interactions between different sites (epistasis) or selection that varies depending upon the other genotypes in the population or the environment (soft selection) (Kondrashov 1995), a simpler resolution is that much of the genome is functionless.

The most widely accepted explanation that attempts to resolve the paradox is that the majority of genomic sequence in most organisms' genomes is not of functional importance, in the sense that mutations in the sequence do not contribute towards discernible phenotypic changes at the whole organism level and do not significantly impact an individual's biological fitness (Eddy 2012). Such apparently functionless sequence has controversially been branded 'junk DNA' (Ohno 1972). This begs the fundamental question of how much of the human and other species' genomes is functional? This is an important question, since identifying DNA sequence that is functional in human genomes is key for interpreting genetic variants that contribute towards traits and disease susceptibility.

1.4 Defining functional sequence

Although functionality is an intuitive idea, it is in fact a slippery concept and it is nontrivial to define clearly what constitutes functional genomic sequence. Like many concepts in molecular biology, such as defining a gene (Gerstein et al. 2007), a rigorous all-encompassing definition that is independent of the context is probably not possible. However, definitions of functional sequence have been coarsely divided into two categories: "causal role" or "selective effect" (Graur et al. 2013).

The "causal role" definition broadly defines functional nucleotides as those that are involved in some discernible biological or biochemical process. This may equate functionality with a phenotype, although phenotype is also a difficult word to dissect, since phenotypes can be overt whole organism phenotypes such as loss of limbs (death being the most extreme and overt) or subtle molecular phenotypes such as transcription (the 'weakest' being the simple act of DNA replication as a biochemical phenotype). The causal definition is intuitive, but ultimately ambiguous.

By contrast, the "selective effects" definition describes functional nucleotides as those whose mutation affects an individual's fitness (so are subject to natural selection). The origin of this concept can be traced back to the population geneticists Ronald Fisher and Sewall Wright, who debated the role of natural selection and genetic drift in shaping genome evolution in the 1930s. Fisher advocated the use of deterministic models that assume selection is perfectly efficient (Fisher 1930). Under these models, evolution proceeds in theoretical populations of infinite size, and assuming consistent directional selection, all deleterious variants are eventually lost and all adaptive variants eventually fixed in a population. Meanwhile, Wright envisaged a much stronger role for the stochastic forces of genetic drift, so he modelled the evolution of DNA sequences in populations of finite size (Wright 1931). With small populations some deleterious variants may not be purged from a population, or may even be fixed within the population, due to chance factors (genetic drift) not governed by the individual's genetics. Furthermore, even highly adaptive variants may be lost by chance since they tend to start at low frequencies in the population. This debate was fully realised in the late 1960s, particularly by Motoo Kimura (Kimura 1968; Kimura 1983), who was the first to postulate that the majority of mutations may be neutral and thus of no selective consequence. This debate revolved around how much of the genome is evolving neutrally (or more precisely, effectively neutrally), and how much is subject to selection, essentially pre-empting the more recent debate about how much of the genome is subject to selection and thus likely to be functional.

In practice, there is likely to be some convergence between the "selective effects" and "causal role" definitions, but there is also a large grey area where these definitions do not agree. Note that in an ideal population, where the effective population size approaches infinity, all the genome would be under selection. For example, the miniscule energy cost required for the replication of each nucleotide would be selected against. In reality,

effective population sizes are finite, and thus there are regions of the genome under selection, and regions that are effectively neutral, where their fate is predominantly governed by genetic drift rather than natural selection. This is formalised as a mutant gene is nearly neutral if $|s| < 1/(2N_e)$, where s is the strength of selection and N_e the effective population size (Kimura 1983). Since this means that an element may be defined as functionless in a small population, but functional in a larger population, it does not entirely agree with our naive intuitions of what constitutes function. However, defining functional nucleotides as those subject to natural selection offers clarity, unlike definitions based on biochemical activity or phenotypes, and utility, since it can be used as a tool for prioritising the investigation of candidate loci associated with diseases and disorders.

Even accepting the "selective effects" definition from a conceptual standpoint, there are many practical difficulties in defining functional sequence as sequence that is subject to selection. First, selection can take several different forms. Negative selection (sometimes called purifying selection) is the removal of deleterious variants by selection, while positive selection (related to adaptive evolution) is the propagation of advantageous variants by selection. (Note that these variants could be advantageous at the cellular level and not necessarily lead to overt phenotypic differences at the whole organism level.) Given that negative and positive selection operate in opposing directions, their signature in genome sequences is expected to be very different. Furthermore, selection is not always unconditional, as it can be dependent on the context, such as the other genotypes present in the population or the environment. Conditional selection can lead to balancing selection, where genetic variants are maintained more-or-less stably at intermediate frequencies in a population, rather than the variants being fixed (where that variant reaches a frequency of 100% within a population), as is predicted by unconditional negative or positive selection. In very rare cases, frequency dependent selection may even lead to cyclic variation in the

frequencies of attributes, at least at the phenotypic level (Sinervo et al. 2001). These are not just theoretical predictions: increasingly there is evidence that genomes tend to have evolved under complex selective scenarios, such as negative selection, selective sweeps, and occasionally balancing or positive selection (Allison 1956; Takahata et al. 1992; Ureta-Vidal et al. 2003; Bersaglieri et al. 2004; Siepel et al. 2005; Eyre-Walker 2006). Since these different types and forms of selection will leave different genetic signatures in the DNA sequence, any study equating functionality with selection needs to consider which signatures of selection they can detect and which will be missed or misinterpreted.

A second issue related to the first is that selection does not act independently on each individual genomic locus, so each nucleotide can be dependent on its neighbouring nucleotides. Negative selection at a particular locus can lead to the removal of nearby genetic variants that are in linkage with the locus. This causes a reduction in genetic diversity around that locus, a process known as background selection (Charlesworth et al. 1993). Conversely, positive selection can lead to the 'genetic hitchhiking' of linked variants as they are swept to higher frequencies in the population (Smith and Haigh 1974). These linkage effects should be accounted for or, at least quantified, particularly when utilising intra-species (polymorphism) data, to ensure that patterns of selection are correctly interpreted.

Third, signatures of selection can be difficult to distinguish from other patterns in genomic sequences. Mutational biases are common and often driven by neutral evolutionary processes. For example, a CpG dinucleotide is hypermutable, because its C nucleotide is a frequent target of DNA methylation and highly prone to be mutated through spontaneous deamination to the T nucleotide (Ehrlich and Wang 1981). Patterns of demography can also recapitulate patterns of selection, particularly when the search for selection is intra-species, rather than inter-species. For example, a signature of negative selection is a skew

towards rare genetic variants, since purifying selection pushes deleterious variants to lower frequencies than would be expected under neutral evolution, but this same signature is recapitulated by a population expansion (Eyre-Walker et al. 2006). Such mutational biases or population effects are confounding variables that should be accounted for to accurately identify genomic regions that are under selection.

The fourth complication relates to a fundamental question in evolutionary biology that asks: what is the unit of selection? Another way of expressing this question is to ask: what is selection actually acting on? This is important since negative selection may not act to retain the DNA nucleotide sequences directly. Instead selection may maintain just the amino acid sequences of proteins, the lengths of DNA sequences (Lu et al. 2012), the structural profile of DNA (Parker et al. 2009), the structure of gene networks, or whole organismal phenotypes. Therefore, functions can be preserved across species despite underlying changes occurring in the DNA sequences. An all encompassing approach to identify signatures of selection would consider evolution through this hierarchical lens with selective pressure acting on these different levels. However, in practice studies tend to focus on detecting selection only at one level.

Despite these aforementioned caveats, many studies have defined function through an evolutionary lens of negative selection acting on DNA sequences due to the generally accepted ubiquitous nature of long-term purifying selection (Ureta-Vidal et al. 2003; Siepel et al. 2005). Although this is a limited definition, it provides a quantitative basis for answering the question of how much of the genome is subject to selection and thus may be functional. I subsequently adopt the notation α_{sel} to represent the quantity of nucleotides in a genome that is estimated to be under negative selection. This definition follows previous studies (Chiaromonte et al. 2003; Meader et al. 2010), although my definition deviates

from theirs in the sense that I use α_{sel} to represent the quantity, and not fraction, of sequence under purifying selection.

1.5 Identifying functional sequence

The question of how much of genomes is subject to negative selection and thus likely to be functional (the estimation of α_{sel}) could be addressed with the rise of whole genome sequencing. This culminated in the human and mouse draft genome sequences being published in the early 2000s (Lander et al. 2001; Venter et al. 2001; Waterston et al. 2002). With these two whole genome sequences there was the first opportunity to examine patterns of selection genome-wide in mammals using comparative genomic approaches.

The initial mouse genome paper (Waterston et al. 2002), using methods detailed in (Chiaromonte et al. 2003), provided the first genome-wide estimate of regions of α_{sel} by estimating the quantity of genomic sequence shared between human and mouse that is subject to negative selection. Their method first required the creation of alignments to identify homologous regions of the human and mouse genomes. Genome alignment is generally the first step after the assembly of genomes that is required for comparative genomic methods. They used the BLASTZ tool to align the two genome sequences (Schwartz et al. 2003), and then the alignments were processed either by chaining and netting or using a reciprocal best alignment approach (Kent et al. 2003). Both processing approaches attempt to remove spurious low quality alignments. Once a series of whole genome alignments were constructed, they looked for the archetypical signature of negative selection by identifying genomic sequences that were inferred to be constrained with respect to point mutations. Constrained sequences are those that evolve significantly more slowly than surrounding neutral sequence. After accounting for confounding effects, such as neutral mutation rate variation, mutations are initially assumed to fall randomly

across the genome, but only those landing in regions subject to negative selection will be preferentially purged by selection.

They looked to identify constrained sequence by first splitting alignments into windows of size Wbp (typically W=50bp). Those windows with fewer than T matches, a match being an identical nucleotide in each genome alignment (typically T=40bp), were discarded. Then a conservation score was obtained for each window that reflects the number of matches in that window compared to that expected under neutral evolution. The rate of neutral evolution was estimated from proximal ancestral repeat (AR) sequence (that is aligned transposable element derived sequence). By using a local neutral rate, rather than a genome-wide one, they partially account for variation in the neutral mutation rate due to mutational biases, such as those associated with varying G+C base compositional content.

They estimated that 5.2% of the human genome, or more specifically 5.2% of 50bp windows, are estimated to be under selective constraint between human and mouse (Chiaromonte et al. 2003). This established an important benchmark for subsequent studies that attempted to estimate α_{sel} between the human and other species' genomes. Additionally, since only approximately 1.2% of the human genome is predicted to consist of protein coding sequence (International Human Genome Sequencing Consortium 2004), their results suggested that the vast majority of the functional components of the human genome lie in the noncoding fraction. However, the methodology was limited by the data available at the time, so their conclusions only pertain to the quantity of constrained, and thus putatively functional, sequence present between human and mouse.

The publication of the dog genome provided a second opportunity to estimate α_{sel} genome-wide (Lindblad-Toh et al. 2005). Following a very similar methodology to that used for the human – mouse comparison (Chiaromonte et al. 2003); it was estimated that 5.3% of sequences are constrained between the human and dog genomes. The similarity of this

estimate of α_{sel} to that between human and mouse is of particular note. Since the human and mouse genomes are more divergent in terms of their sequence than the genomes of human and dog, this implies that the study identified virtually no lineage-specific constrained sequence that has acquired functionality over the divergence of human from mouse, but not over the divergence of mouse and dog.

In the years following the publication of the mouse and dog genomes, many draft mammalian genome sequences were released, thus permitting the estimation of α_{sel} from the comparison of whole genome sequences from multiple species. Adding additional species into sequence constraint analyses provides more power to trace patterns of evolution and potentially allows conservation scores to be predicted at the resolution of a single base pair. However, the particular species used in the analyses will affect the estimates depending on their phylogenetic relationships and history. These further genomic comparisons led to diverse estimates of α_{sel} at 3–15% (Ponting and Hardison 2011; Ward and Kellis 2012; **Figure 1.1**). All these estimates are much larger than the protein coding component of the genome and much smaller than the total size of the human genome. Nonetheless, the five fold variation in these estimates implies that methodological differences, including the choice of species for the analysis, have very large impacts for the estimation of α_{sel}.

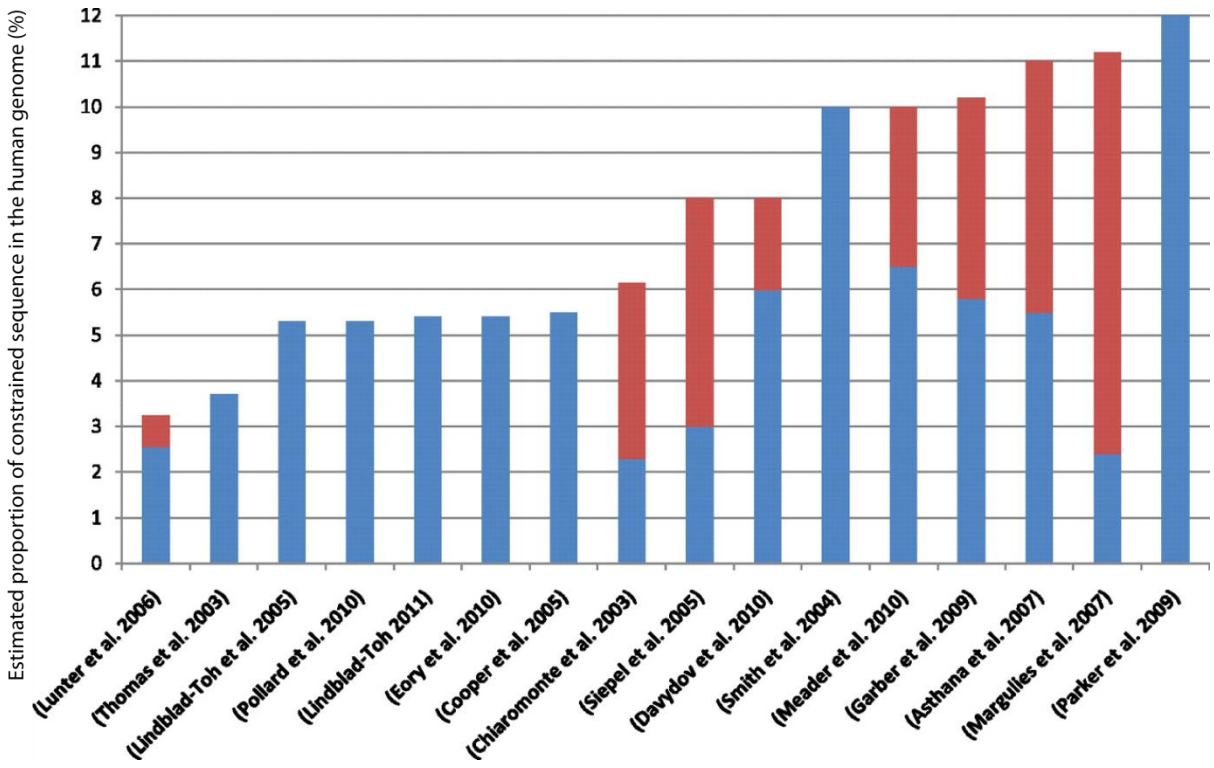

Figure 1.1: Estimates of α_{sel} from previous studies; the red parts of the bars show the lower and upper bound estimates for the studies that provided bounds for their estimates. The figure was taken from Ponting and Hardison (2011), and the references are as follows: (Chiaromonte et al. 2003; Thomas et al. 2003; Smith et al. 2004; Cooper et al. 2005; Lindblad-Toh et al. 2005; Siepel et al. 2005; Lunter et al. 2006; Asthana et al. 2007; Margulies et al. 2007; Garber et al. 2009; Parker et al. 2009; Davydov et al. 2010; Eory et al. 2010; Meader et al. 2010; Pollard et al. 2010; Lindblad-Toh et al. 2011).

1.6 Turnover of functional sequence

Since each species' lineage gains and loses functional elements over time, α_{sel} needs to be understood in the context of divergence between species. The divergence influences the estimate of α_{sel} in two ways. On the one hand, constrained sequence between closely related species, including lineage-specific constrained sequence, is harder to identify than more broadly conserved sequence because of a paucity of informative mutations, thereby reducing detection power. On the other hand, estimates of constraint between any group of species will only include sequence that was present in their common ancestor and predominantly conserved in all the species, with the consequence that turnover of functional sequence erodes α_{sel} estimates as the divergence increases. Assuming that the first effect can be controlled for, higher estimates of sequence constraint that are obtained between more closely related species (Smith et al. 2004; Meader et al. 2010) are thus indicative of the turn over of functional sequence (Ponting et al. 2011). Turnover is the gain or loss of function at a particular locus of the genome, as changes in the physical and genetic environment, and mutations in the sequence at the locus itself, cause the locus to switch from being functional to being non-functional or vice versa (**Figure 1.2**). Protein-coding genes are widely conserved (Waterston et al. 2002) and turnover is thought to play only a minor role in this class. Therefore, any significant turnover, if it exists, would predominantly involve noncoding functional sequences, as illustrated in **Figure 1.3**.

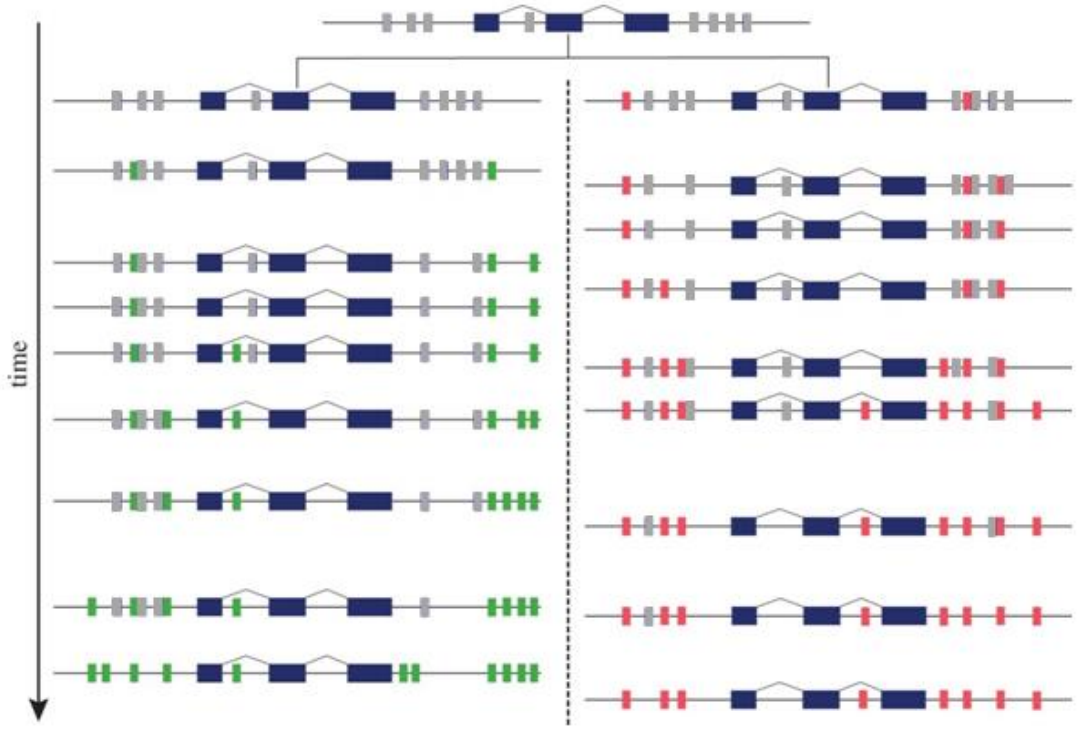

Figure 1.2: A schematic of how the turnover of functional noncoding sequence may proceed. Cis-regulatory elements (grey boxes) are shown flanking protein coding sequences (dark blue boxes). A common ancestral sequence is shown above, and over time the ancestral cis-regulatory elements are replaced by lineage-specific regulatory elements (red or green boxes). Eventually, all the regulatory elements are expected to turnover, so the regulatory elements are no longer recognisable between the two different lineages. The figure was modified from Harmston et al. (2013).

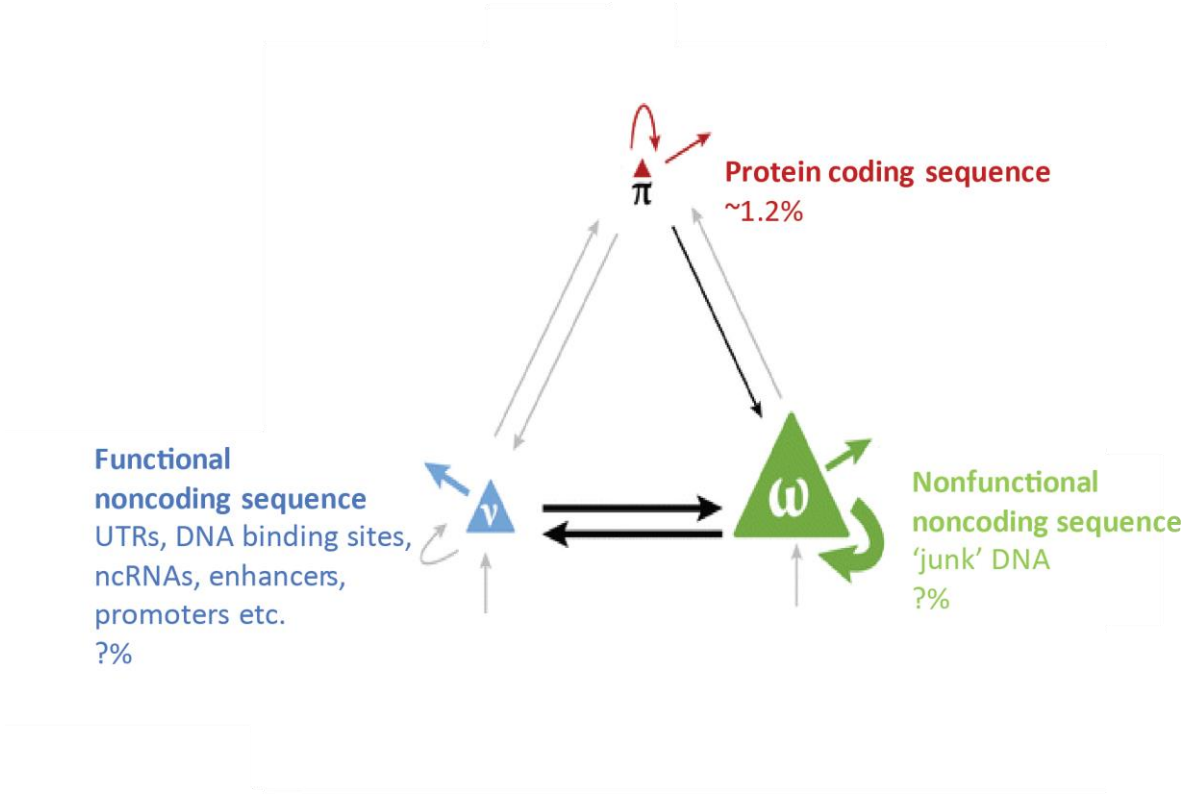

Figure 1.3: Illustration of how sequence changes between (functional) protein coding sequence, functional noncoding sequence, and non-functional noncoding sequence. The size of the triangles reflects the approximate expected relative proportions of sequence falling into each category, and the arrows indicate the transitions between the different categories, and duplication or deletion of sequence within a category. The thicker the arrows, the larger the expected amount of activity, so the vast majority of turnover between functional and non-functional sequence is expected to occur within noncoding sequences. The figure was modified from Ponting et al. (2011).

There are two previous studies that have examined the turnover of constrained genomic sequence in mammals without focussing on specific classes of functional element. Smith et al. (2004) examined the levels of sequence constraint with respect to point mutations across eight different species for a short genomic region of 1.8Mb in length and extending across ten genes. Multiple alignments were constructed across this region with MAVID, an alignment tool that takes a maximum-likelihood approach to progressively align sequences along a guide phylogenetic tree (Bray and Pachter 2004). They used a 50bp size sliding window to arbitrarily identify conserved alignment blocks as those with at least 90% sequence identity (45 or more matches between the two sequences in the 50bp window). The quantity of constrained sequence was then estimated as the excess of conserved blocks over the number of blocks predicted when simulating data under neutral evolution. The simulations were conducted under the Tamura-Nei substitution model implemented using the Evolver package of the programme PAML (Yang 2007) with realistic parameters specified based on mammalian sequence data and simulated additional regional variation in mutation rates. Applying this method across 21 species pairs from eight mammalian species, they find a negative correlation between the estimated α_{sel} and the divergence of the species pair, indicative of the turnover of constrained, and thus putatively functional, sequence. Their measure of pairwise divergence is simply the nucleotide substitution rate across the entire alignment between the two species estimated again under the Tamura-Nei substitution model.

Fitting an exponential model through their data allows estimation of the rate at which constrained sequence turns over. One metric for the rate of turnover is $d_{1/2}$, defined as the mean proportion of neutral substitutions per site that accumulate between the two species' genomes over the time it takes half of the functional sequence to turnover as it is lost and

gained. Using data from this study, the rate of turnover is estimated at $d_{1/2} = 0.14$ (Ponting and Hardison 2011).

This study was pioneering in terms of inferring the turnover of constrained sequence, but it was limited by the small quantity of data accessible at the time. Additionally, they define their sequence identity parameter to fix constraint at 1% of noncoding sequence between human and mouse. Therefore, the calibration of their method means that their estimates of α_{sel} are only meaningful in a relative sense; they do not shed light on the question of how much of genome sequences is constrained and thus likely to be functional in terms of absolute numbers.

The second more recent study that identified signatures of the turnover of constrained sequence improved on the previous work of Smith et al. (2004) by examining sequence constraint genome-wide. Meader et al. (2010) identified constrained sequence genome-wide between pairs of mammalian genomes, specifically examining sequence constraint with respect to insertions and deletions (indels), rather than point mutations. Their methodology is based on an earlier model of indel evolution (Lunter et al. 2006); the model estimates α_{sel} by quantifying the excess of long un-gapped alignment blocks observed in the data compared to those predicted under neutral evolution. The study estimates α_{sel} at 6.5–10% between closely related species pairs (human – macaque), and the rate of turnover is estimated at $d_{1/2} = 0.19$ (Ponting and Hardison 2011). These two studies were significant in the sense that they contextualised α_{sel} in the context of divergence and alerted the scientific community to the potential importance of functional sequence turnover genome-wide. However, the methodological approach adopted by Meader et al. (2010) had a number of shortcomings, including theoretical oversights in the original model and issues of poor data quality, so the quantitative assertions from the study (such as the estimations of α_{sel} and $d_{1/2}$) are thrown into doubt.

In addition to these studies broadly examining sequence constraint turnover, various studies have found evidence for the turnover of functional sequence in mammals through examining the evolution of specific classes of functional element defined with experimental approaches. One of the first studies to identify evidence for turnover found transcription factor binding sites (TFBSs) that lie in human promoter regions of 51 genes are not widely shared across species, with 32–49% of human TFBSs not found in rodents (Dermitzakis and Clark 2002). More recently, 23–41% of TF binding events were estimated to be conserved across human, dog and mouse for four liver TFs (Odom et al. 2007), while for two additional liver TFs, 7–14% of TF binding events are shared between human and mouse, and 15–20% between human and dog (Schmidt et al. 2010). However, not all studies support the notion that turnover of TFBSs is prevalent across mammalian evolution, since signatures of turnover can be misinterpreted due to technical artefacts in some cases, and it has been suggested that up to 94% of binding sites may be conserved between humans and rodents (Balmer and Blomhoff 2009).

There are anecdotal examples of turnover for enhancer elements. Comprehensive studies of human and mouse embryonic heart enhancers found these to be weakly conserved (Blow et al. 2010; May et al. 2012), despite human enhancer sequences largely driving expected tissue-specific expression in mouse embryonic heart tissue (May et al. 2012). Another study found that two mammalian hypothalamic enhancers have no homolog across non-mammalian vertebrates, yet are still able to drive specific expression patterns in zebrafish neurons (Domene et al. 2013). However, no previous studies have compared the rates of turnover across different categories of experimentally defined functional elements. Such research would provide a fundamental insight of how genome evolution proceeds for different types of functional element, but could also form the basis of a quantitative

framework for assessing the relevance of model organisms to inform specific questions of human biology via homology.

To contextualise the rate of turnover across the mammalian lineage, it is useful to have a comparative reference point. This can be provided by examining the rate of turnover across other lineages. There is some evidence for turnover of functional sequence in birds, flies, yeast, plants and bacteria (Mustonen and Lassig 2005; Moses et al. 2006; Doniger and Fay 2007; Kunstner et al. 2011a; Hupalo and Kern 2013). Birds are closely related to mammals and both groups are endotherms. Therefore, the avian lineage is the most obvious group in which to investigate turnover in order to inform mammalian biology. Additionally, it is arguably of importance to examine patterns of evolution for birds due to their involvement in the food industry and their importance in maintaining ecosystem services. Kunstner et al. (2011a) identified constrained regions in the untranslated regions (UTRs) flanking avian genes. They found that more closely related avian species share more constrained flanking sequence than more distantly related species, consistent with the notion of turnover of UTR sequences in birds. However, it is not known if this observed trend of turnover extends more broadly across the avian lineage and across different functional element types. Given that avian genomes are approximately one third the size of mammalian genomes, and have a different karyotype, consisting of macrochromosomes and microchromosomes, it is not clear that the genome-wide patterns of functional sequence evolution in birds will be found to recapitulate those observed in mammals.

There are indications that turnover may also be pervasive outside vertebrates. *Drosophila* species have been extensively studied as model organisms, and the sequencing of 12 *Drosophila* genomes provided an initial opportunity to infer evolutionary patterns genome-wide (Stark et al. 2007). TFBSs show evidence of turnover such that 5% of functional binding sites of the Zeste TF in *Drosophila melanogaster* were turned over in the last ten

million years since the divergence of *D. melanogaster* from the other four *Drosophila* species examined (Moses et al. 2006). More recently, it has been shown that more closely related *Drosophila* species share larger quantities of constrained sequence (Meader 2010). Further outside the vertebrate clade, TFBS in yeast show rapid turnover across four different *Saccharomyces* species (Doniger and Fay 2007); many categories of functional element show substantially increased 'alignability' with decreasing divergence in plants that could indicate functional turnover (Hupalo and Kern 2013); and 5% of TFBSs are predicted to turnover between *E. coli* and *Salmonella* bacterial species (Mustonen and Lassig 2005). This implies that turnover of functional sequence may be ubiquitous across the major kingdoms of life.

1.7 Positive selection

The turnover of functional sequence is related to the concept in molecular evolution of positive selection. I define positive selection as the propagation of advantageous mutations through a population due to natural selection. If turnover proceeds through mutations at the locus then gain-of-function mutations are expected to be positively selected. On the fixation of these positively selected mutations, the locus will become subject to purifying selection to maintain the newly acquired function. Thus, there is a strong link between sequence turnover and positive selection.

There has been considerable interest in identifying the molecular basis of positively selected traits because this aids our understanding of how individuals adapt to their environment and how complex traits arise and persist in populations. In humans, the rate of positive selection is usually estimated to be low, with generally only around a few percent of loci estimated to have evolved under positive selection (Eyre-Walker 2006). These loci cover genes that have been implicated with functions related to immune response, reproduction, olfaction, and nutrition (Nielsen et al. 2005; Voight et al. 2006;

Bakewell et al. 2007; Williamson et al. 2007). There are intuitive evolutionary arguments that can be postulated for why these types of genes may have been involved in human adaptation. Immune related genes may be evolving adaptively due to host-parasite coevolutionary arms races (Ma et al. 2006), reproductive genes may be subject to positive selection due to sexual conflict (Gavrilets 2000), while olfaction and nutrition genes are likely to be subject to differing selective pressures due to the heterogeneous environments in which humans have evolved. There are also more specific individual examples of positively selected genes involved in human traits. For example, positive selection in the *SLC24A5* and *SLC45A2* genes have been implicated in skin pigmentation, with variants in the genes associated with the differences in skin pigmentation patterns between European, African and Asian populations (Sabeti et al. 2007). *FOXP2*, a gene involved in speech development and formation, has experienced a greater than sixty fold increase in substitution rate in humans compared to other primates (Zhang et al. 2002).

However, given the relatively low rate of molecular adaptation estimated in humans (Eyre-Walker 2006), probably due to our small effective population size that reduces the efficacy of selection and due to our relatively long generation time, studies have often examined non-human model organisms to try and understand the molecular basis of beneficial traits and speciation. Adaptation has been studied in many vertebrate species, such as the cichlid fishes (Kocher 2004), but probably the most well-known vertebrate model organisms for studying adaptation are the Darwin's finches, which have been used as textbook examples for illustrating a diverse range of fundamental evolutionary processes including speciation, natural selection, and niche partitioning (Freeman and Herron 2003; Barton 2007; Futuyma 2009).

The Darwin's finches are ideal for such study since the different species show extensive variation in phenotypes and geographical distribution, and they adapt sufficiently rapidly

that some of their traits can be observed to change over a matter of mere decades, a minute period on evolutionary timescales. The varying morphology of the different finches on the various islands of the Galapágos archipelago they inhabit was influential for the inception of Darwin's theory of evolution by natural selection. Since the mid-20th century there has been extensive characterisation of the evolution of the Darwin's finches, including the evolution of traits such as their beak shape and size, and body size (Grant and Grant 1989).

More recently, the molecular basis of some Darwin's finch adaptive traits has been demonstrated. For example, molecular analysis has shown that the ground finch bill morphology correlates with a developmentally earlier and broader gene expression of *Bone morphogenetic protein 4* (*Bmp4*), especially in the large ground finch (*Geospiza magnirostris*). Functional experiments mimicking these changes in *Bmp4* expression using chicken embryos are consistent with its role in this beak trait (Abzhanov et al. 2004). Other experiments elucidated the roles of three further developmental factors, *Transforming Growth Factor beta Receptor Type II* (*TGFβRII*), *beta-Catenin* (*βCat*) and *Dickkopf-3* (*Dkk3*), at later stages of beak development that help in forming the bill shapes (Mallarino et al. 2011). Analyses also revealed an important role of change in the *Calmodulin* (*CaM*) expression pattern for the development of elongated bills of cactus finches (Abzhanov et al. 2006). With the advent of whole genome sequencing there is potential to unlock more information about the genetic basis of Darwin's finch traits, and to provide a genomic perspective on the evolution of this archetypal species.

1.8 Thesis scope and structure

In this thesis I develop and utilise comparative genomic evolutionary methods to provide insights into functional biology by identifying and characterising putatively functional DNA sequences in mammalian and avian genomes using the signatures of natural selection that are left in the genome sequences of extant species.

In **Chapter 2**, I describe the general data sets and methods that are common across multiple chapters of the thesis. The majority of methods are chapter-specific and are detailed within the relevant chapters.

Chapter 3 and **Chapter 4** are based on a project I led in collaboration with Stephen Meader, Chris Ponting, and Gerton Lunter. The project is described in a manuscript that is currently being revised for publication. I start **Chapter 3** by providing additional details on and refinements of a Neutral Indel Model (termed NIM1), originally developed by Lunter et al. (2006) and later Meader et al. (2010), that estimates the quantity of sequence that is constrained with respect to indels between the genome sequences of two species, a quantity I refer to as $\alpha_{selIndel}$. I also describe a second neutral indel model (termed NIM2) that estimates $\alpha_{selIndel}$ using a different approach from NIM1, and provides a partially independent validation of the results of NIM1. I then describe an additional processing step of alignments beyond the standard protocol developed at UCSC (Kent et al. 2003), after it was found that the genomic alignments were contaminated with poorly aligned sequence. This chapter also includes an extensive set of genome simulations that I conducted which validate the robustness of the neutral indel models over a wide variety of different biological parameterisations.

In **Chapter 4**, I apply these novel improved approaches for estimating $\alpha_{selIndel}$ to a wide variety of eutherian species pairs, but focussing on the human genome. I estimate that there is more than three times as much constrained sequence in the human genome than

such sequence that is shared between human and mouse, implying that there is an abundance of sequences with short lived lineage-specific functionality. As expected, most of the sequence involved in this functional turnover is noncoding, while protein coding sequence is stably preserved over long evolutionary timescales. Examining functional elements (such as TFBSs, enhancers, DNase 1 hypersensitivity sites), predominantly defined by experiments of the ENCODE project (Dunham et al. 2012), I find that the rate of functional turnover varies significantly across categories of functional noncoding elements. The results provide a pan-mammalian and whole genome perspective on how rapidly different classes of sequence have gained and lost functionality down the human lineage.

In **Chapter 5**, I identify patterns of sequence constraint and turnover in avian genomes using the same NIM1 approach described in the previous chapters. I find that constrained sequence is enriched in protein coding regions, and sequence constraint is correlated with base composition and chromosome size. I observe a negative correlation between estimates of $\alpha_{selIndel}$ and the divergence between the species pair, as measured by estimating the substitution rates for either ancestral repeats (these are transposable elements that align between the two species) or synonymous sites (these are mutated sites in protein sequence that leave amino acid sequences unchanged). This correlation implies that functional sequence turns over across avian evolution. Comparing these observations to the trends observed in mammals, I infer that the quantities of constrained sequence and rates at which this sequence turns over are similar between the mammalian and avian lineages, but that there may be more constrained sequence in mammalian than avian genomes.

In **Chapter 6**, I present evolutionary rate analyses that examine patterns of positive selection in the *Geospiza magnirostris* (large ground finch) genome, based on our

published work (Rands et al. 2013). I identify two genes, *POU1F1* (POU domain, class 1, transcription factor 1; also known as Pit1, growth hormone factor 1) and *IGF2R* (insulin-like growth factor 2 receptor), with candidate positively selected sites that may have been involved in Darwin's finch beak development. I validate that these mutations are true biological events with sequence data from other Darwin's finch species. I also examine patterns of selection more broadly across the songbird lineage. I find 47 predicted passerine-specific positively selected genes, and these are highly enriched in genes with cilium related functions. Through protein sequence analysis, I infer that these positively selected genes may be examples of adaptively evolving reproductive proteins.

In **Chapter 7**, I summarise the findings of the analysis chapters, and provide some future perspectives.

Chapter 2: Materials and methods

2.1 Genome assemblies

2.1.1 Sources of assemblies

Whole genome sequences assembled into chromosomes or a series of scaffolds were the starting point for the majority of my analyses. Most of the assemblies that I used were released for public use via the University of California Santa Cruz (UCSC) ftp server (hgdownload.cse.ucsc.edu), with the following exceptions (**Table 2.1**). The ferret genome assembly was accessed via the Broad Institute. The large ground finch genome assembly was obtained from the assembly team at the University of California, Davis (UC Davis). The assemblies of the adelie and emperor penguins were provided directly by the Beijing Genomics Institute (BGI), where the penguin genome assemblies were sequenced and assembled.

2.1.2 Qualities of assemblies

The assemblies vary in size from 0.99Gb for a Darwin's finch (*Geospiza magnirostris*) genome to 3.10Gb for the human genome, with the typical mammalian genome assembly of approximately 2.6Gb, almost three times larger than the average avian genome assembly. All assemblies have N50 values of well over a megabase except for the *G. magnirostris* genome, implying that generally the assemblies are of good quality in the sense that they are not highly fragmented (**Table 2.1**). Extensive details of the *G. magnirostris* assembly are given in **Chapter 6**. The assemblies consist of between 10Mb (for the zebra finch assembly) and 337Mb (for the rat rn5 assembly) of 'N' bases that mark either ambiguous bases or un-assembled genomic regions. The relatively large number of 'N' bases in the human genome reflects the large genome assembly size with attempts to cover the difficult to sequence heterochromatic regions, such as the chromosome centromeres and telomeres, in addition to the euchromatic genome.

Table 2.1: The details of the whole genome assemblies that were used in this study.

Common species name	Latin species name	Assembly version[1]	Assembly size (Gb)	N50 value (Mb)	Number of 'N' bases (Mb)	Obtained from
Human	*Homo sapiens*	hg19	3.10	155.27	234.35	UCSC
Mouse	*Mus musculus*	mm10	2.73	130.69	78.09	UCSC
Mouse	*Mus musculus*	mm9	2.73	131.74	105.42	UCSC
Mouse	*Mus musculus*	mm8	2.66	132.09	97.17	UCSC
Rat	*Rattus norvegicus*	rn5	2.91	154.58	336.85	UCSC
Rat	*Rattus norvegicus*	rn4	2.83	143.00	267.83	UCSC
Cattle	*Bos taurus*	bosTau7	2.98	85.12	176.45	UCSC
Dog	*Canis lupus familiaris*	canFam2	2.53	67.21	146.68	UCSC
Horse	*Equus caballus*	equCab2	2.48	91.57	55.74	UCSC
Guinea pig	*Cavia porcellus*	cavPor3	2.72	27.94	59.85	UCSC
Rabbit	*Oryctolagus cuniculus*	oryCun2	2.74	111.80	133.47	UCSC
Bushbaby	*Otolemur garnettii*	otoGar3	2.51	13.85	160.19	UCSC
Panda	*Ailuropoda melanoleuca*	ailMel1	2.30	12.81	54.20	UCSC
White Rhino	*Ceratotherium simum*	cerSim1	2.46	26.28	97.51	UCSC
Ferret	*Mustela putorius furo*	mpf_v1	2.41	93.35	132.85	Broad Institute
Chicken	*Gallus gallus*	galGal4	1.05	90.22	14.08	UCSC
Chicken	*Gallus gallus*	galGal3	1.12	94.23	57.90	UCSC
Zebra finch	*Taeniopygia guttata*	taeGut1	1.23	73.66	10.32	UCSC
Turkey	*Meleagris gallopavo*	melGal1	1.06	74.86	125.89	UCSC
Budgerigar	*Melopsittacus undulatus*	melUnd1	1.12	10.61	30.76	UCSC
Medium ground finch	*Geospiza fortis*	geoFor1	1.07	5.26	24.01	UCSC
Large ground finch	*Geospiza magnirostris*	gm_v1	0.99	0.38	156.09	UC Davis
Emperor penguin	*Aptenodytes forsteri*	af_v1	1.26	5.07	71.90	BGI
Adelie penguin	*Pygoscelis adeliae*	pa_v1	1.23	5.05	54.16	BGI
Anolis lizard	*Anolis carolinensis*	anoCar2	1.80	150.64	97.79	UCSC

[1] UCSC identifier if available

2.1.3 Marking repeats in assemblies

Genomes are highly repetitive, for example roughly half the human genome consists of transposable element (TE) sequences (Lander et al. 2001); sequences that have increased in copy number over evolution due to their transposition and duplication. Masking these repetitive elements in genome assemblies is generally the first step prior to downstream analysis since repetition in the genome makes any sequence searches via homology difficult. This is because repeats can lead to spurious matches when sequences are aligned or mapped, and such searches tend to be computationally expensive. By marking repeats they can be differentially treated from non-repetitive regions during subsequent analyses. Therefore I obtained softmasked genome assemblies from UCSC, where repetitive regions annotated by RepeatMasker are marked as lower case letters in the genome sequences (Smit et al. 1996-2010). RepeatMasker screens DNA sequences for low complexity regions and interspersed repeats using pre-defined libraries. The programme applies a particular case of the Smith-Waterman-Gotoh algorithm, a form of local alignment with affine gap penalties allowing for a separate penalty for opening and extending gaps (Smit et al. 1996-2010). The libraries are based on curated catalogues of repeats defined by the de novo repeat finder Repbase that identifies repeat families using consensus sequences, and Dfam that takes a profile hidden Markov approach to repeat annotation (Wheeler et al. 2013). Where softmasked genome assemblies were not available, repeat annotations were constructed and assemblies masked as described in **Chapter 4** and **Chapter 6**.

2.2 Genome alignments

2.2.1 LASTZ

Where available, whole genome alignments were downloaded from UCSC; otherwise I constructed alignments myself. The alignments were built by applying UCSC's protocol as follows. Alignments were made initially using the LASTZ software, available from http://www.bx.psu.edu/miller_lab/, an updated version of BLASTZ (Schwartz et al. 2003),

which is a derivative of BLAST (Altschul et al. 1990). BLAST is one of the most widely used algorithms for finding high scoring pairwise alignments between a query and target sequence. It follows a seed-based approach to alignment, which is heuristic, and therefore not guaranteed to find the optimally scoring alignments between any two given sequences. However, seed-based approaches will perform well providing that a reasonable assumption is met, namely that the true alignments contain short sequences that are identical (or near-identical) between the two sequences. Heuristic approaches have the advantage over optimal alignment approaches that the methods compute rapidly and therefore are feasible over large alignments, such as whole genome-alignments between different vertebrate species.

Under my implementation of LASTZ, the alignment process started with the input of two softmasked sequences. The target sequence was the assembled genome sequence from one species, and the query sequence the genome sequence of the second species. I chose to use the genome assembly arranged on the fewest chromosomes or scaffolds as the target sequence to reduce the build up of spurious alignments in the early processing stages. To make the alignments computationally tractable, I divided up the target genome sequence by chromosome (or by scaffold if the assembly was not fully assembled into chromosomes). Each chromosome or scaffold of the target sequence was aligned with LASTZ against the entire genome sequence of the second species. The large redundancy that was generated in the alignments by taking this approach was dealt with later in the netting stage of post-alignment processing. Alignment parameters for each whole genome alignment and relevant scoring matrices are given in **Table 2.2** and **Table 2.3**.

Table 2.2: LASTZ parameterisations implemented for the different alignments. BLASTZ parameter names are shown in parentheses. Rows highlighted in bold represent alignments that I constructed, while the other alignments were built by UCSC.

Species Pair	HSP threshold (K)	Gapped threshold (L)	Y dropoff (Y)	Inner score (H)	Gap penalties (O,E)	Seed spec (T)	Scoring matrix
hg19 – equCab2	3000	**3000**	9400	0	400, 30	1	HOXD70
hg19 – cerSim1	3000	3000	9400	2000	400, 30	1	HOXD70
hg19 – otoGar3	**3000**	**3000**	**9400**	**2000**	**400, 30**	**1**	**HOXD70**
hg19 – canFam2	3000	3000	9400	0	400, 30	1	HOXD70
hg19 – ailMel1	2200	6000	3400	2000	400, 30	2	HOXD70
hg19 – bosTau7	3000	3000	9400	0	400, 30	1	HOXD70
hg19 – oryCun2	3000	3000	9400	0	400, 30	1	HOXD70
hg19 – cavPor3	3000	2200	9400	2000	400, 30	1	HOXD70
hg19 – mm10	3000	2200	9400	2000	400, 30	1	HOXD70
mm8 – rn4	3000	2200	9400	2000	400, 30	1	HOXD70
mm9 – rn4 (1)	3000	2200	9400	2000	400, 30	1	HOXD70
mm9 – rn4 (2)	4500	2200	15000	2000	600, 55	2	MR_old
mm10 – rn5	3000	3000	5000	2000	600, 55	2	MR_op
mm10 – equCab2	**3000**	**3000**	**9400**	**2000**	**400, 30**	**1**	**HOXD70**
mm10 – canFam2	**3000**	**3000**	**9400**	**2000**	**400, 30**	**1**	**HOXD70**
mm10 – bosTau7	3000	3000	9400	2000	400, 30	1	HOXD70
canFam2 – mpf_v1	**3000**	**3000**	**5000**	**2000**	**600, 55**	**2**	**MR_op**
canFam2 – equCab2	3000	2200	9400	2000	400, 30	1	HOXD70
canFam2 – bosTau7	3000	2200	9400	2000	400, 30	1	HOXD70

Species Pair	HSP threshold (K)	Gapped threshold (L)	Y dropoff (Y)	Inner score (H)	Gap penalties (O,E)	Seed spec (T)	Scoring matrix
galGal3 – af_v1	**3000**	**3000**	**9400**	**2000**	**400, 30**	**1**	**HOXD70**
galGal3 – pa_v1	**3000**	**3000**	**9400**	**2000**	**400, 30**	**1**	**HOXD70**
galGal3 – taeGut1	3000	2200	9400	2000	400, 30	1	HOXD70
galGal3 – melGal1	3000	2200	9400	2000	400, 30	1	HOXD70
galGal3 – gm_v1	**3000**	**3000**	**9400**	**2000**	**400, 30**	**1**	**HOXD70**
galGal4 – taeGut2	3000	3000	9400	2000	400, 30	1	HOXD70
taeGut1 – af_v1	**3000**	**3000**	**9400**	**2000**	**400, 30**	**1**	**HOXD70**
taeGut1 – pa_v1	**3000**	**3000**	**9400**	**2000**	**400, 30**	**1**	**HOXD70**
taeGut1 – gm_v1	**4500**	**3000**	**15000**	**2000**	**600, 55**	**1**	**HOXD70**
taeGut1 – melUnd1	**3000**	**2200**	**9400**	**2000**	**400, 30**	**1**	**HOXD70**
taeGut1 – geoFor1	3000	2200	9400	2000	400, 30	1	HOXD70

Table 2.3: Substitution scoring matrices used for the LASTZ alignments.

HOXD70					
		Base in query sequence			
		A	C	G	T
Base in target sequence	A	91	-114	-31	-123
	C	-114	100	-125	-31
	G	-31	-125	100	-114
	T	-123	-31	-114	91
MR_old					
		Base in query sequence			
		A	C	G	T
Base in target sequence	A	56	-109	-45	-137
	C	-109	100	-103	-45
	G	-45	-103	100	-109
	T	-137	-45	-109	56
MR_op					
		Base in query sequence			
		A	C	G	T
Base in target sequence	A	100	-139	-54	-170
	C	-139	95	-83	-54
	G	-54	-83	95	-139
	T	-170	-54	-139	100

My alignment protocol with LASTZ consisted of four steps. First, the query and target sequences were separately indexed into overlapping seed words of length L; L = 19bp for my implementation. Any softmasked bases in the seed were skipped over during the seeding because short repetitive sequences were likely to align spuriously to many different locations. The seed words of the two sequences were then compared to identify seeds as seed words that have 12bp of the 19bp as identical matches between the two sequences.

During the second step these seeds were then extended 1bp at a time in each direction according to some scoring scheme. This scoring requires a substitution matrix that rewards matches and penalises mis-matches. This can be derived from the data, but typically is specified a prori based on the HOXD70 matrix (**Table 2.3**), which was originally derived from alignment sequences between human and mouse (Chiaromonte et al. 2002). For alignments between less divergent species' pairs more stringent substitution matrices were used (**Table 2.3**). The HOXD70 matrix accounts for differences in the expected transition/transversion mutation rate: a transition is approximately 2–3 times more permissible than a transversion. The matrix has symmetry imposed on it with respect to target/query sequence or strand. As the seeds were extended 1bp at a time, the cumulative score of the extended seed was calculated according to the substitution matrix, and the extension was stopped when the current score dropped off by more than the x-drop threshold. Extended seeds that reached the high-scoring pair (HSP) threshold were kept as HSPs.

After this gap-free extension step, each HSP was reduced to an anchor point before the third stage of gapped extension began. This anchor point was the central base of the highest-scoring 31bp segment. The reduction to an anchor point, rather than starting from the HSP, allowed for the possibility that the highest scoring alignment actually included gaps in the already gap-free defined HSP. The anchor points were then extended independently in each direction and this extension was scored using the substitution matrix as before, but

additionally with gap penalties. There was an opening gap penalty for the first gap in a sequence, and then an extension gap penalty for subsequent adjacent gaps. The opening gap penalty was typically set as approximately five times higher than the extension gap penalty, due to the reasoning that indels are quite unlikely to arise, but a short indel is not much more likely to have arisen than a slightly longer one. This means that one long gap will be preferentially created in the alignment, rather than several small ones. The anchor points were extended in order, with those deriving from higher HSPs processed first. If the extension of an anchor point reached another anchor point that had yet to be processed, then this anchor point was not processed again later. This gapped extension continued until the cumulative score dropped off below a value specified by the y-drop threshold. At this point if the maximal score reached by the extended anchor point met the gapped threshold then the alignment was kept, otherwise it was discarded. An illustration of the components involved in the seeding and extension steps is given in **Figure 2.1**.

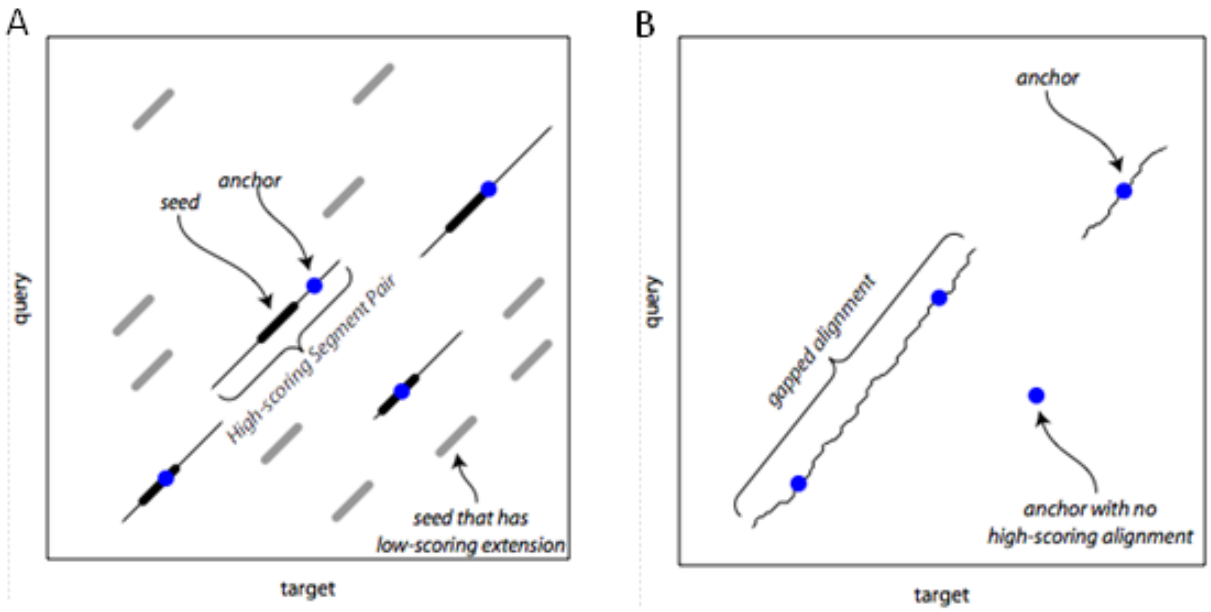

Figure 2.1: Schematic representations of the LASTZ seeding, gap-free and gapped extension steps. A. At the termination of gap-free extension, HSPs have been formed each of which contained a seed and can be reduced to a single anchor point that lies in a particularly high-scoring region of the HSP. B. After gapped extension, longer alignments were created, and overlapped alignments were joined. The figure was taken from the LASTZ website: http://www.bx.psu.edu/miller_lab/dist/README.lastz-1.02.00/README.lastz-1.02.00a.html.

The fourth and final step I implemented with LASTZ was interpolation. This consisted of repeating the gap-free and gapped alignment steps for regions of the sequences that were previously unaligned, but under more sensitive parameterisations than before. The specified inner score was used as the threshold for both what was previously the HSP threshold in the gap-free extension step, and the gapped threshold in the gapped extension step.

2.2.2 Chaining and netting alignments

After the completion of the interpolation stage of LASTZ, an extensive set of whole genome alignments had been constructed, but two further steps, chaining and netting, were also applied. LASTZ alignments may be short and, more importantly, the target and query bases may have been aligned many times. Whilst we would not necessarily expect an exact one-to-one correspondence between the query and target sequences because of the occasional duplication of sequence, the majority of bases that were aligned multiple times at this stage were likely to represent spurious alignments. This was because of the necessary approach adopted in which each chromosome of the target sequence was aligned against the entire genome sequence of the query sequence. The purpose of chaining was to link together existing alignments to create longer contiguous alignments, and that of netting was to reduce the redundancy in the alignments by creating a single-coverage alignment for the target (but not query) sequence. The chaining and netting tools were created by Jim Kent and colleagues at UCSC (Kent et al. 2003). Chains, which are ordered sets of alignment blocks separated by larger gaps, were initially constructed with the axtChain utility, which takes a k-dimensional tree algorithmic approach to create the chains (Zhang et al. 1994). The chains were then combined across the different chromosomes using the chainMergeSort utility. Chains were then converted into nets with the chainPreNet and chainNet programmes. Nets are collections of hierarchical non-overlapping chains with the best scoring chains stacked up first and the unaligned spaces beneath filled in with lower scoring chains. The end of netting results in the genome-wide alignments between the two genome sequences in the form that they can be downloaded from UCSC. Before identifying constrained sequence, I processed these genome alignments with a further trimming step to remove poor quality aligned sequence as described in **Chapter 3**.

2.3 Neutral Indel Model 1

2.3.1 The distribution of inter-gap segments

The Neutral Indel Model 1 (NIM1) was originally coded by Gerton Lunter in 2005, published by Lunter et al. (2006), and applied extensively by Meader et al. (2010). I will describe in **Chapter 3** some additional novel details and refinements of this model. The NIM1 estimates the quantity of sequence that was constrained with respect to insertions and deletions (indels) and shared between a pair of sequences, a quantity I term $\alpha_{selIndel}$. The NIM1 identifies inter-gap segments (IGSs), which are un-gapped alignment blocks, from whole genome pairwise alignments. IGSs are inferred as being examples of contiguous sequence where an indel has not arisen in the sequence since the divergence of the two sequences from their ancestral sequence (**Figure 2.2**). The distribution of IGSs calculated from the alignments is compared to the distribution expected under neutral evolution.

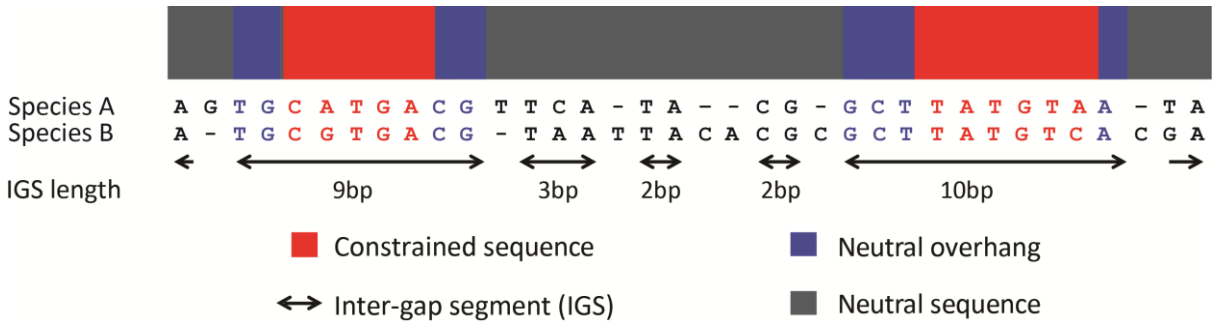

Figure 2.2: Illustration of how IGSs and constrained elements are inferred from a genome alignment. The segment consists of constrained (red blocks) and neutral (blue or grey blocks) segments. Indels falling into the constrained regions are purged by purifying selection, while those falling in neutral sequence accumulate over time. The 'neutral overhang' (blue blocks) represents sequence that lies between adjacent indels that contain a constrained element. Underneath the segment is an alignment with gaps where the indels have fallen. From genome-wide alignments the distribution of IGSs can be deduced, and $\alpha_{selIndel}$ between the species pair can be estimated.

Under neutral evolution, the IGS lengths of intermediate sizes can be fitted to a geometric distribution. The geometric distribution is a particular case of the negative binomial distribution where the number of 'successes' required is one. A success in this case corresponds to the occurrence of a gapped alignment column. Under this distribution the probability of an indel occurring in the genome sequence is assumed to be equal at all positions, and it is assumed that each indel event is independent. The probability of an IGS of 1bp long in neutral sequence is simply 1 - *P* where *P* is the (neutral) indel rate. To generalise, the probability of an IGS of length Nbp arising in neutral sequence is $(1 - P)^N$.

This geometric distribution is fitted to the distribution of IGSs over a particular size range inferred from the alignment. The size range of IGSs over which the geometric distribution is fitted must be small enough to avoid significant contamination from constrained sequence. For relatively small IGS values the quantity of neutral sequence is much larger than the quantity that is subject to selection. This size range must also be large enough to avoid the technical issues of gap attraction and gap annihilation. Gap attraction occurs when multiple gaps are called as a single gap by the alignment algorithm, and gap annihilation is when multiple gaps cancel out and so the resulting alignment is wrongly identified as being contiguous by the alignment algorithm (Lunter 2007; **Figure 2.3**).

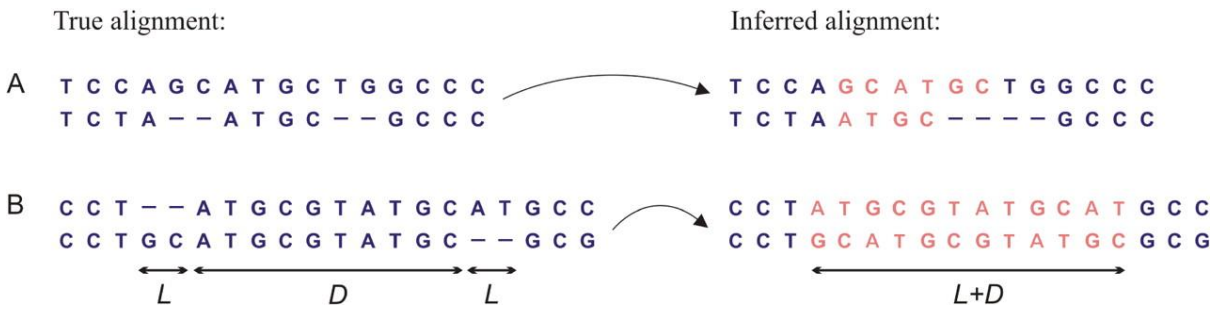

Figure 2.3: Examples of the effects of gap attraction A., where a single gap is incorrectly called by the alignment algorithm instead of two gaps, and gap annihilation B., where no gap is called instead of two gaps. The figure was taken from Lunter (2007). Lunter G, Probabilistic whole-genome alignments reveal high indel rates in the human and mouse genomes, 2007, *Bioinformatics* 23(13): i289-296, by permission of Oxford University Press.

2.3.2 Estimation of $\alpha_{selIndel}$

While the distribution of IGSs in neutral sequence is expected to conform to the geometric distribution, the data from whole genome pairwise alignments show more long IGSs than expected under neutral evolution (**Figure 2.4**). This excess of long IGSs is indicative of constrained sequence.

$\alpha_{selIndel}$ between the two species can then be estimated by calculating the cumulative lengths of all overrepresented IGSs, and then subtracting a correction factor to account for the contribution of neutral sequence to each IGS. Each constrained element within the IGS is expected to be flanked each side by *K* bases of neutral sequence before the next indel is reached (**Figure 2.2**), where *K* is the mean number of bases between indels in neutral sequence, and $K = p^{-1}$, and *p* is the indel mutation rate. The model subtracts this 'neutral overhang' sequence to estimate $\alpha_{selIndel}$ between the pair of genome sequences from the length of each IGS minus *2K* summed over all overrepresented IGSs.

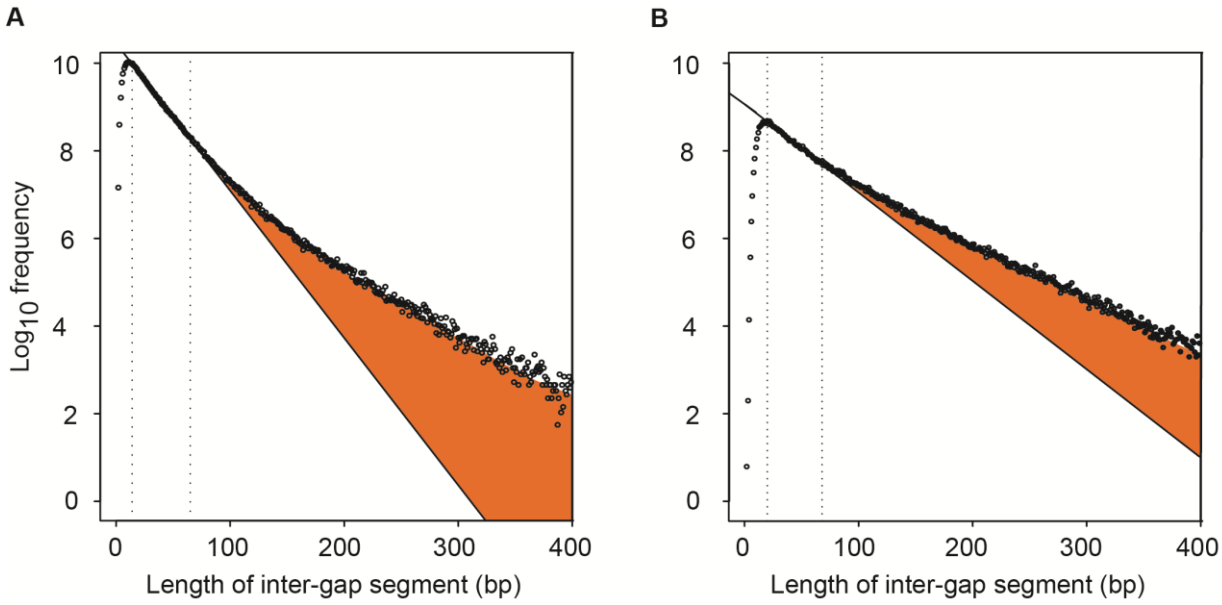

Figure 2.4: Inter-gap segment histograms for A. chicken (galGal3) – Darwin's finch (gm_v1) and B. zebra finch (taeGut1) – Darwin's finch (gm_v1) genome alignments. A representative GC-bin is shown in each case. The estimate of $\alpha_{selIndel}$ for each alignment is the orange area minus the predicted length of the 'neutral overhang' sequence.

2.3.3 Accounting for heterogeneity in the neutral indel rate

During the analysis, the genome was partitioned into 20 bins according to base composition in order to partially account for indel rate mutational biases that covary with G+C content. Generally, the quantity of constrained sequence in a G+C bin is positively correlated with G+C content (Lunter et al. 2006). This relationship can be partially explained by the fact that protein-coding regions tend to be both GC-rich and highly conserved. However, the relationship between G+C content and constrained sequence is very complex, since G+C content correlates with a large number of genomic properties such as the mutation rate, gene density, recombination rate, gene expression levels, and protein secondary structures (Mouchiroud et al. 1991; Fullerton et al. 2001; D'Onofrio et al. 2002; Hardison et al. 2003). The X chromosome is analysed separately from the other chromosomes because it accumulates indels at a lower rate than the autosomes (Makova et al. 2004).

2.4 Estimating the fraction of constrained bases in subsets of the genome

I have described above how NIM1 is used to estimate the amout $\alpha_{selIndel}$ of constrained bases within a genome G consisting largely of neutrally evolving sequence. To estimate $\alpha_{selIndel}$ within a subset $S \subseteq G$ representing a particular functional element type, for instance when estimating $\alpha_{selIndel}$ within coding sequence, I estimated $\alpha_{selIndel}$ within the subsets G and G\S; the difference between the resulting estimates is then the estimate of $\alpha_{selIndel}$ within S. This approach was in preference to the direct estimation of the quantity of constrained sequence in S because many functional element categories, such as coding sequence, are likely to contain only a small fraction of neutrally evolving sequence. Without substantial amounts of neutral sequence the model is unable to calibrate to a background geometric distribution and so is unable to (accurately) estimate $\alpha_{selIndel}$.

2.5 Estimating the rate of neutral evolution

2.5.1 Estimating substitution rates in ancestral repeats

Nucleotide substitution rates were estimated for putatively neutrally evolving sequence. It is imperative to define a proxy for the neutral rate of evolution in order to provide a suitable measure of divergence between two species. As such a proxy I made use of ancestral repeat (AR) sequences since they virtually all show the patterns of indel mutation expected under neutral evolution (Lunter et al. 2006). ARs are TE-derived sequences that are present in the common ancestral sequence of the different species' sequences being examined; in other words they are aligned TE sequences. There is also evidence that ARs may be a better candidate for neutrally evolving sequence than are 4-fold degenerate sites (protein coding sites where any mutation will result in a codon that encodes the same amino acid), at least for avian genome sequences (Kunstner et al. 2011b). I extracted alignments of ARs from the trimmed whole genome alignments using RepeatMasker annotations to identify repetitive elements (Smit et al. 1996-2010). I then calculated the pairwise sequence divergence for the

alignments as the substitution rate estimated using the HKY85 model applied in the PAML package BASEML (Yang 2007).

The HKY85 model (Hasegawa et al. 1985) uses Markov chains to model nucleotide evolution, combining features of the earlier JC69, K80 and F81 substitution models (Jukes and Cantor 1969; Kimura 1980; Felsenstein and Churchill 1996). The Markov process assumes that the probability of mutation at a particular locus is independent of previous mutations that have occurred, and so is only dependent on the current state. Sequence identity could be used as an appropriate proxy of divergence for very closely related sequences, but probabilistic modelling is necessary for even slightly divergent sequences to account for multiple hits, where a single site is mutated several times, and also parallel, convergent, or back mutations (**Figure 2.5**). This is achieved using transition probabilities that explore all possible ways by which evolution could have proceeded (Yang 2006).

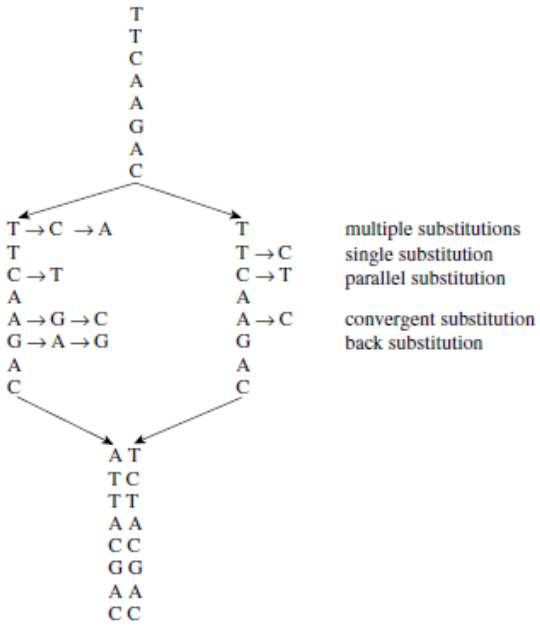

Figure 2.5: Illustration of how multiple-hits, parallel mutations, convergent evolution, and back-mutation may occur over the course of evolution. The figure was taken from Yang (2006).

In addition to this probabilistic modelling that was the basis of the JC60 model, the HKY85 model permits a variable transition to transversion substitution ratio by incorporating an additional parameter into the model, as in the K80 model. This improves the model since transitions are generally 2–3 times more likely to occur than transversions. Finally, the HKY85 model also allows for unequal base frequencies, as in the F81 model. This means that the mutation rate need not be symmetric: for example, A → C may not equal C → A. This is essential to realistically model evolution in biological sequences since base composition varies widely between different (mammalian) species (Romiguier et al. 2010). The HKY85 model can be summarised in a substitution-rate matrix (Haswgawa et al. 1985; Yang 2006; **Figure 2.6**).

	A	C	G	T
A	-	$\beta\pi_C$	$\alpha\pi_G$	$\beta\pi_T$
C	$\beta\pi_A$	-	$\beta\pi_G$	$\alpha\pi_T$
G	$\alpha\pi_A$	$\beta\pi_C$	-	$\beta\pi_T$
T	$\beta\pi_A$	$\alpha\pi_C$	$\beta\pi_T$	-

Figure 2.6: Substitution-rate matrix for the HKY85 model (Hasegawa et al. 1985), where α is the transition rate, β is the transversion rate, and π is the frequency of each nucleotide.

2.5.2 Calculating synonymous substitution rates

I also examined synonymous substitution rates (dS) across protein coding regions for selected species pairs. dS is another proxy for the rate of neutral evolution, since synonymous substitutions are mutations within protein coding regions that do not change the amino acid produced. This occurs because there is redundancy in the genetic code: there are 64 different codons, but only 20 different amino acids produced, hence multiple codons can produce the same amino acid. However, there is increasing evidence that dS may not represent a suitable neutral proxy for at least some vertebrate genomes (Kunstner et al. 2011b).

Synonymous substitutions can be subject to selection due to synonymous codon usage bias, where certain codons are favoured over others that produce the same amino acid due to translational efficiency, although the strongest evidence for this occurrence is in bacterial and yeast genomes (Sharp and Li 1987). Additionally, when regulatory elements overlap coding sequence synonymous sites can be conserved (Lin et al. 2011). Estimates of dS for a mammalian species pair were made by calculating the median dS of all one-one gene orthologues in the Ensembl Compara database. Genes with dS greater than one were excluded from the calculation as estimates of dS are not likely to be accurate in such cases due to multiple hits causing the saturation of sites. An alternative approach for determining a cut-off above which to exclude genes would be to examine the patterns of mutation in the coding sequences. Sequences with high dS estimates cease to show a linear relationship when the transition/transversion ratio at 4-fold degenerate sites is plotted against the dS (Axelsson et al. 2008). For the Darwin's finch (*Geospiza magnirostris*) and other avian species, dS was estimated directly from alignments as described in **Chapter 6**.

2.6 Modelling the turnover of constrained sequence

To help describe and interpret the turnover of constrained sequence I used a time-homogeneous birth-death model for sequence turnover on a genomic scale that was developed by Gerton Lunter. I applied this model to specific sequence classes, such as protein coding genes or transcription factor binding sites (TFBSs), allowing me to relate the rates of turnover for particular types of functional element. The model makes the following assumptions: for a particular class of functional elements, both the total amount of functional sequence and the rate of turnover are constant over time, and the turnover rate (weighted by the length of the elements) is identical for all elements in the class. Specifically, within a class of functional sites comprising a nucleotides, in a small time interval dt a number bdt of sites dispense with function, while an identical number gain function. Thus, while the total

amount of functional nucleotides in the class remains constant over time, the amount that is currently functional and retains homology to functional nucleotides in the ancestral species at divergence *d* (meaning the amount that was constrained and has not turned over during this time) is $a \exp(-b\ d)$. Note that to arrive at this result I make an "infinite sites" assumption, namely that the genome can be considered infinitely large compared to *a*; otherwise one would need to account for reversions back to functionality of neutral but previously functional material. Such reversions appear unlikely under an indel model where *a priori* back-mutation is expected to be rare compared to under a substitution model.

The turnover half life, the divergence at which half the functional sequences in the class have turned over, is calculated as $\log_e(2) / b$. Fitting the data to this model under the assumption of independent normally distributed errors in the observations provides estimates and error bounds on parameters *a* and *b*. This model is time reversible, so that the same formulas hold for the amount of mutually constrained sequence between two extant species at divergence *d*, where *d* is calculated by adding the divergences along the two branches. I express the divergence in time units corresponding to one expected nucleotide substitution per site in neutrally evolving sequence.

2.7 Other bioinformatic tools, computational infrastructure, and software

During my DPhil I used a wide variety of different computational resources. I will very briefly acknowledge here the main suites of bioinformatic tools, computational hardware, and software that I used frequently but do not mention elsewhere. I used the bedTools suite to manipulate genomic coordinates and interval data (Quinlan and Hall 2010). These manipulations include merging overlapping genomic coordinates and adding or subtracting coordinates that represent particular genomic features (such as extracting TFBSs that do not intersect with protein coding sequences). I applied a number of UCSC utilities for transforming sequence data, particularly the tools for converting aligned sequences between

different formats such as the AXT pairwise alignment format and multiple alignment format (Kent et al. 2003). I also used a wide variety of scripts released in the CGAT Code repository, predominantly sequence manipulation tools written by Andreas Heger (Sims et al. 2014).

I worked on a local computer running GNU/Linux operating system, and when necessary accessed Windows applications via the Unit's Windows server. I ran my programmes on the computational infrastructure provided for Chris Ponting's group, which consists of a number of high memory servers and a cluster with 61 nodes, and kept data on the Isilon and near-line storage systems. I wrote the vast majority of my scripts in the Python programming language, with the Ruffus library to facilitate the creation of pipelines (Goodstadt 2010), and many other available packages and libraries for particular tasks. I also used the R statistical package and some shell scripting. All statistics and graphs were calculated or generated in either R or Microsoft Excel. I used the postGreSQL database software to store selected data sets, and a Zope software interface to view the output of some data analyses. I backed up scripts with the subversion software, and I used the eclipse and vim text editors to edit code. Figures were assembled with the graphical editor Adobe Illustrator or Microsoft PowerPoint.

Chapter 3: Improved methods for estimating the quantity of sequence constrained between two genomes

3.1 Abstract

With the sequencing of the human and mouse genomes it became possible to examine patterns of selection genome-wide in mammals using comparative genomic approaches. This led to the development of a number of methods for identifying genomic regions that have evolved under selective constraint, meaning that they have evolved more slowly than the background rate of neutral evolution due to the purging of deleterious alleles. Identifying constrained genomic regions is of interest since such sequences are likely to remain of functional importance and are highly enriched for biological elements, most conspicuously, but not restricted to, protein coding sequences. However, identifying constrained sequence is non-trivial due to both biological difficulties (such as defining an appropriate neutral standard from which to calibrate sequence constraint) and technical issues (such as alignment errors that artefactually inflate the apparent rate of evolution). Here I present improved methods for estimating the quantity of sequence that is mutually constrained with respect to insertion and deletion mutations between a pair of genome sequences, a quantity termed $\alpha_{selIndel}$. The work builds on two previous studies that developed a Neutral Indel Model (termed NIM1) for estimating $\alpha_{selIndel}$. I introduce an alignment processing criterion that I demonstrate improves the robustness of $\alpha_{selIndel}$ estimates. I describe refinements that were made for the application of NIM1, and introduce a second complementary Neutral Indel Model (NIM2) that produces concordant estimates of $\alpha_{selIndel}$. My estimates of $\alpha_{selIndel}$ were consistent in the presence or absence of either non-reciprocally aligned sequence or indel hotpot regions; and I show through extensive genome simulations that the two models generate reliable estimates of $\alpha_{selIndel}$ under a wide range of biological parameterisations.

3.2 Introduction

Evolutionary studies often equate functionality with signatures of long-term purifying selection. While it is undisputed that functional regions have evolved under complex selective regimes including selective sweeps (Bersaglieri et al. 2004) or ongoing balancing selection (Allison 1956; Takahata et al. 1992), and it appears likely that loci exist where recent positive selection or reduction of constraint has decoupled deep evolutionary patterns from present functional status (Pollard et al. 2006; McLean et al. 2011), it is also widely accepted that purifying selection persisting over long evolutionary times has been the prevailing mode of evolution (Ureta-Vidal et al. 2003; Siepel et al. 2005). From this it appears reasonable to make the assumption that at any given moment, the vast majority of nucleotides contributing to an organism's phenotype are under purifying selection. While acknowledging the caveats, this justifies the operational definition of functional nucleotides used here, as those that are subject to purifying selection. Sequences can be identified as subject to purifying selection if they are constrained, meaning that they evolve significantly more slowly than the rate of neutral evolution. Such selectively purified sequences are enriched for biological elements including protein coding regions, small RNAs, and untranslated regions (Abecasis et al. 2012). A large number of different methods have been developed to estimate the amount of constrained sequence between two or more genome sequences (Ponting and Hardison 2011; Ward and Kellis 2012), a quantity I refer to as α_{sel}.

The approach of Lunter et al. (2006) and Meader et al. (2010) for estimating α_{sel} examines sequence constraint with respect to insertions and deletions (indels) rather than using signatures left by point mutations, which is an important distinction of this method from others (Chiaromonte et al. 2003; Margulies et al. 2003; Thomas et al. 2003; Smith et al. 2004; Cooper et al. 2005; Lindblad-Toh et al. 2005; Siepel et al. 2005; Asthana et al. 2007; Margulies et al. 2007; Garber et al. 2009; Parker et al. 2009; Davydov et al. 2010; Eory et al.

2010; Pollard et al. 2010; Lindblad-Toh et al. 2011; Ward and Kellis 2012). I term the model underlying this approach the Neutral Indel Model 1 (NIM1). The NIM1 is a quantitative model describing the distribution of distances between neighbouring indels (intergap segments; IGSs) in neutrally evolving sequence, and it provides an excellent description of the observed frequency of medium-sized IGSs (Lunter et al. 2006). However, across whole genomes longer IGSs are strikingly overrepresented compared to this expectation under neutrality, presumably as a result of the presence of functional genomic segments under purifying selection in which indel mutations are unlikely to become fixed. By quantifying this overrepresentation it is possible to estimate $\alpha_{selIndel}$, the quantity of nucleotides contained with respect to indels within these functional segments. The model, which seeks to account for G+C content and sex chromosome-dependent mutational biases, performs well for simulated data, and accurately identifies ancestral repeats (ARs; aligned transposable element sequences) as neutrally evolving (Lunter et al. 2006; Meader et al. 2010).

Purifying selection is stronger with respect to indel than single nucleotide variants in genomic regions such as coding sequences and 3' UTRs (Montgomery et al. 2013), implying that indel models may provide greater power than substitution approaches for identifying functional genomic sequences. Additionally, there are sequences where the particular nucleotide base present is unimportant, yet the presence of a base is important to facilitate the occurrence of a process. For example, substitutions at 4-fold degenerate sites in protein coding sequences do not affect the amino acid produced, so although these sites are no less crucial for production of the protein than other amino acid sites, they are evolving under weaker selection. This redundancy also occurs in non-coding sequences, such as when intron lengths are under selective pressure to optimise the efficiency of splicing (Parsch 2003). This limitation of point mutation approaches is overcome by methods that infer sequence constraint with

respect to indels, since for indel models the lengths of the ungapped segments are important, rather than the bases present at each site.

Point mutation models do have an advantage over indel approaches in the sense that single nucleotide changes occur with approximately 8–14 times the frequency of indel events (Lunter 2007), so point mutation models can provide greater power and potentially base-pair resolution of constraint. However, this base-pair resolution is not necessary to detect constrained elements, and requires multiple sequence alignments across a broad range of species, so neglects lineage-specific constraint. Ultimately, the best approach to identify constrained sequence would be one that uses information from both point mutations and indels, and such approaches are being developed (Satija et al. 2010). However, in these approaches it is somewhat arbitrary how the different information is weighted and they are not yet fully computationally scalable for very large data sets (Satija et al. 2010). Consequently, in the absence of such fully integrated methods, indel models offer several advantages, or at least differences, compared to models based on point mutations.

However, the methodology of Lunter et al. (2006) was found to be incomplete in several ways, and I present here four improvements for the estimation of $\alpha_{selIndel}$ building on previous implementations (Lunter et al. 2006; Meader et al. 2010). First, I describe two issues that improve the original derivation of the NIM1 model, but find that when considered together their contributions to $\alpha_{selIndel}$ estimates cancel, thereby not invalidating the original results. I present an alternative derivation of the model that properly accounts for these issues. Second, I introduce a new likelihood Neutral Indel Model (NIM2) that estimates $\alpha_{selIndel}$ by directly fitting a probabilistic model to the observations. Application of NIM2 provides a partially independent validation of the revised NIM1 estimates. Third, I conducted a comprehensive simulation study that underscores the validity, accuracy, and robustness of the NIM1 and the NIM2. Based on these results, I calculate a narrower and more reliable confidence interval

for the NIM1 $\alpha_{selIndel}$ estimates than obtained previously. Finally, I find that earlier $\alpha_{selIndel}$ estimates were upwardly biased as a consequence of poor quality alignments, and I show that a likelihood approach to discarding poor quality aligned sequence substantially improves the robustness of $\alpha_{selIndel}$ estimates to alignment artefacts and variation in genome assembly quality.

3.3 Materials and methods

3.3.1 An updated Neutral Indel Model 1 (NIM1)

The Neutral Indel Model (NIM1), originally of Lunter et al. (2006), estimates $\alpha_{selIndel}$ between a pair of species' genome sequences. As described in **Chapter 2**, the model examines the distribution of IGSs from a set of whole genome pairwise alignments using a regression approach over a range of medium IGS lengths to estimate the parameters of a predicted geometric distribution of IGSs in neutral sequence. $\alpha_{selIndel}$ is then estimated as x - 2K bp summed over all the long IGSs inferred to be in excess above that predicted under neutral evolution, where x is the length of the overrepresented IGS, and K is the estimated mean spacing between indels in neutral sequence. 20 equally populated G+C content bins are each analysed separately, as is the X chromosome, to account, in part, for mutational variation.

I describe here two novel theoretical features of the model and an additional approximation, none of which were previously identified by either Lunter et al. (2006) or Meader et al. (2010). These features are: (A) that thresholding can bias the expected lengths of the neutral overhang and, (B) that neutral segments are depleted from the background distribution due to the presence of constrained segments; I will now explain these features, and show that their contributions largely cancel out. The distance between successive indels (inter-gap segments, IGSs) in neutrally evolving sequence approximates a geometric distribution (Lunter et al. 2006). Note that this holds true even in the presence of indel hotspots (Montgomery et al. 2013), provided that the hotspot locations are themselves generated by a uniform random process. The mean IGS length in neutral sequence was denoted by K. The NIM1 consists of

fitting a geometric model $h(x)$ to the observed IGS histogram $H(x)$, over the range of x that is dominated by neutral sequence; here x is the IGS length. Consider a functional segment consisting of c bases that does not accept indels. This segment contributes an IGS of expected length $c + 2K$, whose greater length is because of "neutral overhang" from either end of the segment to the nearest indel in neutral sequence; the expected distance to this first indel is K nucleotides on either side. The estimator α_{selIndel} for the total amount of constrained sequence is $[x - 2K][H(x) - h(x)]$, summed over all x.

Feature (A) arises because in practice, in order to avoid the effects of gap attraction (Lunter et al. 2008), only contributions for $x > T$ for some threshold T are included in this sum. This causes an upward bias to the expected neutral overhang, making the $2K$ correction too small, and resulting in an inflated estimate of α_{selIndel}. Feature (B) arises because of the presence of constrained segments in the genome which skew the distribution of the remaining purely neutral segments. Imagine placing constrained segments randomly into an otherwise neutrally evolving genome with indels already located within them; because the segments are placed randomly, longer IGSs are more likely to be the recipients of such sequence than shorter ones, thereby depleting the set of neutral IGSs of longer segments. Since not only IGSs containing constrained sequence but also neutral IGSs contribute to $H(x)$, this results in a lower estimate of α_{selIndel}. The effect of (A) and (B) when taken together is that the estimator α_{selIndel} becomes conservative. To see this, consider an IGS of length $x + c$ of which c nucleotides are conserved, assume that $h(x)$ is correctly estimated, and that the threshold $T \leq 2K$. Three cases are to be considered. (i) If $x \geq T$, the segment contributes $x + c - 2K$ to α_{selIndel}. Due to effect (B) a segment x that would have contributed $x - 2K$ is now missing from $H(x)$, reducing α_{selIndel} by $x - 2K$. The total contribution is $(x + c - 2K) - (x - 2K) = c$, as desired. (ii) If $T - c \leq x < T$, the segment x missed due to effect (B) is not included in the estimate (since $x < T$), so the contribution is $x + c - 2K$ which is less than c since $x < T \leq 2K$.

(iii) If $x < T - c$, the segment makes no contribution. In no case is the contribution of the constrained segment overestimated, and hence $\alpha_{selIndel}$ underestimates the amount of constrained sequence.

These two new features of the model are the conclusions from an extensive discussion of the theoretical aspects of the model led by Gerton Lunter and Phil Green (University of Washington), with involvement from myself, Chris Ponting, and Stephen Meader. This dissection of the model also revealed an approximation that was not previously noted. *K* is not known *a priori*, but is estimated from the data. If there are short ($x < T$) unaccounted for unclustered functional elements across the genome, these elements would remove indels randomly across the genome. This leads to a distribution of IGSs that can be modeled by a geometric distribution, but with a reduced inferred indel rate and therefore an increased *K*. Overestimating *K* increases the subtracted correction *2K* factor for the neutral overhang hence lowering the estimate of $\alpha_{selIndel}$. However, it is also possible that *K* could be underestimated. There will still inevitably be residual variation in the neutral indel rate that is not accounted for by the partitioning of the autosomal genome based on G+C content, and this will tend to lead to an overestimation of the indel rate. This in turn leads to an underestimation of *K* and thus an overestimation of $\alpha_{selIndel}$. Another way of understanding this is that genomic regions with a lower neutral indel rate will sometimes be falsely called as constrained rather than neutrally evolved. This highlights that a critical dimension to explore further in the genome simulations is the effect of a heterogeneous neutral indel rate on estimates of $\alpha_{selIndel}$.

My implementation of the NIM1 differs from that of the preceding studies in the manner in which I calculate the bounds of the estimates. The previous approaches constructed the upper and lower bound estimates based on the uncertainty in the degree of clustering of functional elements. The lower bound estimate was derived assuming that functional elements are

unclustered (each overrepresented IGS contributes $x - 2K$ bp towards the $\alpha_{selIndel}$ estimate), while the upper bound was derived assuming a high degree of clustering (each overrepresented IGS contributes $x - K$ bp). In my revised approach, I construct a 95% confidence interval around the lower $x - 2K$ bp estimate. The previous upper bound estimate assumes an unrealistically high degree of clustering of functional elements. Furthermore, by providing a 95% confidence interval for the $\alpha_{selIndel}$ estimate of NIM1, the estimate is directly comparable to the estimates of the second neutral indel model that I now describe.

3.3.2 Neutral Indel Model 2 (NIM2)

I introduce an alternative Neutral Indel Model (NIM2) that estimates $\alpha_{selIndel}$ using a maximum likelihood approach. The model was conceived by Gerton Lunter and Phil Green, and coded by Gerton Lunter. The model considers that indel mutations fall randomly across a wholly neutrally evolving genome into which conserved indel-free elements, representing functional sequence, are subsequently inserted. Although this construction does not reflect the actual series of evolutionary events, the resulting distribution of indels and conserved elements within neutral sequence is analogous, and the mathematical derivation is simplified. From this starting point, the distribution of medium-length and long IGSs is modeled, and $\alpha_{selIndel}$ is estimated by subtracting the cumulative length of the observed segments from the estimated length of the underlying neutral segments.

To account for gap attraction, the NIM2 describes the observed IGS counts above a lower length threshold i_{min}. Its parameters are the total segment count N, the neutral average inter-gap distance K, and the per-base probability of insertion of functional sequence p. N is fixed to be the observed segment count, and as with NIM1 i_{min} and i_{max} are fixed describing the range $[i_{min}, i_{max}]$ in which neutrally evolving sequence dominates, using the same range I used for NIM1.

Before introducing functional sequence, the expected number of IGSs of length i is $N_i =$

$\frac{N}{K}(1-\frac{1}{K})^{i-i_{\min}}$. An IGS of length *i* has probability *p*(*i*+1) of being the recipient of a functional segment, reducing the expected count to $\widetilde{N}_i = \frac{N}{K}\left(1-\frac{1}{K}\right)^{i-i_{\min}}(1-p)^{i+1}$; later for technical convenience the approximation 1-*p*(*i*+1) ≈ (1-*p*)$^{i+1}$ is used, valid as long as $p \ll \frac{1}{i}$. The expected number of segments of length above $i_{\max}$, $\widetilde{N}_{\text{above}}$, consists of contributions from the $\widetilde{N}_i$ in that range, and from neutral segments in the range [$i_{\min}$,$i_{\max}$] into which a functional segment was inserted. The observed counts are then modeled N_i^{obs} ($i \in [i_{\min}, i_{\max}]$) and $N_{\text{above}}^{\text{obs}} = \sum_{i>i_{\max}} N_i^{\text{obs}}$ as Bernoulli-distributed random variables with expectations $\widetilde{N}_i$ and $\widetilde{N}_{\text{above}}$; this choice does not enforce the sum of observed counts to equal *N*, but ignoring the small anti-correlation between observations that result from this constraint makes little difference in practice. Numerical methods implemented in R were used to infer maximum likelihood parameters for the given observed data. Finally, the inferred number of functional nucleotides was computed as the total number of nucleotides covered by observed IGSs, less the predicted total number of nucleotides covered by neutral segments N_i.

Compared with the NIM1, the advantage of the NIM2 model is that features (A) and (B) are modeled explicitly. On the other hand, my implementation of NIM2 does assume a single genome-wide neutral indel rate, and thus does not account for known location-dependent mutational biases that NIM1 partially controls for by binning the genome based on G+C content and by analysing the X chromosome separately.

3.3.3 Simulating genome evolution

Genome simulations were conducted to test the accuracy and robustness of the NIM1 and NIM2. For each simulation, a 200Mb genome was simulated in 5kb blocks with a G+C content distribution matching that of the human genome. 5% of this simulated genome was annotated as constrained, and was evolved at half the rate of the surrounding neutral sequence with respect to substitutions. This simulated genome was then used as ancestral sequence from which to evolve two descendant genomes at a specified divergence that represents the

neutral substitution rate. Substitutions were modelled under the HKY85 model (with transition/transversion ratio of 2), and indel evolution was modelled by sampling indel lengths from a geometric distribution. Gerton Lunter coded the original simulation scripts, which were subsequently modified by myself and Stephen Meader.

All relevant parameters are described in **Table 3.1**. The following parameters were fixed at a single value across all simulations. The functional length was fixed at 158.6bp, reflecting the mean length of GERP++ conserved elements (Davydov et al. 2010). The fixation probability was fixed at a value of 0.1, based on estimates from protein coding sequences (Brandstrom and Ellegren 2007). The substitution/indel ratio was fixed at 12, reflecting the values obtained across my alignments.

There is likely to be variation in the autosomal neutral indel rate due to mutational biases that are not accounted for by binning on G+C content. Consequently, I incorporated a degree of neutral indel rate heterogeneity into the model using a residual indel rate variation parameter (**Table 3.1**). To estimate the parameter value I first masked out all vertebrate PhastCons conserved elements from the trimmed hg19 to mm10 LASTZ alignment and then examined the distribution of indels in concatenated 5kb alignment blocks of the remaining putatively neutral sequence. I find that there are a mean and standard deviation of 164 and 20 indels per 5kb in these masked alignments. If there were to be no rate variation, and the indel frequencies were drawn from Poisson distributions, then the variance in the number of indels per 5kb would be 140 (a standard deviation of approximately 12). I observe a standard deviation of 20 (a variance of 400), so an excess variance of 400-140 = 260. The variance of a uniform distribution across [x-L,x+L] is $L^2 / 3$; setting this equal to 260 gives L=28, which translates to a value for the residual indel rate variation parameter of 28/140 = 0.2.

Table 3.1: The definitions of the parameterisations that were considered across the genome simulations.

Parameter name	Parameter description
Species divergence	Mean proportion of substitutions per site that have changed between the species pair
Functional element clustering coefficient	Proportion of functional segments that are followed by another
Indel fixation probability	Fixation probability of an indel touching a functional segment
Functional expected length	Expected length of functional segment drawn from the gamma distribution
Functional shape	Shape parameter of gamma distribution for functional material
Intervening expected length	Expected length of neutral intervening segment drawn from the gamma distribution
Intervening shape	Shape parameter of gamma distribution for intervening neutral material
Substitution/indel ratio	The number of times higher the substitution rate was compared to the indel rate
Residual indel rate variation	This parameter is a value from 0 to 1 which allows the neutral indel rate to vary per simulated block within a set of boundaries. For example, set to 0.5, for a neutral indel rate of 0.2, there would be a an interval centred on 0.2 (in this case 0.1-0.3) from which the neutral indel rate is picked for each simulated block.

3.3.4 Alignment trimming

Accurate estimation of $\alpha_{selIndel}$ requires high quality whole genome alignments to infer the distribution of IGSs. When available, whole genome pairwise alignments were downloaded from the UCSC Genome Informatics website (http://genome.ucsc.edu). Otherwise, I constructed alignments following UCSC's protocol (Kent et al. 2003). Initial alignments were constructed with LASTZ (http://www.bx.psu.edu/miller_lab/) and these alignments were subsequently chained and netted using tools from UCSC. A detailed explanation of the alignment methodology is given in **Chapter 2**.

I introduce an additional processing step to remove (trim-off) poor quality aligned sequence from the whole genome alignments. I first rescored each alignment to generate a new substitution matrix using a log-odds ratio approach similar to a previous method (Chiaromonte et al. 2002) (**Figure 3.1**). The score of each alignment column x-over-y is

given as the log odds ratio s(x,y) = log (p(x,y) / $q_1(x)q_2$(y)), where p(x,y) is the number of x-over-y alignment columns as a proportion of the total number of gap-free alignment columns, while $q_1(x)$ and q_2(y) are the total observed frequencies of the bases x and y. The substitution scores are transformed so that the largest value is 100 to make the values consistent with gap penalties that were used subsequently.

```
for each gap-free local alignment:
        for each column, x-over-y, of the alignment:
                observe(x,y)
npairs   = n1(A) + n1(C) + n1(G) + n1(T)
for x {A, C, G, T}:
        q1(x)   = n1(x) / npairs
        q2(x)  = n2(x) / npairs
        for y {A, C, G, T}:
                p(x, y) =  m(x, y) / npairs
for x {A, C, G, T}:
        for y {A, C, G, T}:
                s(x, y) =   log ( p(x,y) / q1(x)*q2(y) ) (scale so largest entry is 100)
```

Figure 3.1: Schematic showing how the log odds substitution matrix is calculated. The figure was modified from Chiaromonte et al. (2002). Copyright @ Pacific Symposium on Biocomputing 2002.

The approach is different from that of Chiaromonte et al. (2002) in the respect that I did not impose symmetry on the scoring matrices with respect to strand or species. Strand symmetry (such as AC = TG) is expected, but species symmetry (such as AC = CA) is not necessarily evident *a priori*, since it may not be observed if the G+C content is different between the two species. Consistent with this expectation, data from mouse – rat and human – mouse alignments show stronger strand than species symmetry (**Figure 3.2**).

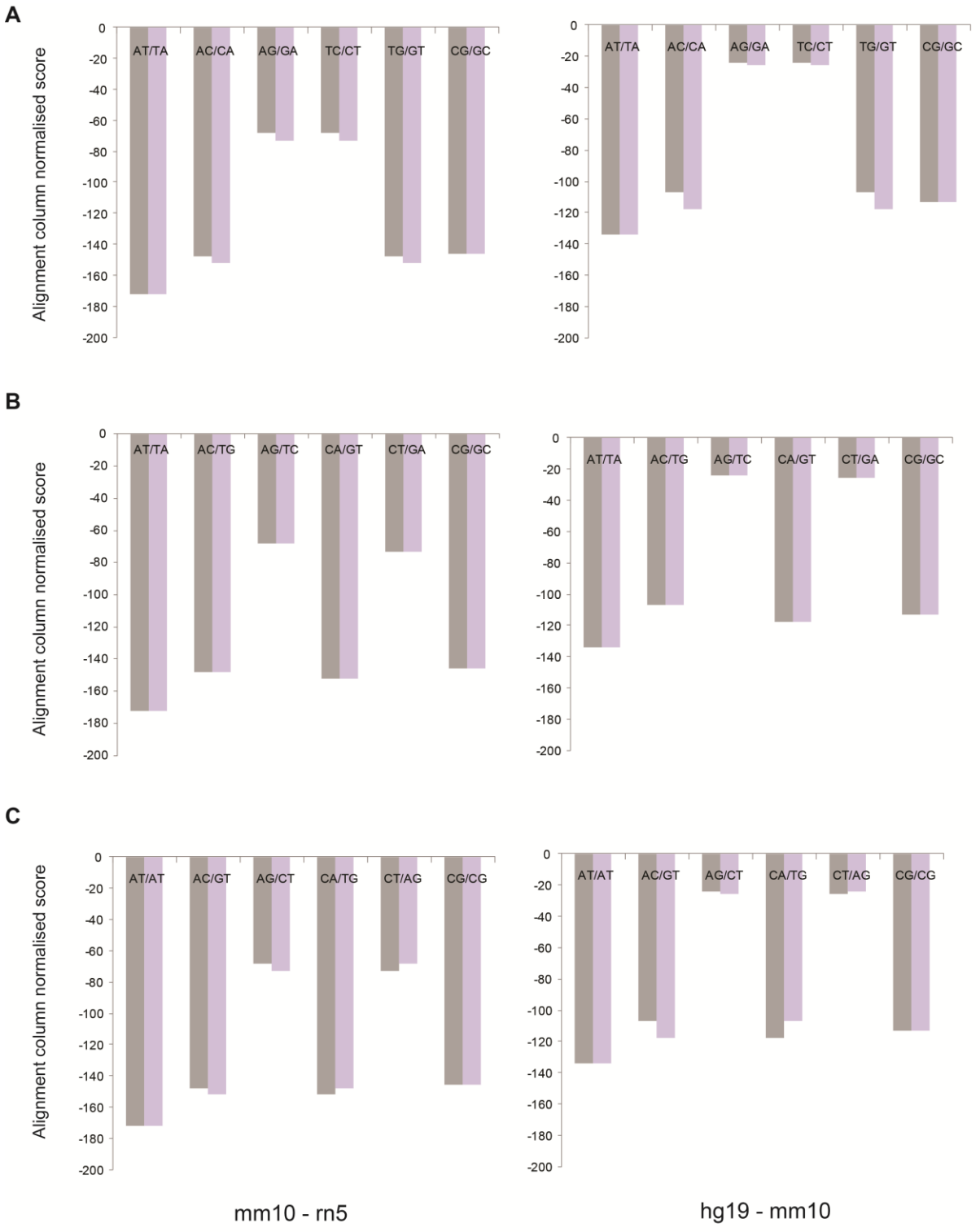

Figure 3.2: The normalised log odds scores of different alignment columns from the trimmed whole genome mouse – rat (mm10 – rn5) and human – mouse (hg19 – mm10) alignments. I compare: A. the species symmetry, B. the strand symmetry, and C. the species and strand symmetry.

Alignment columns that contained 'N' bases (*x*N columns) were scored as 0. Since N bases are proportionally rare in genome sequences the scoring of N has a very small impact on the generated alignments. For example, for mouse – rat alignments (mm10 – rn5), 1739Mb aligned when *x*N bases were scored at 0 and 1741Mb when they were scored at 100. Subsequently alignments of fewer than 50 alignment columns in length were removed, although this removed only small quantities of sequence and had a minimal impact on estimates of $\alpha_{selIndel}$ (**Table 3.2**). The generated substitution matrices for all alignments are displayed in **Table 3.3**.

Table 3.2: Statistics for mouse – rat and human – mouse trimmed alignments before and after removing alignment blocks of fewer than 50 alignment columns in length

Mm10 – rn5 alignments		
	All alignments	Alignments longer than 50 columns
Ungapped length (Mb)	1741	1739
Sequence identity (%)	85.2	85.2
Proportion repetitive	0.284	0.284
G+C content	0.417	0.417
Hg19 – mm10 alignments		
	All alignments	Alignments longer than 50 columns
Ungapped length (Mb)	876	871
Sequence identity (%)	70.7	70.1
Proportion repetitive	0.219	0.219
G+C content	0.407	0.408

Table 3.3: The normalised log odds substitution matrices that were calculated for different eutherian genome alignments. The species identifiers correspond to the genome assemblies used as shown in Table 2.1.

mm10 – rn5						**mm9 – rn4(2)**					
		Base in query						Base in query			
		A	C	G	T			A	C	G	T
Base in target	A	79	-148	-68	-172	Base in target	A	80	-148	-68	-171
	C	-152	99	-146	-73		C	-151	100	-150	-72
	G	-73	-146	100	-152		G	-72	-150	99	-151
	T	-172	-68	-148	79		T	-171	-68	-148	80
mm9 – rn4(1)						**mm8 – rn4**					
		Base in query						Base in query			
		A	C	G	T			A	C	G	T
Base in target	A	79	-148	-67	-172	Base in target	A	80	-150	-68	-173
	C	-152	100	-153	-72		C	-153	100	-154	-72
	G	-72	-153	99	-152		G	-72	-154	100	-152
	T	-172	-67	-148	70		T	-173	-68	-150	80
mm10 – equCab2						**mm9 – canFam2**					
		Base in query						Base in query			
		A	C	G	T			A	C	G	T
Base in target	A	80	-116	-24	-136	Base in target	A	80	-114	-24	-134
	C	-110	100	-112	-28		C	-110	100	-112	-25
	G	-28	-112	99	-110		G	-25	-112	99	-110
	T	-136	-24	-116	80		T	-134	-24	-114	80
mm9 – bosTau7						**hg19 – cerSim1**					
		Base in query						Base in query			
		A	C	G	T			A	C	G	T
Base in target	A	80	-115	-25	-144	Base in target	A	77	-132	-43	-162
	C	-110	100	-114	-26		C	-135	100	-127	-51
	G	-26	-113	99	-110		G	-51	-127	99	-135
	T	-133	-25	-115	80		T	-162	-43	-132	77
hg19 – bosTau7						**hg19 – oryCun2**					
		Base in query						Base in query			
		A	C	G	T			A	C	G	T
Base in target	A	77	-120	-36	-149	Base in target	A	79	-112	-32	-149
	C	-129	100	-121	-39		C	-131	100	-118	-42
	G	-39	-121	99	-129		G	-42	-118	99	-131
	T	-149	-36	-120	77		T	-149	-32	-112	79
hg19 – canFam2						**hg19 – cavPor3**					
		Base in query						Base in query			
		A	C	G	T			A	C	G	T
Base in target	A	77	-120	-35	-153	Base in target	A	77	-112	-30	-142
	C	-131	100	-120	-39		C	-123	100	-113	-29
	G	-39	-120	99	-131		G	-29	-113	99	-123
	T	-153	-35	-120	77		T	-142	-31	-112	77

hg19 – equCab2						**hg19 – mm10**					
		Base in query						Base in query			
		A	C	G	T			A	C	G	T
Base in target	A	78	-127	-40	-161	Base in target	A	79	-107	-24	-134
	C	-135	100	-124	-50		C	-118	100	-113	-26
	G	-50	-123	99	-135		G	-26	-113	99	-118
	T	-161	-40	-127	78		T	-134	-24	-107	78
hg19 – otoGar3						**hg19 – ailMel1**					
		Base in query						Base in query			
		A	C	G	T			A	C	G	T
Base in target	A	78	-124	-42	-157	Base in target	A	78	-120	-37	-157
	C	-135	100	-126	-43		C	-134	100	-122	-45
	G	-43	-126	99	-135		G	-45	-122	99	-134
	T	-157	-42	-124	78		T	-157	-37	-120	78
canFam2 – musFur1						**canFam2 – equCab2**					
		Base in query						Base in query			
		A	C	G	T			A	C	G	T
Base in target	A	78	-142	-63	-180	Base in target	A	78	-138	-45	-169
	C	-151	99	-132	-66		C	-133	99	-126	-51
	G	-66	-132	99	-151		G	-50	-126	100	-133
	T	-180	-63	-142	78		T	-169	-45	-138	78
canFam2 – bosTau7											
		Base in query									
		A	C	G	T						
Base in target	A	78	-128	-40	-156						
	C	-126	99	-123	-40						
	G	-39	-123	100	-126						
	T	-156	-40	-128	78						

I then used the generated substitution matrices, with gap opening and extension penalties derived from the original UCSC alignments, to discard the maximal non-positively scoring terminal segments of the alignment blocks and any non-positively scoring IGSs. Trimming removes terminal and internal alignment segments that are more likely to have arisen under a model of independent evolution than of evolution from a common ancestor. Subsequent analyses were carried out following the discarding of all trimmed sequence (**Table 3.4**). I also excluded alignments that were led by sequence not mapped to chromosomes.

Table 3.4: The quantity of aligned sequence trimmed off for each pairwise eutherian alignment. The trimmed start and end sequences were the quantities of maximally non-positively scoring terminal sequence at the starts and ends of alignment blocks. The trimmed internal sequence was the quantity of non-positively scoring IGSs. Note that the statistics are prior to the removal of alignments that are fewer than 50 alignment columns long.

Species Pair	Quantity of aligned sequence (Mb)			
	Kept	Trimmed start	Trimmed end	Trimmed internal
hg19 – equCab2	1577.4	35.7	36.4	5.8
hg19 – cerSim1	1610.2	34.8	35.6	6.1
hg19 – otoGar3	1498.6	47.2	47.5	7.3
hg19 – canFam2	1445.4	42.8	43.2	5.9
hg19 – ailMel1	1414.5	16.5	17.9	4.7
hg19 – bosTau7	1264.0	47.1	48.0	5.5
hg19 – oryCun2	1185.3	47.7	48.6	4.9
hg19 – cavPor3	1153.2	55.2	55.7	5.3
hg19 – mm10	883.2	71.5	71.8	4.3
mm10 – rn5	1743.0	17.9	18.5	4.7
mm10 – equCab2	821.2	45.0	45.5	4.0
mm10 – canFam2	677.1	50.5	50.9	3.9
mm10 – bosTau7	608.8	44.0	44.3	3.4
canFam2 – mpf_v1	1682.2	16.3	18.1	5.6
canFam2 – equCab2	1610.9	30.7	31.1	6.4
canFam2 – bosTau7	1287.7	45.9	46.4	6.3

3.4 Results

3.4.1 Genome simulations demonstrated the accuracy and robustness of the NIMs

To test the performance of the updated NIM1 and new NIM2, I next set up an extensive series of simulations. While the NIM1's analysis shows that under the stated assumptions the original regression analysis will provide unbiased results, I revisited other potential causes of bias that were not included in the model analysis by simulating data under a more detailed model and over a broader range of parameters than previously explored (Meader et al. 2010). Specifically, I investigated the following parameters: species divergence, functional element clustering coefficient, indel fixation probability, size and shape parameters of the length distributions for both functional elements and intervening neutral sequences, and the residual

indel rate variation; parameter definitions are provided in **Table 3.1**. I deemed it particularly important to test how robust the models are to variations in the residual indel rate because there is heterogeneity in the neutral indel rate that is not captured simply by partitioning the genome based on G+C content (Montgomery et al. 2013).

For each set of parameters, the evolution of two genomes sharing a common ancestral genome was simulated, with a constant 5% of the sequence evolving more slowly due to constraint. To reduce computation time, I simulated 200Mb of sequence and scaled up the results to produce estimates for genomes of 3Gb in size. Where possible, I estimated the parameter values required for the simulations from real data; when appropriate parameter values were difficult to obtain, a range of realistic values was used (**Materials and methods**). Following their simulated evolution, the descendant genomes were aligned using LASTZ and the NIM1 was applied to determine how the inferred amount of constrained sequence compared to the known true amount.

For every one of 160 parameter combinations, the NIM1 $\alpha_{selIndel}$ estimate was conservative, but was never less than 80% of the true value (blue diamonds in **Figure 3.3**, **Table 3.5**). Importantly, the simulations also demonstrate that the model's ability to accurately infer the amount of constrained sequence is not substantially diminished at the extremes of the divergence range, although varying the sequence divergence has a larger impact on the resulting estimate of $\alpha_{selIndel}$ than the other parameters (**Figure 3.4**, **Table 3.5**). Consequently, I conclude that the NIM1 is robust: it does not overestimate $\alpha_{selIndel}$, even over the relatively small evolutionary distances I considered here. Furthermore, any trends in $\alpha_{selIndel}$ exceeding approximately 20% across the range of evolutionary distances cannot be attributed to any of the potential sources of bias I included in my simulations.

The simulations also validate the revised manner in which I calculate the confidence interval for estimating $\alpha_{selIndel}$ with NIM1. The modification I implemented was to construct a 95%

confidence interval around what was previously the lower bound NIM1 estimate (**Materials and methods**). My modified estimate was always conservative under all the simulation scenarios, whereas the previous implementation of the NIM1 sometimes overestimated the true value of $\alpha_{selIndel}$ (purple triangles in **Figure 3.3**, **Table 3.1**). Additionally, altering the clustering of functional elements in the simulations actually had only a minor effect on the estimate of $\alpha_{selIndel}$ (**Figure 3.3**), contrary to the theoretical justification for the original NIM1 upper and lower bounds.

Under the same simulations NIM2 produced relatively robust estimates of $\alpha_{selIndel}$, although it is predicted to overestimate $\alpha_{selIndel}$ at lower divergences and exhibits some diminution of power to detect constrained sequence with increasing divergence (red circles in **Figure 3.3**). Therefore, although NIM2 provides a partially-independent validation for the results of NIM1, NIM2 performs worse than NIM1 in the simulations. This is probably because NIM2 assumes a global neutral indel rate without attempting to correct for biases in the neutral indel rate that are known to correlate with genomic G+C content. I focus on the NIM1 results in subsequent analyses.

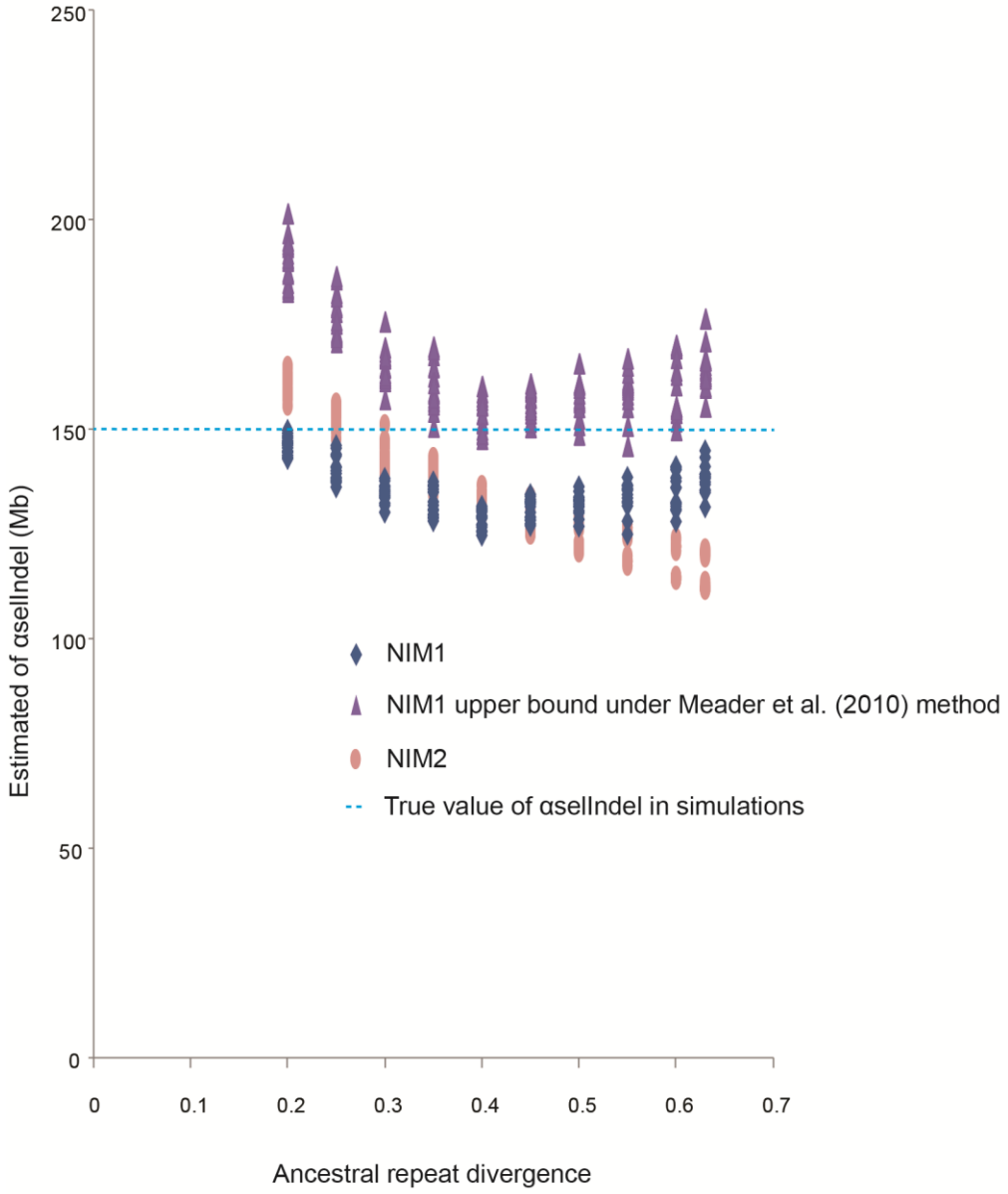

Figure 3.3: Estimates of $\alpha_{selIndel}$ by NIM1 and NIM2 on simulated data under a variety of parameterisations. The true value of $\alpha_{selIndel}$ was fixed at 150Mb across all the simulations (dashed horizontal line).

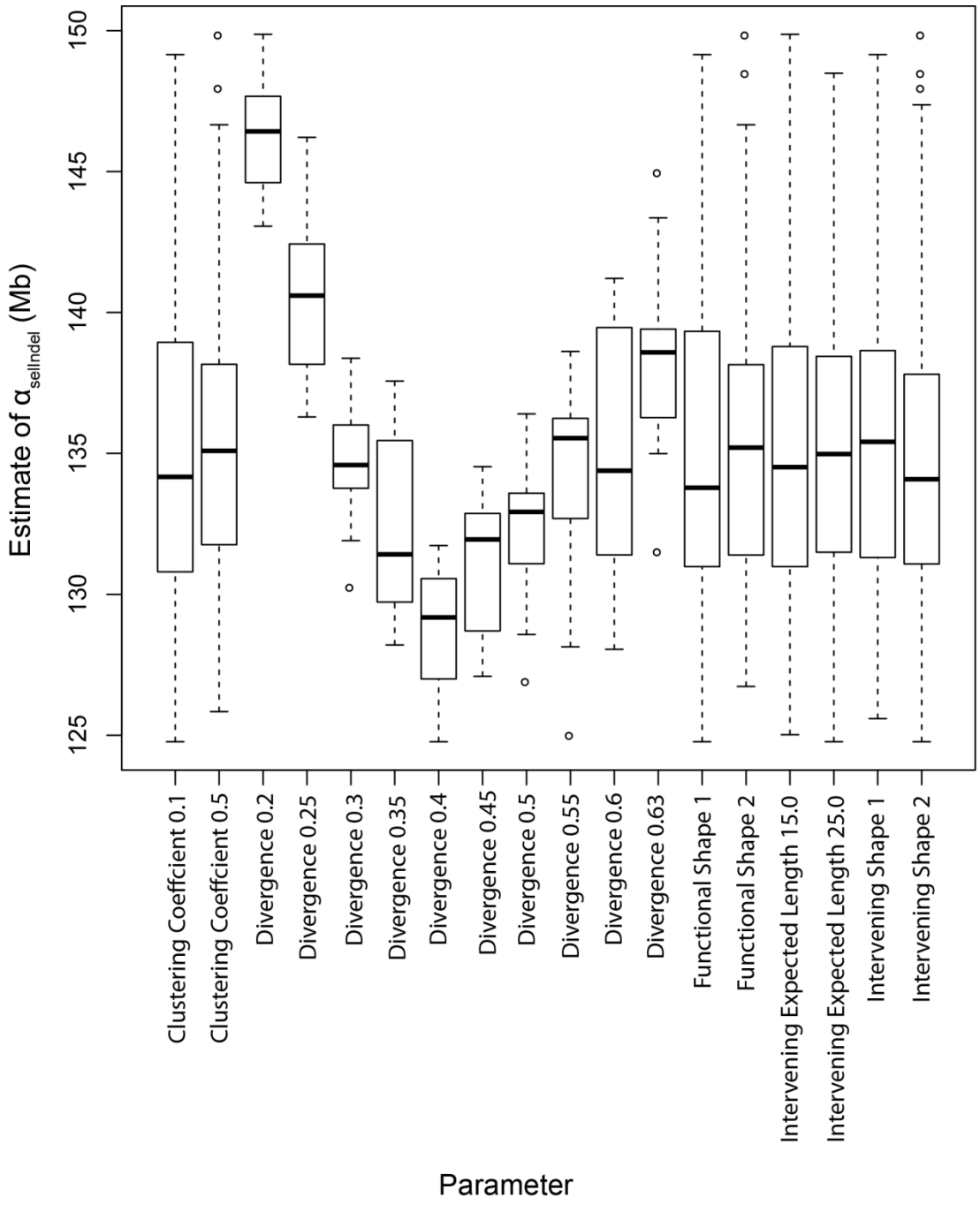

Figure 3.4: Effect of varying different parameterisations on the estimated quantity of constrained sequence by NIM1 in the genome simulations. Each box plot shows the distribution of $\alpha_{selIndel}$ estimates for which that particular parameter was fixed at the given value. The box hinges represent the upper and lower quartiles and the notches extend to +/-1.58 of IQR/sqrt(n) (approximating a 95% confidence interval), where IQR is the interquartile range and n is the number of simulations.

Table 3.5: Estimates of $\alpha_{selIndel}$ by NIM1 on simulated data. My implementation of NIM1 always estimated $\alpha_{selIndel}$ accurately or conservatively, although there was variation in estimates across the different parameterisations. The previous implementation by Meader et al. (2010) can overestimate $\alpha_{selIndel}$.

NIM1 estimate of $\alpha_{selIndel}$ (Mb)	Upper bound estimate of $\alpha_{selIndel}$ (Mb) under Meader et al. (2010) implementation	Parameterisation				
		Divergence	Clustering Coefficient	Functional Shape	Intervening Expected Length (bp)	Intervening Shape
133.9	160.9	0.5	0.5	2	25	2
138.9	164.3	0.63	0.5	2	15	1
135.3	161.3	0.5	0.5	2	15	2
134.5	159.3	0.45	0.5	1	15	2
133.4	160.3	0.45	0.5	2	25	2
132.8	158.1	0.45	0.5	1	25	1
135.9	161.6	0.55	0.5	1	25	2
136.5	162.1	0.55	0.5	2	15	1
135.6	161.0	0.55	0.5	2	15	2
131.8	155.6	0.55	0.5	2	25	1
128.1	151.1	0.55	0.5	2	25	2
133.3	157.2	0.5	0.5	1	15	1
132.7	155.7	0.6	0.5	1	15	2
131.7	154.3	0.6	0.5	1	25	1
132.4	155.2	0.6	0.5	1	25	2
136.1	160.8	0.6	0.5	2	15	1
135.4	160.2	0.63	0.5	2	25	2
128.6	151.3	0.5	0.5	1	25	2
130.1	152.8	0.5	0.5	2	15	1
134.3	165.0	0.3	0.5	2	25	2
131.9	157.2	0.35	0.5	1	15	1
128.2	150.9	0.35	0.5	1	15	2
132.8	159.1	0.35	0.5	1	25	1
133.9	162.5	0.3	0.5	1	25	2
137.8	168.8	0.3	0.5	2	15	1
130.3	157.5	0.3	0.5	2	15	2
136.6	167.9	0.3	0.5	2	25	1
132.6	157.8	0.45	0.5	2	15	2
134.1	161.2	0.45	0.5	2	25	1
128.3	150.8	0.45	0.5	1	25	2
132.9	158.0	0.45	0.5	2	15	1
136.8	165.1	0.35	0.5	1	25	2
129.7	154.5	0.35	0.5	2	15	1
131.7	158.1	0.4	0.5	2	25	2
131.6	155.6	0.45	0.5	1	15	1

NIM1 estimate of $\alpha_{selIndel}$ (Mb)	Upper bound estimate of $\alpha_{selIndel}$ (Mb) under Meader et al. (2010) implementation	Parameterisation				
		Divergence	Clustering Coefficient	Functional Shape	Intervening Expected Length (bp)	Intervening Shape
144.6	183.8	0.2	0.5	1	15	1
144.9	183.1	0.2	0.5	1	25	1
148.0	187.9	0.2	0.5	1	15	2
144.6	185.2	0.2	0.5	2	15	1
143.9	184.2	0.2	0.5	1	25	2
146.7	191.0	0.2	0.5	2	25	1
149.9	195.3	0.2	0.5	2	15	2
141.0	172.7	0.25	0.5	1	15	1
145.7	187.4	0.2	0.5	2	25	2
137.5	163.0	0.63	0.5	2	15	2
132.9	157.1	0.5	0.5	1	25	1
129.3	154.0	0.4	0.5	2	25	1
131.1	152.2	0.6	0.5	1	15	1
139.4	165.6	0.63	0.5	1	25	2
143.4	171.3	0.63	0.5	1	25	1
127.3	150.2	0.4	0.5	2	15	2
138.7	163.5	0.63	0.5	1	15	1
137.7	164.3	0.6	0.5	2	25	2
130.7	154.2	0.6	0.5	2	25	1
132.3	155.7	0.6	0.5	2	15	2
135.5	160.3	0.55	0.5	1	25	1
138.5	164.7	0.63	0.5	2	25	1
136.7	161.5	0.55	0.5	1	15	2
139.5	173.1	0.25	0.5	1	25	1
143.7	178.3	0.25	0.5	1	15	2
145.6	182.8	0.25	0.5	2	15	1
137.6	171.1	0.25	0.5	1	25	2
138.4	173.0	0.25	0.5	2	25	1
146.2	186.1	0.25	0.5	2	15	2
136.3	163.7	0.3	0.5	1	15	1
137.9	173.5	0.25	0.5	2	25	2
135.7	165.2	0.3	0.5	1	25	1
134.4	161.8	0.3	0.5	1	15	2
137.6	168.0	0.35	0.5	2	25	1
135.0	162.9	0.35	0.5	2	15	2
126.7	149.1	0.4	0.5	2	15	1
130.8	156.0	0.4	0.5	1	25	2
125.8	148.0	0.4	0.5	1	25	1
128.8	151.7	0.4	0.5	1	15	2
129.1	152.0	0.4	0.5	1	15	1
126.9	149.0	0.5	0.5	2	25	1
141.2	167.2	0.63	0.5	1	15	2

NIM1 estimate of $\alpha_{selIndel}$ (Mb)	Upper bound estimate of $\alpha_{selIndel}$ (Mb) under Meader et al. (2010) implementation	Parameterisation				
		Divergence	Clustering Coefficient	Functional Shape	Intervening Expected Length (bp)	Intervening Shape
135.1	165.3	0.35	0.5	2	25	2
135.9	160.8	0.55	0.5	1	15	1
133.0	156.7	0.5	0.5	1	15	2
133.1	161.1	0.5	0.1	2	25	2
135.0	160.6	0.63	0.1	2	15	1
133.8	161.2	0.5	0.1	2	15	2
132.3	159.7	0.45	0.1	1	15	2
129.2	156.5	0.45	0.1	2	25	2
127.5	152.1	0.45	0.1	1	25	1
133.7	159.8	0.55	0.1	1	25	2
134.5	161.3	0.55	0.1	2	15	1
135.9	163.8	0.55	0.1	2	15	2
132.5	158.6	0.55	0.1	2	25	1
136.8	165.5	0.55	0.1	2	25	2
130.6	155.6	0.5	0.1	1	15	1
128.1	150.1	0.6	0.1	1	15	2
141.0	170.4	0.6	0.1	1	25	1
140.4	169.6	0.6	0.1	1	25	2
140.6	169.9	0.6	0.1	2	15	1
145.0	176.7	0.63	0.1	2	25	2
131.8	157.7	0.5	0.1	1	25	2
131.5	158.8	0.5	0.1	2	15	1
133.7	167.2	0.3	0.1	2	25	2
129.7	156.9	0.35	0.1	1	15	1
130.9	158.8	0.35	0.1	1	15	2
129.1	156.2	0.35	0.1	1	25	1
134.8	167.3	0.3	0.1	1	25	2
135.3	169.8	0.3	0.1	2	15	1
133.8	167.5	0.3	0.1	2	15	2
138.4	176.0	0.3	0.1	2	25	1
130.2	157.8	0.45	0.1	2	15	2
127.1	152.7	0.45	0.1	2	25	1
129.0	154.1	0.45	0.1	1	25	2
128.4	155.0	0.45	0.1	2	15	1
130.7	159.3	0.35	0.1	1	25	2
136.5	170.0	0.35	0.1	2	15	1
131.3	160.6	0.4	0.1	2	25	2
132.7	160.2	0.45	0.1	1	15	1
149.1	197.1	0.2	0.1	1	15	1
147.0	193.8	0.2	0.1	1	25	1
147.4	194.4	0.2	0.1	1	15	2
143.1	190.5	0.2	0.1	2	15	1

NIM1 estimate of $\alpha_{selIndel}$ (Mb)	Upper bound estimate of $\alpha_{selIndel}$ (Mb) under Meader et al. (2010) implementation	Parameterisation				
		Divergence	Clustering Coefficient	Functional Shape	Intervening Expected Length (bp)	Intervening Shape
146.6	192.2	0.2	0.1	1	25	2
143.4	191.2	0.2	0.1	2	25	1
146.2	194.5	0.2	0.1	2	15	2
138.6	175.5	0.25	0.1	1	15	1
148.5	201.8	0.2	0.1	2	25	2
131.5	155.7	0.63	0.1	2	15	2
132.3	158.9	0.5	0.1	1	25	1
127.3	153.9	0.4	0.1	2	25	1
141.2	170.3	0.6	0.1	1	15	1
139.4	166.8	0.63	0.1	1	25	2
136.9	162.9	0.63	0.1	1	25	1
129.2	156.3	0.4	0.1	2	15	2
138.4	164.9	0.63	0.1	1	15	1
138.5	167.3	0.6	0.1	2	25	2
130.8	155.1	0.6	0.1	2	25	1
136.2	163.3	0.6	0.1	2	15	2
138.6	167.2	0.55	0.1	1	25	1
135.6	162.4	0.63	0.1	2	25	1
125.0	146.3	0.55	0.1	1	15	2
141.1	179.8	0.25	0.1	1	25	1
141.0	179.5	0.25	0.1	1	15	2
140.2	179.4	0.25	0.1	2	15	1
136.3	172.3	0.25	0.1	1	25	2
144.1	186.8	0.25	0.1	2	25	1
141.0	182.0	0.25	0.1	2	15	2
135.3	167.7	0.3	0.1	1	15	1
137.4	176.3	0.25	0.1	2	25	2
131.9	162.5	0.3	0.1	1	25	1
132.4	163.1	0.3	0.1	1	15	2
135.8	169.4	0.35	0.1	2	25	1
128.9	157.9	0.35	0.1	2	15	2
130.8	159.1	0.4	0.1	2	15	1
124.8	149.2	0.4	0.1	1	25	2
130.3	157.3	0.4	0.1	1	25	1
129.4	156.0	0.4	0.1	1	15	2
125.6	149.3	0.4	0.1	1	15	1
136.4	166.0	0.5	0.1	2	25	1
139.3	166.7	0.63	0.1	1	15	2
130.9	161.3	0.35	0.1	2	25	2
132.9	157.5	0.55	0.1	1	15	1
133.2	160.2	0.5	0.1	1	15	2

3.4.2 Alignment trimming improved alignment quality and estimates of $\alpha_{selIndel}$

I considered whether the manner in which the pairwise genome alignments were constructed influenced the previous estimates of $\alpha_{selIndel}$. To assess this I estimated $\alpha_{selIndel}$ with NIM1 and NIM2 on un-trimmed alignments as provided by UCSC and on trimmed alignments that I trimmed using a log-odds approach similar to that described by Chiaromonte et al. (2002) (**Materials and methods**).

I estimated $\alpha_{selIndel}$ for four different un-trimmed mouse-rat genome alignments produced by UCSC using different alignment parameterisations and/or genome assemblies. In particular, the mm8 – rn4 and mm9 – rn5(1) alignments used less stringent alignment parameterisations compared to the mm9 – rn5(2) and mm10 – rn5 alignments (**Table 2.2** for a full list of parameters). I used mouse – rat genome alignments since these are the most closely related species I can examine, and thus provide the most stringent test for the model due to their scarcity of indels. Estimated values of $\alpha_{selIndel}$ varied markedly depending on the assembly versions used and on the alignment parameters applied (**Figure 3.5, Table 3.6**). To assess whether alignment quality influenced these results, I removed ('trimmed') poorly aligned sequence from these genome alignments, reasoning that such sequence is likely to be enriched with alignment and/or assembly errors. Alignment trimming produced lower and more uniform NIM1 and NIM2 estimates across the various alignments (**Figure 3.5A**, **Figure 3.6**). Trimming removed substantial amounts of sequence that contained transposable elements and that were aligned with unexpectedly low sequence identity (**Figure 3.5B**, **Table 3.5**), and resulted in a sizeable reduction in the numbers of short IGSs (**Figure 3.5C**). These findings are compatible with an interpretation that trimmed sequence emanates from poor-quality non-orthologous alignments.

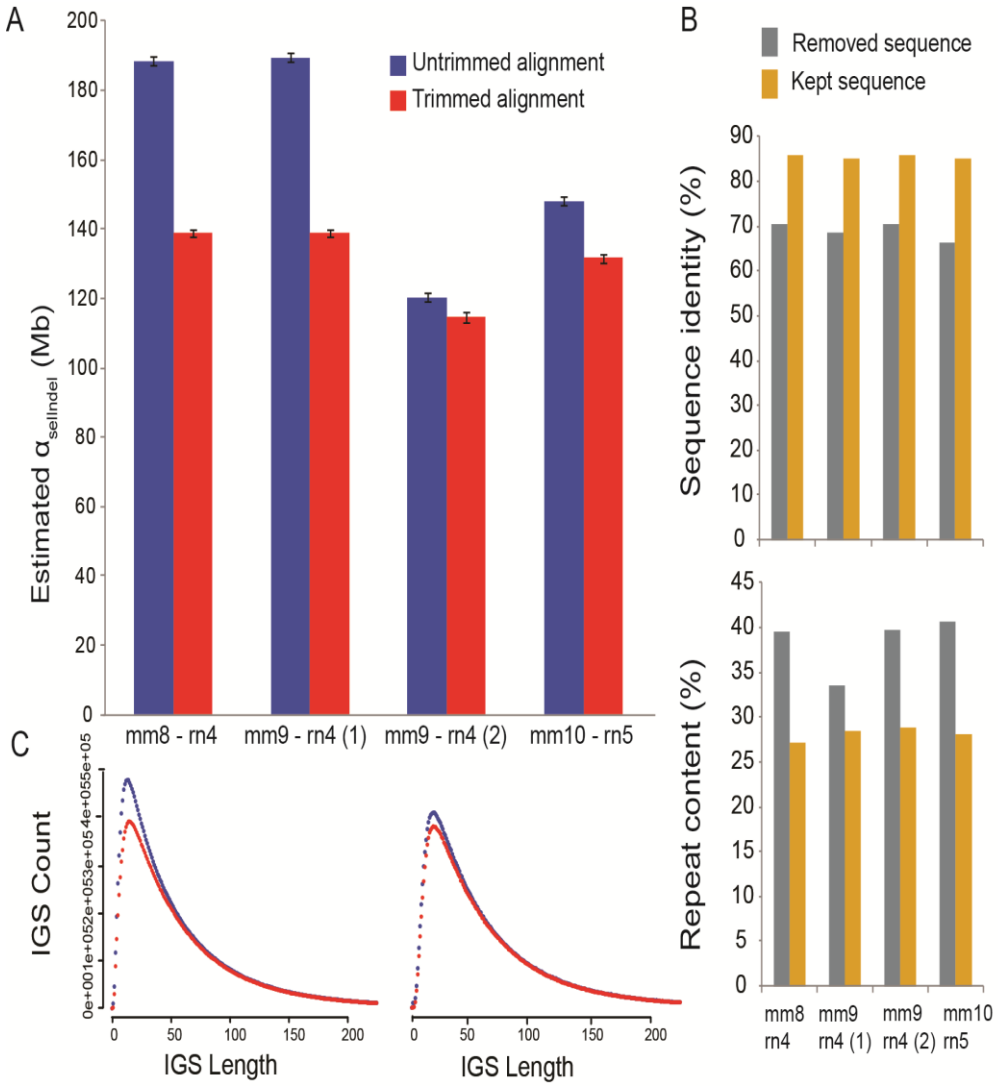

Figure 3.5: Trimming of alignments improves the consistency across alignments. The four different alignments were generated by UCSC with different genome assemblies and under different parameterisations. A. $\alpha_{selIndel}$ estimated by the NIM1 on different mouse-rat alignments. The estimates on the alignments trimmed using a log-odds approach (red) were less variable than on the untrimmed alignments (blue). This pattern was also observed when $\alpha_{selIndel}$ is estimated with NIM2 (Figure 3.3). B. The trimmed off sequence was of substantially worse quality then the remaining sequence, as shown by the removed sequence's low sequence identity and high repetitive content. C. Trimming removed more short IGSs from the mm8 – rn4/mm9 – rn4(1) (mm8 – rn4 shown left), than from the mm9 – rn4(2)/mm10 – rn5 (mm10 – rn5, right) alignments.

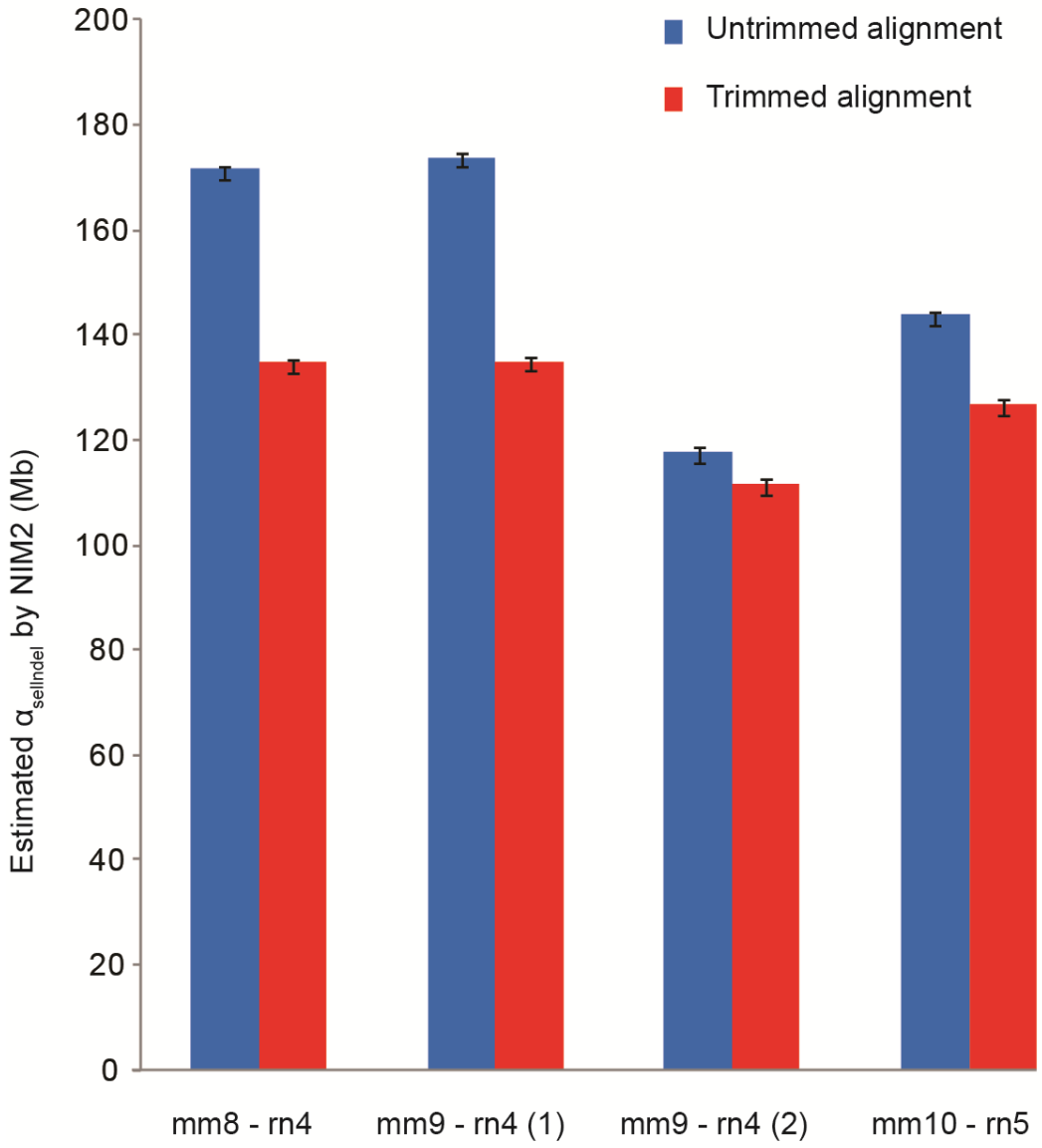

Figure 3.6: The quantity of constrained sequence estimated by NIM2 ($\alpha_{selIndel}$) on untrimmed and trimmed alignments. The trimmed alignments provide more consistent results. This trend is also seen when NIM1 is used to estimate $\alpha_{selIndel}$ (Figure 3.5).

Table 3.6: Sequence quality statistics of different mouse – rat alignments for untrimmed sequence, non-maximally positively scoring sequence trimmed off the starts and ends of alignment blocks, and internally trimmed negatively scoring IGSs. The alignments remaining after trimming are of higher quality than the trimmed-off aligned sequence in the sense that they are both less divergent and consist of proportionally fewer transposable element derived sequences.

Sequence type	Ungapped sequence length (Mb)				Substitution Divergence			
	mm8 rn4	mm9 rn4 (1)	mm9 rn4 (2)	mm10 rn5	mm8 rn4	mm9 rn4 (1)	mm9 rn4 (2)	mm10 rn5
Remaining sequence	1681	1696	1689	1743	0.144	0.145	0.150	0.148
Trimmed start	46.00	48.42	8.441	17.88	0.280	0.281	0.309	0.295
Trimmed end	46.61	49.08	8.695	18.46	0.280	0.281	0.312	0.295
Trimmed internal	7.649	7.857	3.415	4.673	0.473	0.473	0.463	0.465

Sequence type	Mean IGS length (bp)				Proportion Repetitive			
	mm8 rn4	mm9 rn4 (1)	mm9 rn4 (2)	mm10 rn5	mm8 rn4	mm9 rn4 (1)	mm9 rn4 (2)	mm10 rn5
Remaining sequence	64.2	64.3	71.8	68.3	0.272	0.287	0.281	0.284
Trimmed start	22.4	22.6	20.6	20.8	0.395	0.396	0.411	0.329
Trimmed end	22.3	22.4	20.0	20.0	0.394	0.395	0.408	0.333
Trimmed internal	25.3	25.5	32.7	26.9	0.416	0.412	0.387	0.372

3.4.3 NIMs yielded concordant estimates of $\alpha_{selIndel}$ across diverse eutherian species

I applied the two neutral indel models to estimate $\alpha_{selIndel}$ on trimmed whole genome alignments between a wide variety of eutherian species pairs for which high quality genome assemblies are available. Estimates of $\alpha_{selIndel}$ were largely concordant between the two models (**Table 3.7**). By contrast, the previous $\alpha_{selIndel}$ estimates (Meader et al. 2010) are 60–90% higher than my improved estimates (**Table 3.7**). These differences are largely

attributable to alignment trimming, but also to the revised manner in which the confidence interval is calculated for the $\alpha_{selIndel}$ estimates (**Materials and methods**). The quantity of constrained sequence is positively correlated with its G+C content, as illustrated by the larger areas lying between the geometric distribution and the data when IGS histograms are plotted across 20 equally populated GC-bins for each of the genome alignments (**Appendix Figures A.1–A.16**).

Table 3.7: Estimates of $\alpha_{selIndel}$ between different eutherian species pairs under different models. The species symbols correspond to the genome assemblies used as shown in Table 2.1.

Species pair	Estimated quantity of $\alpha_{selIndel}$ (Mb)		
	NIM1	NIM2	Meader et al. 2010
hg19 – equCab2	110.5 - 112.0	118.9 - 120.1	150.8 - 200.8
hg19 – cerSim1	110.8 - 112.1	119.7 - 120.9	N/A
hg19 – otoGar3	106.8 - 108.2	109.1 - 110.2	N/A
hg19 – canFam2	100.8 - 101.9	101.7 - 102.6	121.8 - 151.1
hg19 – ailMel1	101.4 - 102.5	105.5 - 106.4	N/A
hg19 – bosTau7	89.8 - 90.6	90.7 - 91.6	114.3 - 143.6
hg19 – oryCun2	88.8 - 89.7	93.0 - 93.9	N/A
hg19 – cavPor3	81.9 - 82.7	81.2 - 82.0	N/A
hg19 – mm10	68.8 - 69.4	66.6 - 67.1	81.4 - 96.2
mm10 – rn5	130.4 - 132.9	125.6 - 127.5	189.0 - 258.4
mm10 – equCab2	68.9 - 69.5	66.5 - 67.1	76.3 - 91.0
mm10 – canFam2	64.9 - 65.5	60.8 - 61.3	71.1 - 83.0
mm10 – bosTau7	62.9 - 63.4	56.4 - 56.9	63.8 - 74.5
canFam2 – mpf_v1	135.6 - 137.7	141.2 - 142.9	N/A
canFam2 – equCab2	117.4 - 118.9	126.5 - 127.8	147.6 - 194.5
canFam2 – bosTau7	92.5 - 93.6	91.3 - 92.2	114.8 - 144.0

3.4.4 $\alpha_{selIndel}$ estimates were robust to the presence of indel hotspot regions

Given the impact that heterogeneity in the neutral indel rate can have on methods for detecting selection (Kvikstad and Duret 2014), I also examined the performance of the NIM1 applied to data from which known indel hotspot genomic locations were removed. The hotspot locations were identified by Montgomery et al. (2013), who found genomic regions of elevated indel rate using a local indel rate model applied to human polymorphism data

from the 1000 genomes project (Abecasis et al. 2012). Estimates of $\alpha_{selIndel}$ were not substantially influenced by the presence/absence of these indel hotspot regions (**Figure 3.7**), and consequently I retained such sequences in subsequent analyses.

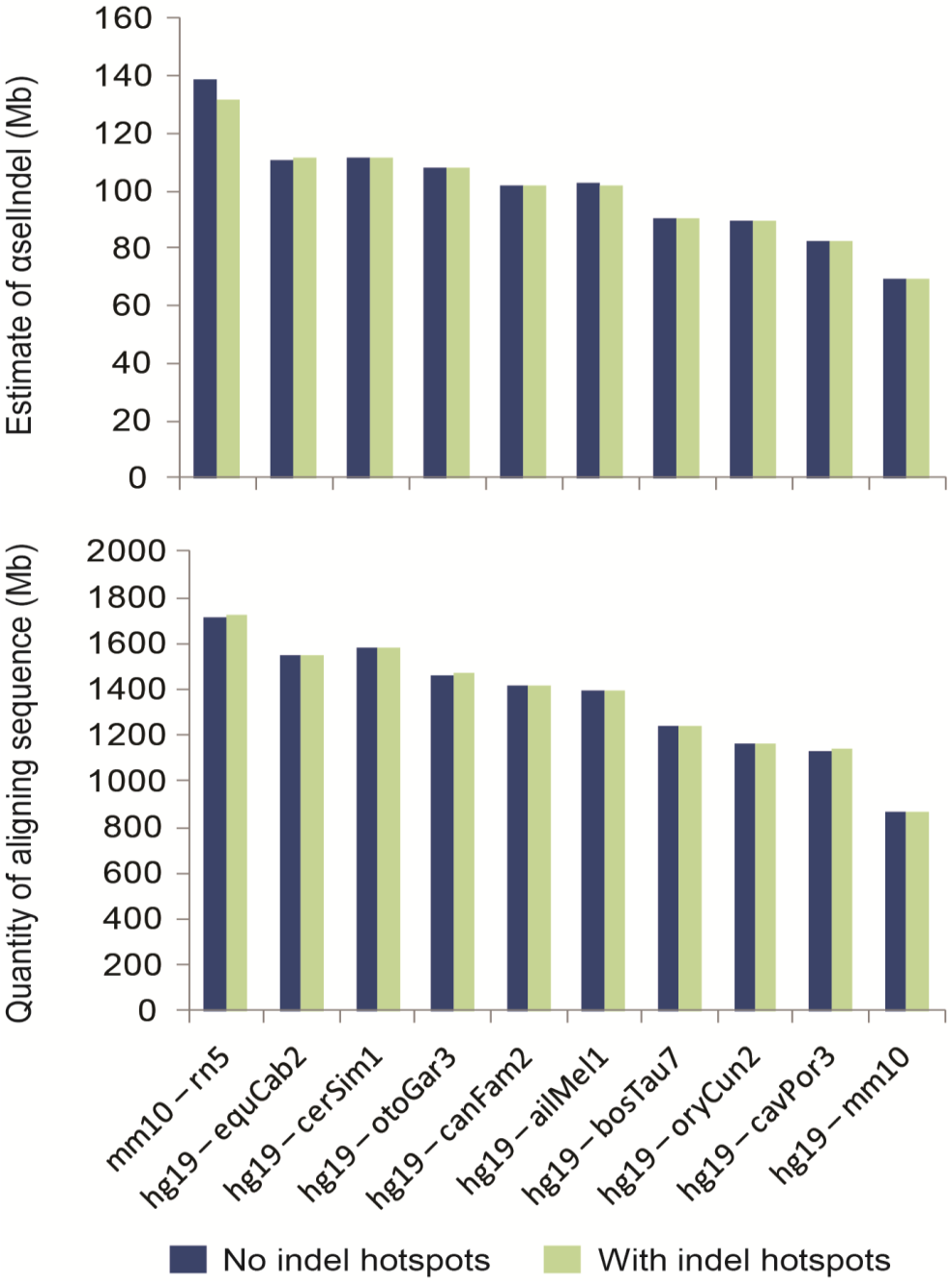

Figure 3.7: Estimates of $\alpha_{selIndel}$ and the quantity of aligned sequence in the presence and absence of indel hotspot locations defined by Montgomery et al. (2013) in the human and mouse genomes.

3.4.5 Non-reciprocally aligned sequence had a small effect on $\alpha_{selIndel}$ estimates

I removed non-reciprocally aligned sequence to test the impact this had on estimates of $\alpha_{selIndel}$. Non-reciprocally aligned sequence is sequence that aligns when the first sequence (from species 'A') is used as the query and the second sequence (from species 'B') as the target, or when species B sequences are the query and species A sequences as the target, but not in both cases. If alignments were perfect, then all sequence would be reciprocally aligned, but in reality non-reciprocally aligned sequence will exist because of difficulties in defining orthology and paralogy relationships due to the repetitive nature of genome sequences. The rationale for removing non-reciprocally aligned sequence is that such sequence is presumably lower-confidence aligned sequence than the reciprocally aligned sequence. However, removing this lower quality non-reciprocally aligned sequence from mouse – rat (and rat – mouse) or human – mouse (and mouse – human) alignments had only a small effect on estimates of $\alpha_{selIndel}$, and so the main analyses were conducted with such sequence retained.

Table 3.8: NIM1 estimates of $\alpha_{selIndel}$ and the quantity of aligned sequence on alignments processed to remove non-reciprocally aligned sequence. In parentheses are the differences from the $\alpha_{selIndel}$ estimates from the trimmed alignments where non-reciprocally aligned sequence was not discarded.

Species Pair	Estimate of $\alpha_{selIndel}$	Aligned sequence (Mb)
mm10 – rn5	136.0 (+5)	1607.4 (-118)
hg19 – mm10	62.6 (-6.5)	751.7 (-114.9)

3.4.6 Enredo-Pecan-Ortheus alignments contained an abundance of long indels

A further way to investigate the robustness of the NIM1 is to test the effect of using a completely different alignment methodology on estimates of $\alpha_{selIndel}$. To this end, I extracted mouse – rat alignments from multi-way Enredo-Pecan-Ortheus (EPO) alignments (Paten et al. 2008). LASTZ, the software I used to construct whole genome pairwise alignments in the main analyses, took a seed based approach to alignment and clusters together short alignments to create larger ones. By contrast, EPO alignments were built with a consistency approach where additional outgroup sequences are used to resolve uncertainties in pairwise alignments. Enredo was used to create a homology map of collinear sequences between genome sequences handling large scale genomic re-arrangements. Then Pecan aligned the genome sequences following a probabilistic consistency methodology. (Note that the Ortheus component constructs ancestral sequences which I did not use here.) The resulting EPO alignments were strikingly different from the LASTZ alignments in the respect that LASTZ alignments consisted of far more smaller alignment blocks; an example given by Paten et al. (2008) shows that LASTZ created approximately 500-fold fewer alignment blocks than Enredo, but the N50 of alignment blocks (that is the alignment block length such that if the alignment blocks were ordered by length, then half of all alignment columns would be covered by the time the alignment block of this length is reached) was about eight-fold less for LASTZ alignments.

I found estimates of $\alpha_{selIndel}$ on EPO mouse – rat alignments were much higher than those based on the LASTZ alignments prior to trimming, but are considerably lower after trimming (**Table 3.9**). This is due to the vast quantity of aligned sequence trimmed off, 666Mb in total, the vast majority trimmed off being terminal sequences rather than internally removed IGSs. Consequently it would appear that the EPO alignments had an abundance of negatively scoring terminal sequence. Visual inspection of the alignments makes it clear that this is a

consequence of the vast number of gapped alignment columns. While the sequence identity remains quite high for the trimmed off terminal sequence at 83.7%, the alignments contained a total of almost 760Mb of gapped alignment columns. An interpretation of this is that the EPO software tends to create long gaps between genuinely homologous sequences, rather than break up the alignment into more alignment block chunks. Although the inconsistency between the estimates of $\alpha_{selIndel}$ on the LASTZ and EPO alignments implies that my approach is not fully robust to the alignment methodology used, the differences appear to be attributable to the excess of artefactual long indels that the EPO alignments contained. Therefore, subsequently analyses focus on the LASTZ alignments and I do not here consider the EPO alignments further.

Table 3.9: NIM1 estimates of $\alpha_{selIndel}$ and the quantity of aligned sequence from EPO alignments. In parentheses are the differences from the LASTZ alignments. However, note that the LASTZ alignments were conducted on the rn5 rat assembly rather than the earlier rn4 assembly used for construction of the EPO alignments.

EPO Alignments	Estimate of $\alpha_{selIndel}$ (Mb)	Aligned sequence (Mb)
Untrimmed mm10 – rn4	209.7 (+61.5)	1442.1(-319.5)
Trimmed mm10 – rn4	94.5 (-37.1)	776.1 (-949.3)

3.5 Discussion

Although the first estimate of α_{sel} was made approximately a decade ago (Chiaromonte et al. 2003), there are many complications to ensure such estimates are robust, including both biological (such as neutral mutation rate variation and lineage-specific effects) and technical (such as sequencing or alignment errors) factors that cause evolutionary rates to vary and that may crucially lead to false inference of the background neutral mutation rate (Green and Ewing 2013). These factors have been demonstrated to confound a recent high profile estimate of α_{sel} and will confound other earlier estimates too (Ward and Kellis 2012; Green and Ewing 2013; Ward and Kellis 2013). I have demonstrated that these issues affected the

previous estimate of Meader et al. (2010), and have introduced new and improved methods and refinements for calculating $\alpha_{selIndel}$. The alignment trimming I introduced in particular had a substantial impact on the resulting estimate of $\alpha_{selIndel}$.

Qualitatively my results reinforce a main conclusion of many previous studies, such as (Siepel et al. 2005; Meader et al. 2010): the proportion of constrained sequence shared between eutherian species pairs is in considerable excess of the quantity of protein coding sequence, and thus the majority of functional sequence in eutherian and other mammalian genomes is predicted to be noncoding. However, quantitatively I provide a more robust estimate of α_{sel} for species pairs than previous studies, and I demonstrate my approach is not substantially affected by the presence of indel hotpots or the removal of non-reciprocally aligned sequence. The method also performs consistently when different mouse and rat genome assemblies are inputs for the model, unlike the previous indel approach of Meader et al. (2010). Note that even the older mouse and rat genome assemblies I use are still of relatively good quality compared to many de novo assemblies created with next generation sequencing technologies; it is unknown whether the model will perform well on lower quality genome assemblies. Therefore, I avoided applying the model to alignments involving low coverage genome assemblies, such as the 29 mammalian assemblies of 2X coverage produced by the Broad Institute (Lindblad-Toh et al. 2011).

Another layer of evidence for the robustness of my approach for estimating $\alpha_{selIndel}$ comes from the genome simulations I conducted, which were much more extensive than those of Meader et al. (2010). These provided corroborating evidence that $\alpha_{selIndel}$ estimates are robust to a wide variety of biological parameterisations that were estimated from biological data where possible, or otherwise simulated over a range of plausible values. Although these simulations did not attain the complex genome architecture found in real vertebrates genomes, with no large-scale synteny rearrangements and intricate repeat structures, they

were as or more realistic than widely used previous models simulating genome evolution, such as PAML's Evolver or Seq-Gen that models only substitutions (Rambaut and Grassly 1997; Yang 2007), indel-Seq-Gen that only simulated indel evolution (Strope et al. 2007), or INDELible that is of similar complexity to the model I used with an integrated substitution and indel model (Fletcher and Yang 2009).

There do remain issues that could impact on estimates of α_{sel} that may warrant further investigation. The estimates of α_{sel} are affected by the alignment methodology, and whilst I attribute this to the poor quality of the EPO alignments in the respect that the algorithm creates long un-gapped alignments, this could be examined further in this context and those of other previous studies. Generally, our understanding of the patterns of indel mutations is in its infancy. Improvements in alignment algorithms could better deal with problems of gap attraction and gap annihilation (Lunter et al. 2008). Furthermore indel homoplasy, where the same indels occur in different lineages due to convergent evolution, may be common (Kvikstad and Duret 2014). These indels will be called as shared ancestral events and not as being lineage-specific. Despite these further potential considerations, the methods detailed here are a marked improvement on previous approaches for estimating α_{sel}, as illustrated by several strands of evidence that these current estimates are relatively accurate and robust.

Chapter 4: Variation in rates of turnover across functional element classes in the human lineage

4.1 Abstract

Recent large scale projects have been conducted to annotate and catalogue the biochemical activity of the human genome. Previous studies have also shown that a small proportion of the human genome is widely conserved across diverse mammalian species. Furthermore, it has been demonstrated that the quantity of constrained, and thus putatively functional, genomic sequence is strongly negatively correlated with divergence between the species under consideration. This is consistent with the notion that functional sequence is turning over rapidly as it is lost and gained over time. Turnover has been shown experimentally for a handful of specific classes of functional elements, such as transcription factor binding sites that are often species-specific. Here I applied two Neutral Indel Models that identify constrained sequence with respect to insertions and deletions (indels) to provide the first genome-wide catalogue of the constraint and turnover across different classes of human functional element. I found that protein coding sequence showed by far the highest levels of constraint and the lowest levels of turnover consistent with the notion that such sequence is static over long evolutionary time periods. By contrast, noncoding sequence is estimated to turnover rapidly, such that half of functional noncoding sequence is lost and re-gained over 127My. Long noncoding RNAs, enhancers and promoters show evidence of particularly rapid turnover, while DNase I hypersensitive sites have been relatively stable. Through extrapolations I estimate that 7.1–9.2% of human genomes is subject to present-day purifying selection and thus likely to be currently functional.

4.2 Introduction

"What proportion of the human genome is functional?" remains a contentious question (Ponting and Hardison 2011; Pennisi 2012; Graur et al. 2013). A related and arguably more important issue is to improve our limited understanding of the dominant evolutionary processes that operate at a genomic scale. One possibility is that genome evolution is characterised by pervasive purifying selection, punctuated by rare episodes of gene loss and positive selection in response to changing evolutionary pressures. This model is widely accepted and provides a good description of the evolution of protein coding genes (Waterston et al. 2002). Nevertheless, it is questionable whether it adequately describes the evolution of the functional noncoding portion of the genome. For instance, recent large-scale functional screens and sequencing projects have revealed large numbers of biochemically active noncoding elements that are not widely conserved across species (Dermitzakis and Clark 2002; Lindblad-Toh et al. 2011; Dunham et al. 2012).

Restricting consideration to those studies that define functional nucleotides as those that are subject to purifying selection, various estimates of the proportion of functional nucleotides in the human genome have been published, ranging from 3% to 15% (Ponting and Hardison 2011; Ward and Kellis 2012). This fraction (or quantity, α_{sel}) is usually estimated from genomic comparisons between species. Since each species' lineage gains and loses functional elements over time, α_{sel} needs to be understood in the context of divergence between species. The divergence influences the estimate of α_{sel} in two ways. On the one hand, constrained sequence between closely related species, including lineage-specific constrained sequence, is harder to detect than more broadly conserved sequence because of a paucity of informative mutations, which reduces detection power. On the other hand, estimates of constraint between any two species will only include sequence that was present in their common ancestor and is predominantly constrained in both extant species' genomes, with the

consequence that turnover of functional sequence leads to diminishing α_{sel} estimates as the species divergence increases. Assuming that the first effect can be controlled for, higher estimates of sequence constraint that are obtained between more closely related species (Smith et al. 2004; Meader et al. 2010) are thus indicative of the turnover of functional sequence (Meader et al. 2010). Here I considered turnover to mean the loss or gain of purifying selection at a particular locus of the genome, when changes in the physical or genetic environment, or mutations at the locus itself, cause the locus to switch from being functional to being non-functional or vice versa.

Two previous studies have made quantitative estimates of the overall rate of turnover (Smith et al. 2004; Meader et al. 2010), reviewed by Ponting and Hardison (2011). The estimate by Smith et al. (2004) was derived from an analysis of point mutations in alignments across a 1.8Mb genomic region. While a high rate of turnover was inferred, the authors emphasised the preliminary nature of their work as a consequence of the limited amount of data available to them at that time. Later, Meader et al. (2010) performed genome-wide analysis with a Neutral Indel Model (here referred to as NIM1) to estimate the quantity, termed $\alpha_{selIndel}$, of human sequence that was constrained with respect to insertions or deletion mutations (indels). This study also found a high rate of turnover, and estimated that at least 6.5% of human genomes is functional.

I apply two neutral indel models explained in **Chapter 3** to pairwise alignments between the genomes of diverse eutherian mammals. I obtain a revised estimate of 7.1–9.2% for the proportion of human genomes that is presently subject to purifying selection, equating to 220–286Mb of constrained sequence. I also take advantage of the additional high-quality eutherian genome sequences that have become available since the previous study to provide more reliable estimates of the rate of turnover of functional sequence in these species. Improvements in biochemical annotation of genomic sequence, predominantly due to the

experiments of the ENCODE consortium (Dunham et al. 2012), mean that I can investigate turnover rates within particular classes of functional elements.

4.3 Materials and methods

4.3.1 Coding and untranslated region sequences

Coding sequence for human (hg19), mouse (mm10), and dog (canFam2) and untranslated region (UTR) annotations for human (hg19) were obtained from Ensembl version 72 (http://www.ensembl.org/index.html). UTRs can affect translational efficiency and the stability of mRNA sequences. UTRs overlapping coding sequence were not considered in the UTR analyses. Coding sequence and UTR annotations were constructed combining information from genomic DNA, cDNA, EST, and protein data with Ensembl's gene prediction pipeline, which consists of four steps (Curwen et al. 2004). (1) species-specific proteins were mapped to the genome assembly of interest and Genewise was used to build a transcript structure for the aligned protein. (2) a set of reference proteins was taken from closely related species and these were used to build transcript models in regions where transcripts were not identified in the first step. (3) species-specific cDNA and EST sequences were aligned to the genome with genebuild to create further transcript models, with non-translated cDNA forming the UTR sequence. (4) all transcript models were combined and merged to create full gene models. For the human and mouse annotations, a number of additional transcript datasets were merged to create the final predictions, including manually curated transcripts (Curwen et al. 2004).

4.3.2 Conserved elements

Human (hg19) PhastCons conserved elements were taken from the vertebrate PhastConsElements46way track downloaded from UCSC Genome Informatics (http://genome.ucsc.edu/). PhastCons is a phylogenetic hidden Markov model (phylo-HMM) that classifies elements into one of two states, either conserved or neutral, based on a

calibration that fixes the total fraction of sequence that is conserved at 5% (Siepel et al. 2005).

Human (hg19) GERP++ conserved elements were downloaded from the Sidow laboratory website (http://mendel.stanford.edu/SidowLab/downloads/gerp/). GERP++ is a position specific model that first uses large cross-species alignments to assign conservation scores to each individual alignment column. Highly scoring alignment columns are then clustered together using a peak-calling type approach to determine the location of conserved elements (Davydov et al. 2010).

4.3.3 ENCODE derived annotations

Many human (hg19) annotations were taken from the ENCODE data available at UCSC Genome Informatics (http://genome.ucsc.edu/ENCODE/). Specifically, the transcription factor binding sites (TFBSs) were acquired from the ENCODE merged sets (wgEncodeRegTfbsClusteredV2.bed). The TFBSs were identified through chromatin immunoprecipitation sequencing (ChIP-seq). The results from different ChIP-seq experiments were combined across different cell types and TFs by clustering together different BS peak-calls. The DNase1 hypersensitivity sites (DNase HSs) were also merged sets (wgEncodeRegDnaseClusteredV2.bed). DNase HSs mark regions of open-chromatin through cleaving by DNase I enzymes, and are likely markers of cis-regulatory elements.

Promoter and enhancer elements, which are involved in the initiation and enhancement of transcription, were extracted from the ENCODE HMM Chromatin State segmentation tracks, and were merged and normalised across these samples: wgEncodeBroadHmmGm12878HMM.bed.gz, wgEncodeBroadHmmH1hescHMM.bed.gz, wgEncodeBroadHmmHepg2HMM.bed.gz, wgEncodeBroadHmmHmecHMM.bed.gz, wgEncodeBroadHmmHsmmHMM.bed.gz, wgEncodeBroadHmmHuvecHMM.bed.gz, wgEncodeBroadHmmK562HMM.bed.gz, wgEncodeBroadHmmNhekHMM.bed.gz, and

wgEncodeBroadHmmNhlfHMM.bed. These nine cell lines represent two cancerous and seven non-cancerous cell lines. The HMM defines enhancers, promoters and other chromatin states based on chromatin marks of histone modifications.

4.3.4 Repeat elements

Repetitive element annotations were taken from RepeatMasker, which identifies transposable elements and low complexity regions (Smit et al. 1996-2010). Unlike the other mammalian genome assemblies, which were downloaded from UCSC, the ferret genome assembly (mpf_v1) was devoid of repeat annotation. I masked repeats in the genome assembly using the generic carnivore libraries produced by RepeatMasker (Smit et al. 1996-2010). This will miss denovo repeats specific to the ferret genome, but this is unimportant in this context, since the purpose of marking repeats is only to speed up the alignment, which is achieved since LASTZ skips over masked repeats during the seeding stage of alignment.

4.3.5 Long noncoding RNAs

Human long noncoding RNAs (lncRNAs) were taken from a recently published large catalogue (Hangauer et al. 2013). LncRNAs are transcripts of greater than 200 nucleotides in length that do not code for the production of proteins. These lncRNAs were predicted from a total of 144 different RNA sequencing (RNA-seq) data sets across 23 different human tissues, consisting of a total of over 4.5 billion uniquely mapped reads.

4.3.6 Comparing the rates of turnover between two functional element types

Using the birth-death model to examine the turnover of functional sequence that I introduced in **Chapter 2**, I estimated parameters *a* and *b* for the time-homogeneous turnover model $\alpha = ae^{-bd}$ by weighted linear regression. Here, *a* represents the total fraction of sequence subject to present-day purifying selection and *b* is the rate of turnover. To compute *p* values for differences in turnover rates, I fitted two nested models to the two data sets using the length of the 95% confidence interval to calculate the weight for each data point, where the null

model has a single turnover rate parameter *b*, while the alternative has independent *b* parameters for the two data sets; I then performed a likelihood ratio test to assess significance.

4.3.7 Modelling the turnover of pan-mammalian conserved elements

To assess whether putatively functional sequence lacking evidence for pan-mammalian conservation, more specifically sequence covered by PhastCons (Siepel et al. 2005) or GERP++ (Davydov et al. 2010) conserved elements, shows enrichment for NIM1-inferred conservation I cannot apply the turnover model $\alpha = ae^{-bd}$ directly as the alternative model. This is because I have purposely removed sequence showing evidence for pan-mammalian selection, resulting in a downward shift of α independent of *d*. This could be modelled by introducing an additional offset parameter to the model, but it was found that this caused the model to be over-determined because of the relatively few data points that were available. Instead, as an alternative I used a linear model $\alpha = a - bd$; since the turnover model is concave, this leads to conservative estimates for the *a* parameter. I used the same model for the null, since there is no expectation as to the dependence of the observations of artefactual constraint α on the divergence under the null hypothesis, so a linear model appears a robust choice.

4.4 Results

4.4.1 Rapid turnover of functional sequence across eutherian evolution

I estimated $\alpha_{selIndel}$ between mammlian genomes by applying two neutral indel models, the Neutral Indel Model 1 (NIM1) and the Neutral Indel Model 2 (NIM2) to whole genome alignments across diverse eutherian species (**Figure 4.1**). The alignments were processed as described in **Chapter 3** to remove poorly aligned sequences.

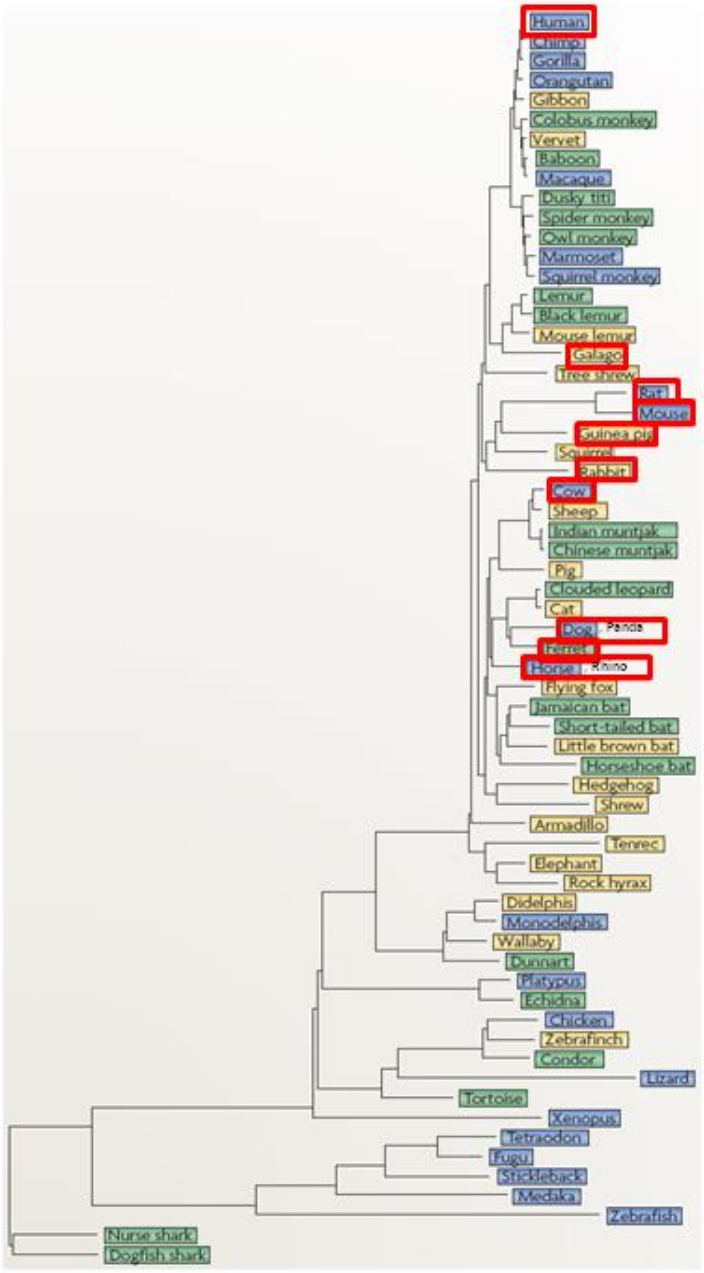

Figure 4.1: Phylogeny showing the relationship between the different vertebrate species. The species whose genome sequences were examined in the constrained sequence analyses are enclosed in red boxes. The branch lengths are based on the substitution divergence of 4-fold degenerate sites. The figure was modified from Margulies and Birney (2008). Reprinted by permission from Macmillan Publishers Ltd: Nature Genetics, Margulies and Birney, copyright 2008.

To examine whether the amount of shared constrained sequence varied as a function of divergence, I took the putatively neutrally evolving standard of ancestral repeat (AR) sequence (defined as sequence derived from transposable elements whose insertion predates the species' last common ancestor) as a proxy of divergence. These are an appropriate proxy because virtually all ARs show the patterns of indel mutation expected under neutral evolution (Lunter et al. 2006). My estimates of divergence using either ARs or synonymous sites as a neutral proxy were concordant; the results are thus insensitive to the choice of putatively neutral sequence (**Figure 4.2**). To convert this divergence to years, I applied a substitution rate of 2.2×10^{-9} per site per year (Kumar and Subramanian 2002). This will be a more appropriate value for the human lineage, on which I focus, than on rodent lineages whose per-year substitution rate and thus turnover rates per year are substantially higher. To convert this time to the age of the most recent common ancestor, it should further be halved since two branches lead from this ancestor.

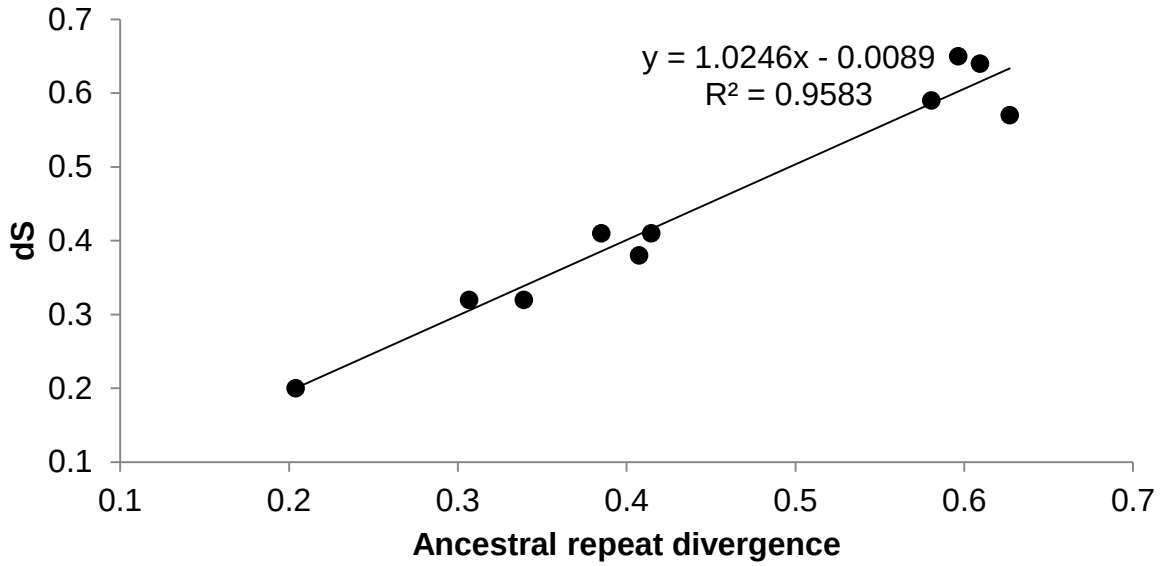

Figure 4.2: Strong positive correlation between ancestral repeat (AR) divergence and synonymous substitution rate (dS) for the following mammalian species pairs: human – cow, human – dog, human – horse, human – mouse, mouse – rat, mouse – cow, mouse – horse, mouse – dog, dog – cow and dog – horse. The correlation implies that my results are robust to the choice of neutral standard.

I observe a strong negative correlation between estimates of $\alpha_{selIndel}$ and the divergence of the two species compared in the alignment (**Figure 4.3**). This is consistent with substantial turnover of functional sequence and thus with earlier conclusions (Smith et al. 2004; Meader et al. 2010), and is inconsistent with simulation results under a scenario in which turnover is absent.

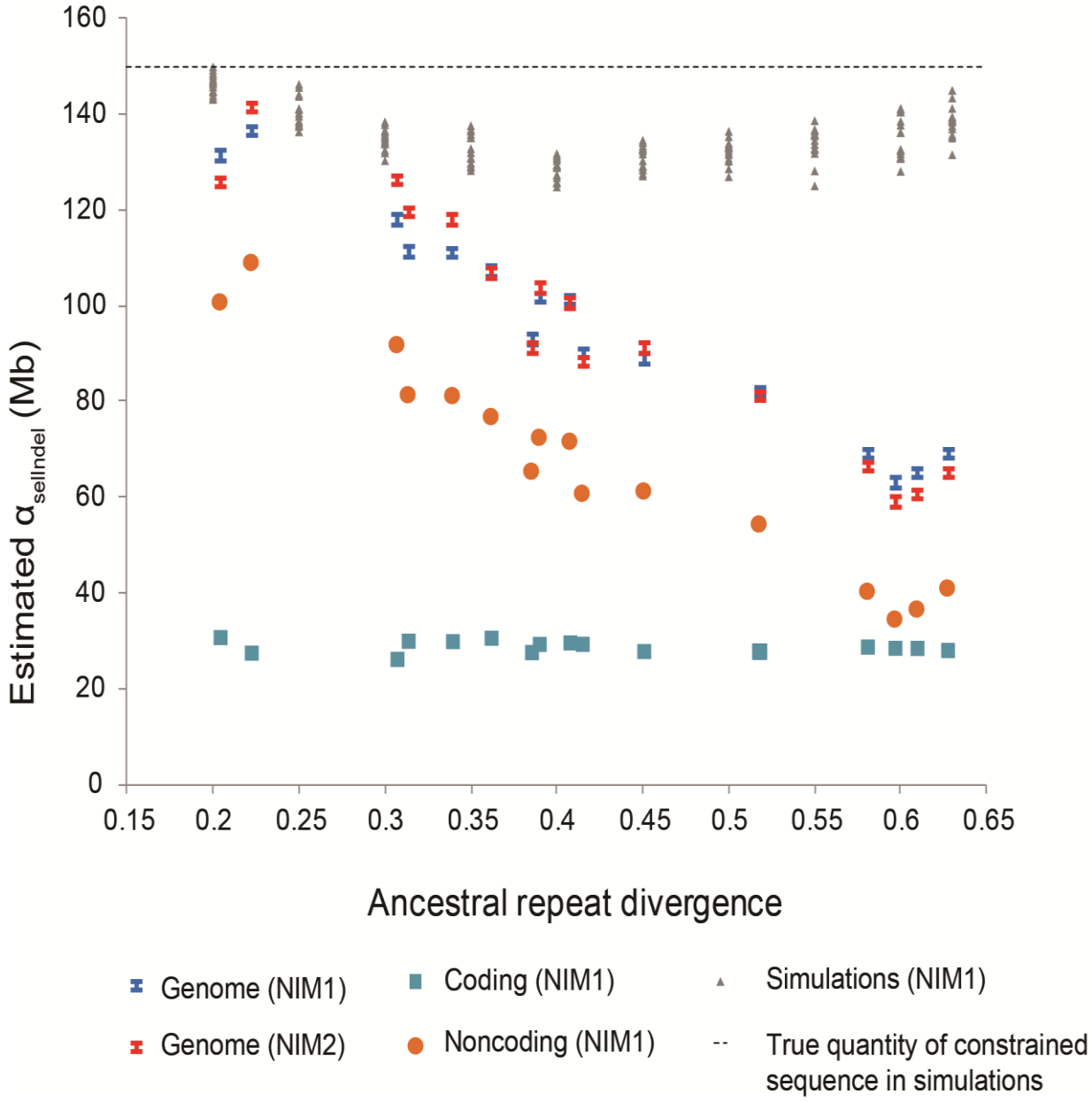

Figure 4.3: Evolutionary turnover of constrained sequence. Estimated quantity of constrained sequence ($\alpha_{selIndel}$) estimated by NIM1 (in blue) and NIM2 (red) plotted against ancestral repeat divergence for different pairs of eutherian species' genomes. Estimated using the NIM1, coding sequence (blue squares) is seen to be broadly conserved, while constrained noncoding sequence (orange circles) shows a strong negative correlation between $\alpha_{selIndel}$ and divergence, implying that it turns over rapidly.

4.4.2 Technical artefacts could not explain observed signatures of turnover

However, it is *a priori* conceivable that these observations are nevertheless the result of an unspecified technical artefact that I failed to consider, leading to a lack of specificity when identifying constrained sequence in particular for less divergent species pairs. I sought to exclude this possibility, by showing that lineage-specific constrained sequence identified by NIM1 is enriched for sequence with biochemical annotation for function.

To argue this, let us hypothesize that functional sequence does *not* exhibit turnover to any significant degree. I will show that this hypothesis is incompatible with my observations. Invoking the premise that function equates to present-day constraint, sequence with associated experimental evidence for biochemical function in human ("putative functional sequence") but that also lacks evidence for long-term, pan-mammalian conservation, would represent either: a false positive of the functional experiment and in fact be neutrally evolving sequence, or a false negative of the algorithms to detect pan-mammalian conservation and actually be pan-mammalian but weakly conserved, functional sequence. (The third possibility, short-lived constrained sequence that was not identified by algorithms not optimized for this type of constraint, is excluded by hypothesis.) In the first case, I do not expect such neutral sequence to be enriched among purportedly constrained sequence identified by the NIM1, as their distinguishing feature (the ability to cause a false positive in a functional experiment) is not expected to correlate with a spurious lack of indels in neutral sequence. In the second case, I would not expect the NIM1, which uses only the signal of indels between pairs of species, to detect such pan-mammalian but weakly conserved sequence with better power than algorithms that integrate the signal of single-nucleotide substitutions across multiple species, and that were specifically designed to detect the signal of pan-mammalian conservation. Either way, under the assumption that functional sequence exhibits turnover to a minimal degree, I concluded that putative functional sequence that

shows no signature of pan-mammalian constraint would not be enriched with NIM1-constrained sequence.

Instead, I observed precisely this enrichment, as I now discuss. I used two substitution-based methods that identify pan-mammalian conserved sequence: the two-state phylogenetic HMM PhastCons (Siepel et al. 2005), and the position specific model GERP++ (Cooper et al. 2005; Davydov et al. 2010). Between closely related species, NIM1 identified substantial amounts (e.g. 24Mb between human and horse) of putatively constrained sequence that were not detected by either PhastCons or GERP++ as being pan-mammalian conserved (**Figure 4.4**). As putatively lineage-specific functional sequence I took mutually exclusive sets of predicted enhancers (530Mb), TFBSs (79Mb), and DNase HSs (116Mb) as defined from experimental evidence by the ENCODE project (Dunham et al. 2012), that in addition do not overlap either PhastCons or GERP++ conserved elements. Within each of these sets, NIM1 identified significantly higher fractions of constrained sequence compared to a control set of sequence ($p = 8\times10^{-8}$, $p = 1\times10^{-4}$, $p = 2\times10^{-10}$, **Figure 4.5A, and 4.5B**). The control set was defined as sequence not covered by PhastCons or GERP++ elements, not within 50bp of enhancers, DNAse HSs, TFBSs, or promoters as defined by ENCODE, not within 50bp of protein coding exons or UTRs as defined by Ensembl, and not within 50bp of the lncRNAs from (Hangauer et al. 2013) (1148Mb in total for this control set).

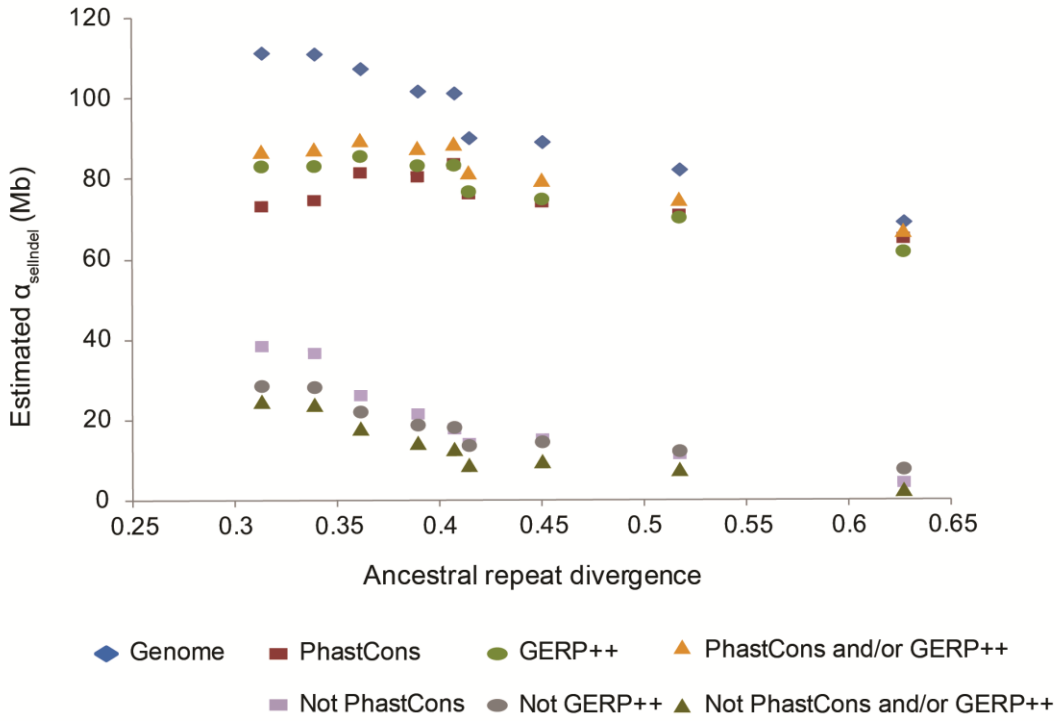

Figure 4.4: The overlap of constrained sequence with pan-mammalian conserved sequences. The quantity of constrained sequence estimated by NIM1 that overlaps sequence identified as conserved by either PhastCons and/or GERP++. Much of the lineage-specific constrained sequence identified by NIM1 was not detected by these other methods that mainly have power to identify pan-mammalian conserved sequences.

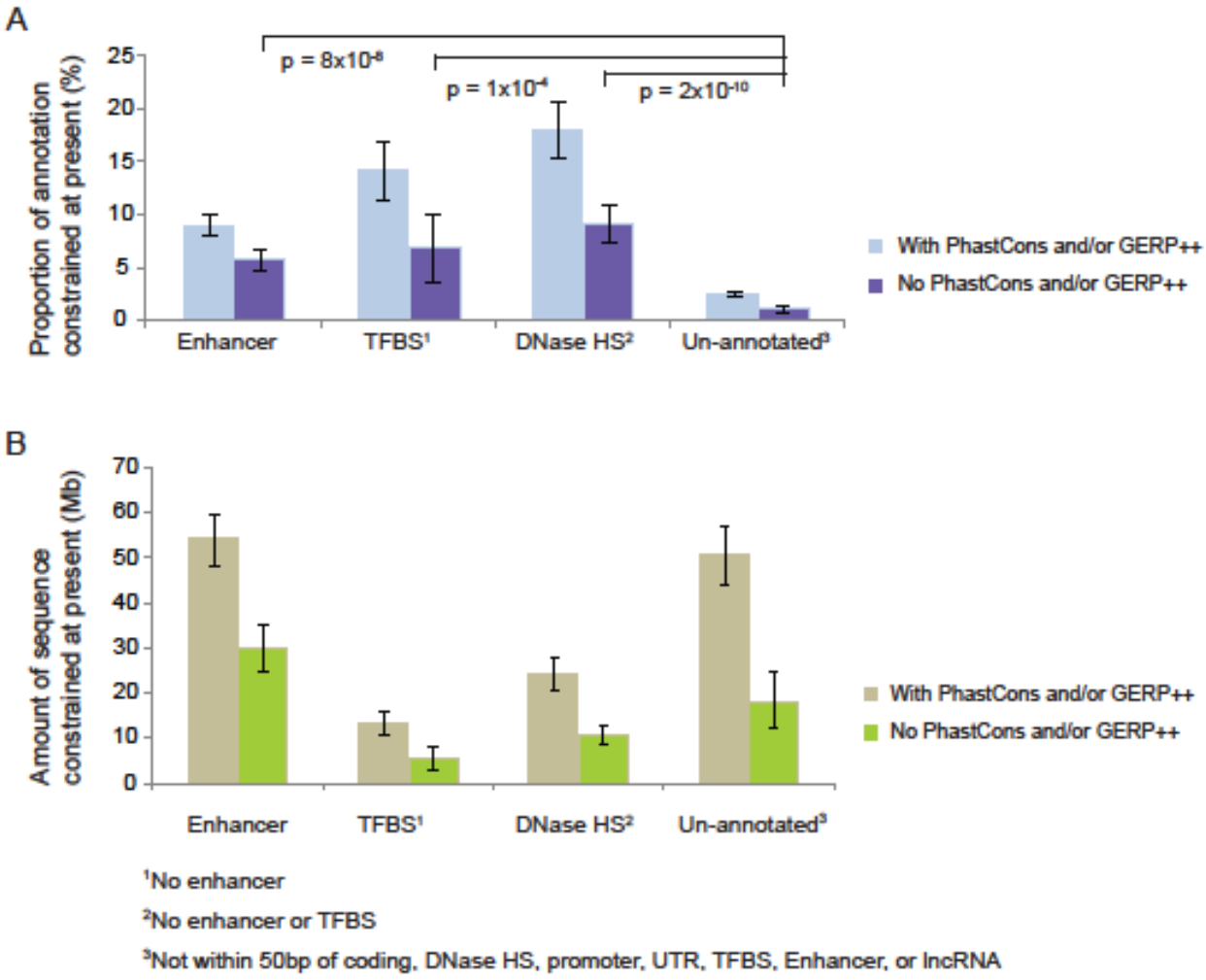

Figure 4.5: The intersection of constrained sequence with pan-mammalian conserved sequences and selected biochemical annotations. A, B. The proportions and quantities of constrained sequence at the present for different types of biochemically annotated and un-annotated sequences, with and without PhastCons or GERP++ conserved elements, estimated using linear extrapolations (Materials and methods). The un-annotated not PhastCons/GERP++ sequence was depleted of NIM1 constrained sequence compared to the biochemically annotated not PhastCons/GERP++ sequence, implying that the NIM1 had power to detect functional lineage-specific constrained sequence.

In summary, the observation of significant enrichment of NIM1-constrained sequence within putatively functional sequence that did not show evidence for pan-mammalian constraint is incompatible with the hypothesis that functional sequence does not exhibit turnover to a significant degree. I therefore concluded that the signature of functional turnover identified by NIM1 (**Figure 4.3**) is not driven by its lack of specificity for low species divergence, but instead indicates that turnover of functional sequence is pervasive.

For completeness, I showed more directly that NIM1 is highly specific also for divergent species pairs. Over larger evolutionary distances, virtually all sequence identified as being constrained by NIM1 was also found by PhastCons and GERP++; for example, 97% of NIM1 constrained sequence was estimated to be conserved by PhastCons and/or GERP++ between the divergent genomes of human and mouse. Since PhastCons and GERP++ are substitution methods, this also suggests that there was little difference between the genomic regions that are constrained with respect to indels and those that are constrained with respect to point mutations. Moreover, this shows that the additional constrained sequence identified by the NIM1 is likely to represent lineage-specific constraint, rather than indel-specific constraint.

4.4.3 Contrasting estimates for turnover rates in coding and noncoding sequence

I next estimated, using the NIM1 and NIM2, the fraction of constrained sequence within coding and noncoding sequences (**Materials and methods**). Within protein coding sequence the models indicated, as expected, that selective constraint is pervasive (**Figure 4.3**). The NIM1 estimated that 80–88% of human or mouse annotated coding sequence has been under selective constraint with respect to indels across eutherian evolution (**Figure 4.6**). Lower proportions (68–71%) of dog annotated coding sequence were inferred to have been under constraint (**Figure 4.6**), likely reflecting the lower abundance of transcriptional evidence and thus the less refined gene annotations for the dog genome (Derrien et al. 2012). While NIM2 showed qualitatively the same trends, its estimates of protein coding constraint under this

model were somewhat lower (72–81% for the human or mouse genome, 64% for the dog genome). Results from application of the two models were consistent with the accepted notion that most protein coding gene sequence is highly conserved (Nielsen et al. 2007). Since NIM1 consistently showed greater sensitivity in identifying protein coding sequence than NIM2 (**Figure 4.6**), I focused on NIM1 estimates for all subsequent analyses.

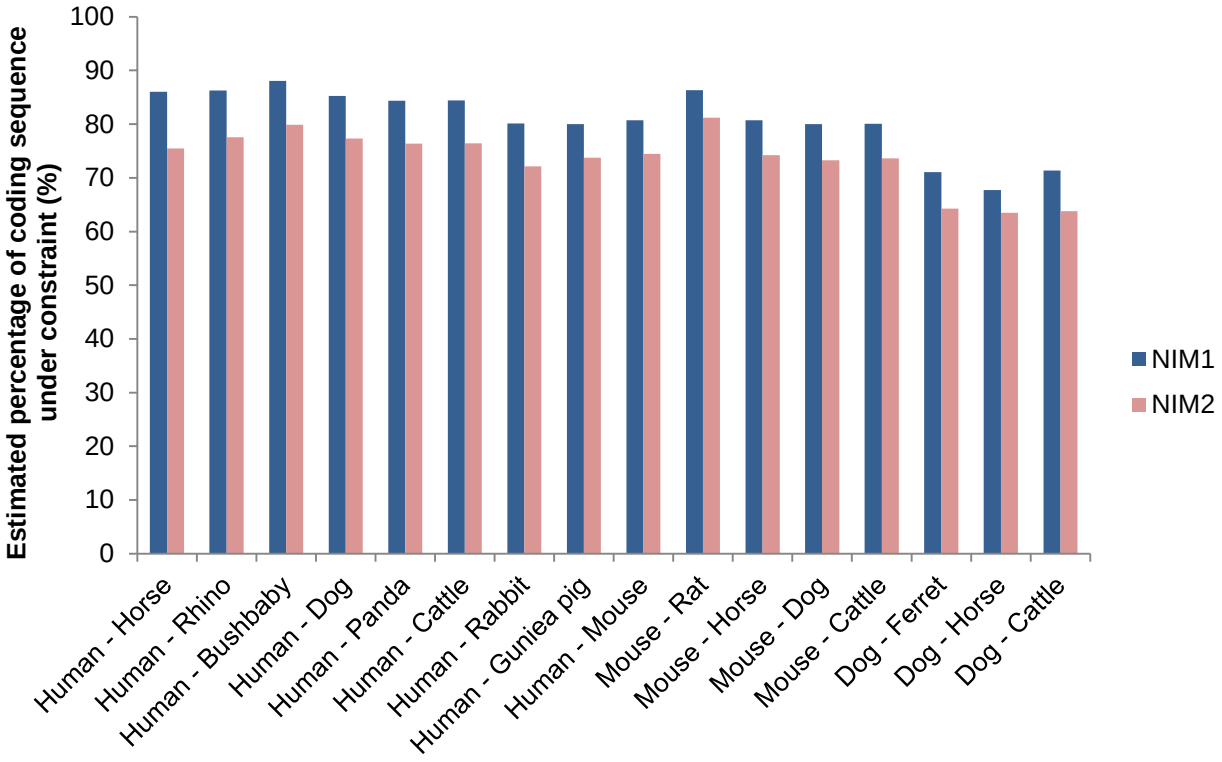

Figure 4.6: The proportions of coding sequence that were inferred to be under constraint by NIM1 or NIM2 for different pairs of eutherian genomes. NIM1 consistently identified a greater percentage of coding sequence as being constrained compared to NIM2.

In contrast to protein coding sequence, estimates for the extent of constraint in noncoding sequence showed a pronounced drop-off with increasing divergence (orange filled circles in **Figure 4.3**); an observation compatible with turnover occurring predominantly within the noncoding functional fraction of the genome. When applying the time-homogeneous turnover model to these data, I estimated the turnover rate parameter *b* for noncoding sequence at 2.48 turnover events per neutral substitution (2.26–2.71, 95% confidence interval). Equivalently, the turnover half life $d_{1/2}$ is estimated at 0.28 (0.25–0.31) in units corresponding to one expected neutral substitution per site; in natural units this is equivalent to 127My (116–139My). The present estimate represents a slower turnover rate than a previous estimate of $d_{1/2}$ = 0.19 (86My) made by Ponting and Hardison (2011) with data from Meader et al. (2010).

Ponting and Hardison (2011) assumed that coding sequence exhibits no turnover. Here I observed a very low yet significantly non-zero rate of turnover in coding sequence, b = 0.24 (0.14–0.33) events per neutral substitution, corresponding to $d_{1/2}$ = 2.9 (2.1–5.0), or in natural units 1300My (950–2250My). These estimates represent an average across the undoubtedly variable rates of turnover for different classes and types of protein coding gene sequence. Nevertheless, under this simple model, I found that protein coding sequence is relatively evolutionarily stable, showing long-term conservation, while present-day constrained noncoding sequence is less stable, being relatively rapidly gained and lost in a lineage-specific manner.

4.4.4 Constraint and turnover among human functional element classes

I investigated in more detail whether various classes of functional element, identified in human, primarily by the ENCODE project (Dunham et al. 2012), exhibit contrasting levels of constraint, and whether these constrained element classes show a propensity to turn over at different rates. Of the functional classes I considered, promoters, UTRs and DNAse HSs

showed intermediate levels of turnover, while TFBSs, enhancers, lncRNA sequence and un-annotated sequences (defined as not within 50bp of ENCODE DNAse HSs, TFBSs loci, lncRNAs from (Hangauer et al. 2013), Ensembl coding sequence, or UTRs) showed relatively high levels of turnover (**Figure 4.7**, **Figure 4.8, Figure 4.9**). The fraction of sequence that the model inferred to be under constraint also varied across these categories, with intermediate fractions inferred for UTRs, DNAse HSs and TFBSs, lower fractions for lncRNAs and enhancers, while un-annotated sequence showed the lowest inferred fractions of conserved elements (**Figure 4.7**). Constrained sequence in this category may represent lineage-specific, novel candidate functional sequences that were not identified by the ENCODE project, for instance because of their function in tissues or developmental stages not investigated by ENCODE. Finally, transposable element-derived sequences show very small amounts of constraint, and my methods have little power to detect turnover in this class due its low levels of constraint.

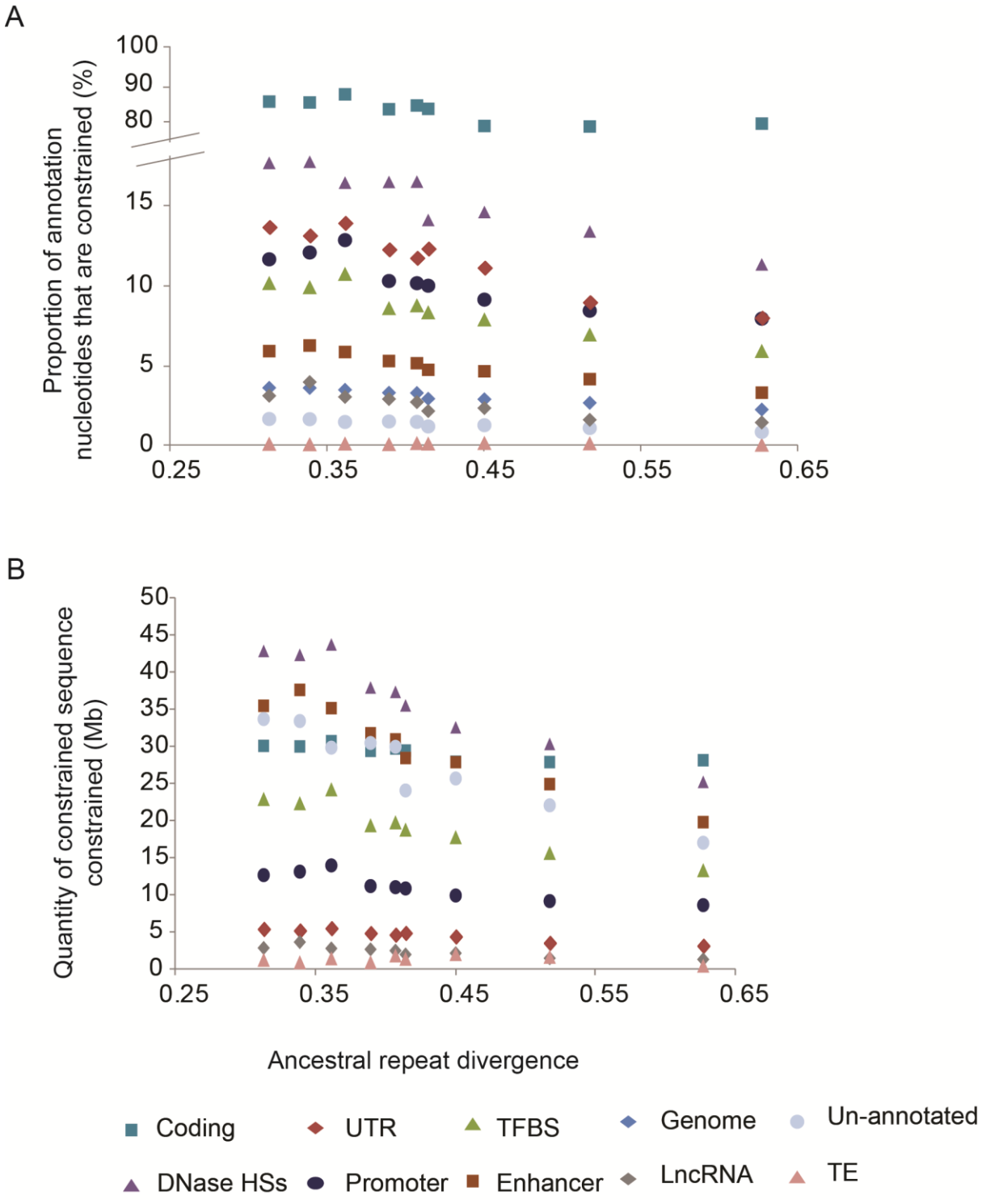

Figure 4.7: Constraint and turnover for different classes of human functional element. A. The proportion, and B. The quantity, of annotation bases inferred as being constrained plotted against divergence.

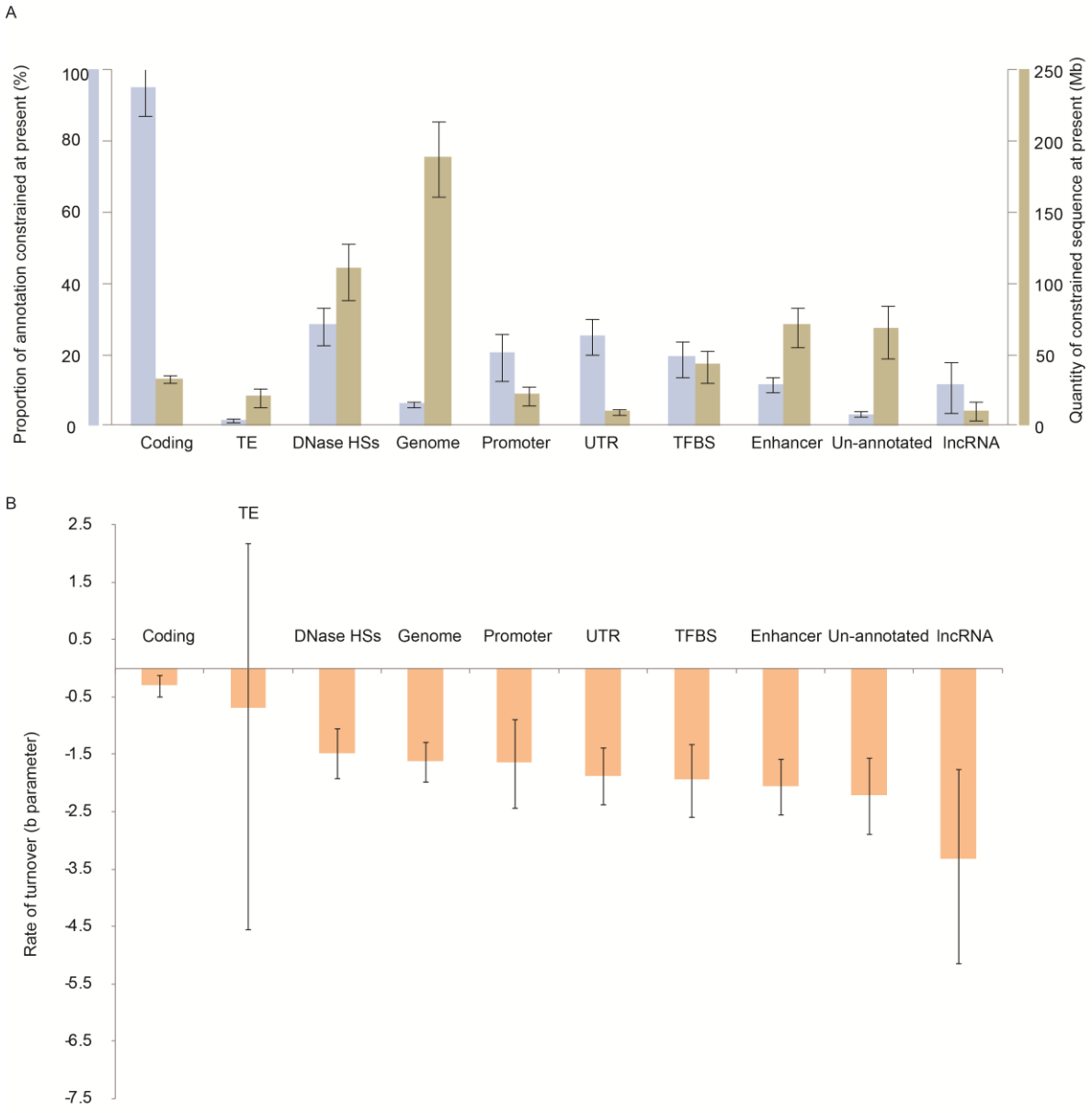

Figure 4.8: A. The total proportions and quantities of constrained sequence estimated for the present day by extrapolation for different element types. B. The estimated rate of turnover (b parameter) for different types of constrained element. Error bars show the 95% confidence intervals.

A

	Coding	Repeat	DNase HSs	Genome	Promoter	UTR	TFBS	Enhancer	No annotation	lncRNA
Coding	X	0.360	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Repeat	0.360	X	0.784	0.847	0.822	0.146	0.977	0.960	0.872	0.547
DNase HSs	0.000	0.784	X	0.111	0.857	0.179	0.137	0.005	0.000	0.004
Genome	0.000	0.847	0.111	X	0.847	0.252	0.162	0.003	0.004	0.007
Promoter	0.000	0.822	0.857	0.749	X	0.1459	0.019	0.049	0.104	0.008
UTR	0.000	0.146	0.179	0.252	0.1459	X	0.960	0.616	0.001	0.002
TFBS	0.000	0.977	0.137	0.162	0.019	0.960	X	0.456	0.335	0.031
Enhancer	0.000	0.960	0.005	0.003	0.049	0.616	0.456	X	0.364	0.021
No annotation	0.000	0.872	0.000	0.004	0.104	0.459	0.335	0.364	X	0.045
lncRNA	0.000	0.547	0.004	0.007	0.008	0.002	0.031	0.021	0.045	X

B

	Coding	Repeat	DNase HSs	Genome	Promoter	UTR	TFBS	Enhancer	No annotation	lncRNA
Coding	X	0.263	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Repeat	0.263	X	0.940	0.818	0.778	0.978	0.973	0.952	0.850	0.425
DNase HSs	0.000	0.940	X	0.193	0.258	0.730	0.781	0.300	0.120	0.007
Genome	0.000	0.818	0.193	X	0.755	0.130	0.174	0.024	0.014	0.002
Promoter	0.000	0.778	0.258	0.755	X	0.188	0.216	0.073	0.034	0.003
UTR	0.000	0.978	0.730	0.130	0.188	X	0.968	0.536	0.222	0.011
TFBS	0.000	0.973	0.781	0.174	0.216	0.968	X	0.536	0.230	0.011
Enhancer	0.000	0.952	0.300	0.024	0.073	0.536	0.536	X	0.441	0.018
No annotation	0.000	0.850	0.120	0.014	0.034	0.222	0.230	0.441	X	0.056
lncRNA	0.000	0.425	0.007	0.002	0.011	0.011	0.011	0.018	0.056	X

C

	Coding	Repeat	DNase HSs	Genome	Promoter	UTR	TFBS	Enhancer	No annotation	lncRNA
Coding	X	0.461	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.001
Repeat	0.461	X	0.140	0.408	0.185	0.187	0.118	0.110	0.245	0.237
DNase HSs	0.000	0.140	X	0.436	0.211	0.877	0.722	0.261	0.138	0.383
Genome	0.000	0.408	0.436	X	0.845	0.880	0.561	0.223	0.014	0.745
Promoter	0.000	0.185	0.211	0.845	X	0.237	0.101	0.052	0.258	0.060
UTR	0.000	0.187	0.877	0.880	0.237	X	0.955	0.887	0.863	0.008
TFBS	0.000	0.118	0.722	0.561	0.101	0.955	X	0.580	0.448	0.281
Enhancer	0.000	0.110	0.261	0.223	0.052	0.887	0.580	X	0.475	0.410
No annotation	0.000	0.245	0.138	0.014	0.258	0.863	0.448	0.475	X	0.818
lncRNA	0.001	0.237	0.383	0.745	0.060	0.008	0.281	0.410	0.818	X

Turnover rate of annotation in row is faster/slower than the annotation in column

	p>0.1	0.05< p <0.1	0.01 < p <0.05	0.001 < p < 0.01	0 < p < 0.001
Faster					
Slower					

Figure 4.9: Comparisons of the rates of turnover of different constrained element types. A. P-values are computed by considering the ratio of observations, which under the hypothesis that the turnover rate is equal, should fit a model with b = 0. B. P-values are computed using a likelihood ratio test to compare a model where the b parameter is shared between the two annotations to one where b is independent for the annotations. C. The same computation as B. except that the length of the NIM1 95% confidence interval was used to calculate the weight for each data point.

4.4.5 Distribution of functional classes in present-day functional DNA

I next examined how constrained sequence in the human genome is distributed cumulatively for selected categories of functional element. I did this by fitting the model of functional turnover to the observed data and extrapolating to the present day. In this way I also inferred the reciprocal quantities of sequence that, when comparing to another species or human ancestor at a particular divergence, are presently functional in human yet have lost (or not gained) constraint in the lineage leading to the ancestor or other species, under the assumption that the total quantity of functional sequence in genomes has remained constant over time and across species (**Figure 4.10**). 8.6Mb (26%) of constrained coding nucleotides was estimated to have lost constraint (and has thus turned over) since the divergence of humans from monotremes approximately 228 million year ago (AR divergence = 1), while 200.0Mb (79%) of the constrained noncoding human genome was inferred to have lost constraint over the same period. DNAse HSs covered more indel constrained sequence at all divergence ranges than all other annotated noncoding sequence combined, implying that DNAse HSs are an abundant and informative biochemical marker of functionality outside protein coding regions. Enhancers also showed a marked contribution towards the constrained human genome, while TFBSs, promoters, UTRs and lncRNAs contributed considerably less sequence once their overlap with other annotations was removed. In **Figure 4.10** I summed the quantities of constrained sequence estimated from independent NIM1 runs for different annotation types. For consistency this approach required mutually exclusive annotation sets, in contrast to those used in **Figure 4.7** and **Figure 4.8**, making the results not directly comparable. In addition, the un-annotated sequence here was required to not extend to 50bp either side of each annotation. Overlaps between the major different annotations are shown in **Figure 4.11**.

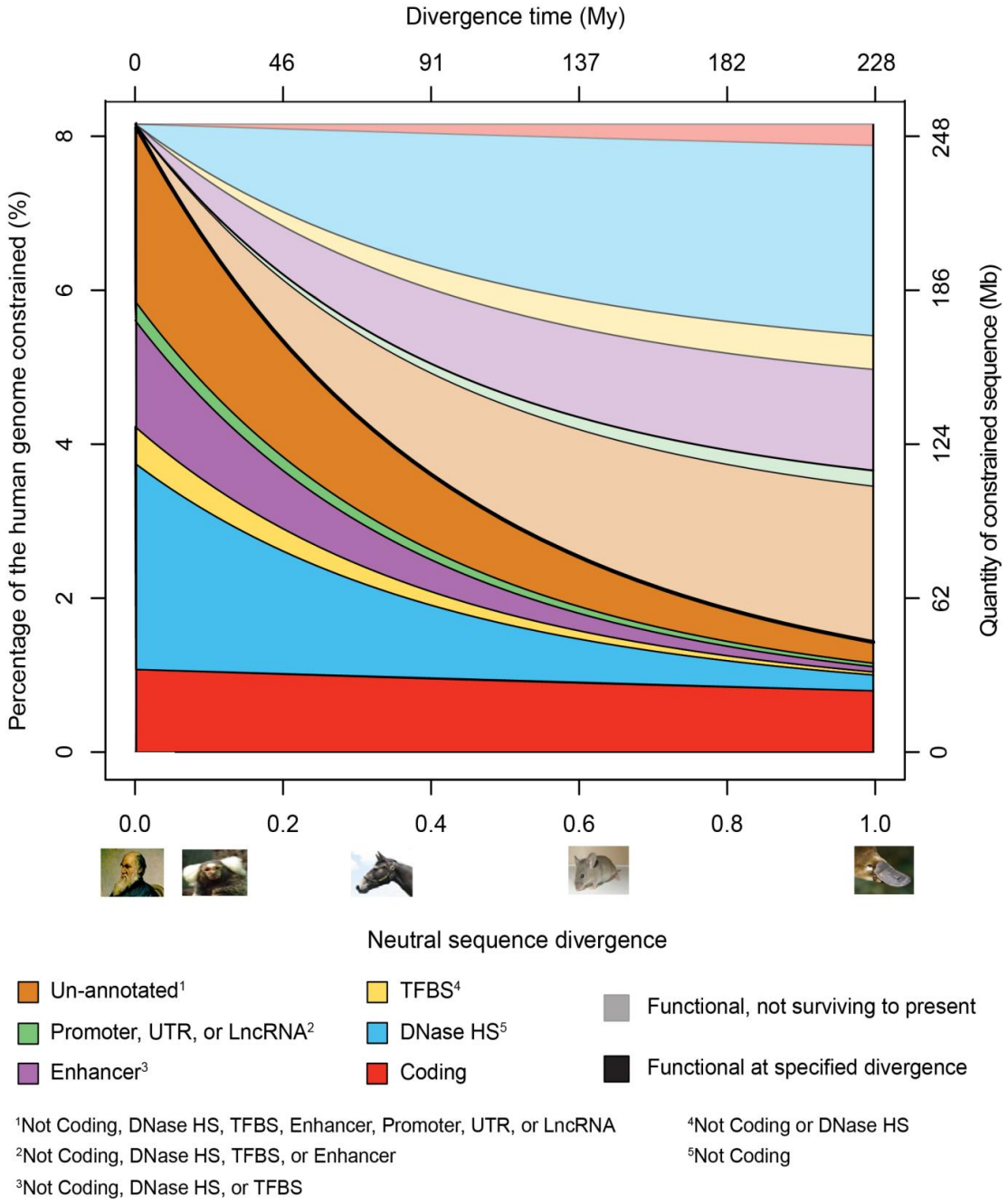

Figure 4.10: Model-based inference of turnover by functional class. Schematic summary of how constrained sequence has been retained (deeper colours) or gained (lighter colours) in the human lineage over time (or its proxy, neutral sequence divergence, X-axis) and how it has been distributed across various categories of functional element. In addition to showing the reduced quantity of preserved constrained sequence with increasing divergence, I infer the reciprocal quantity of sequence that is assumed to have been gained over human lineage evolution.

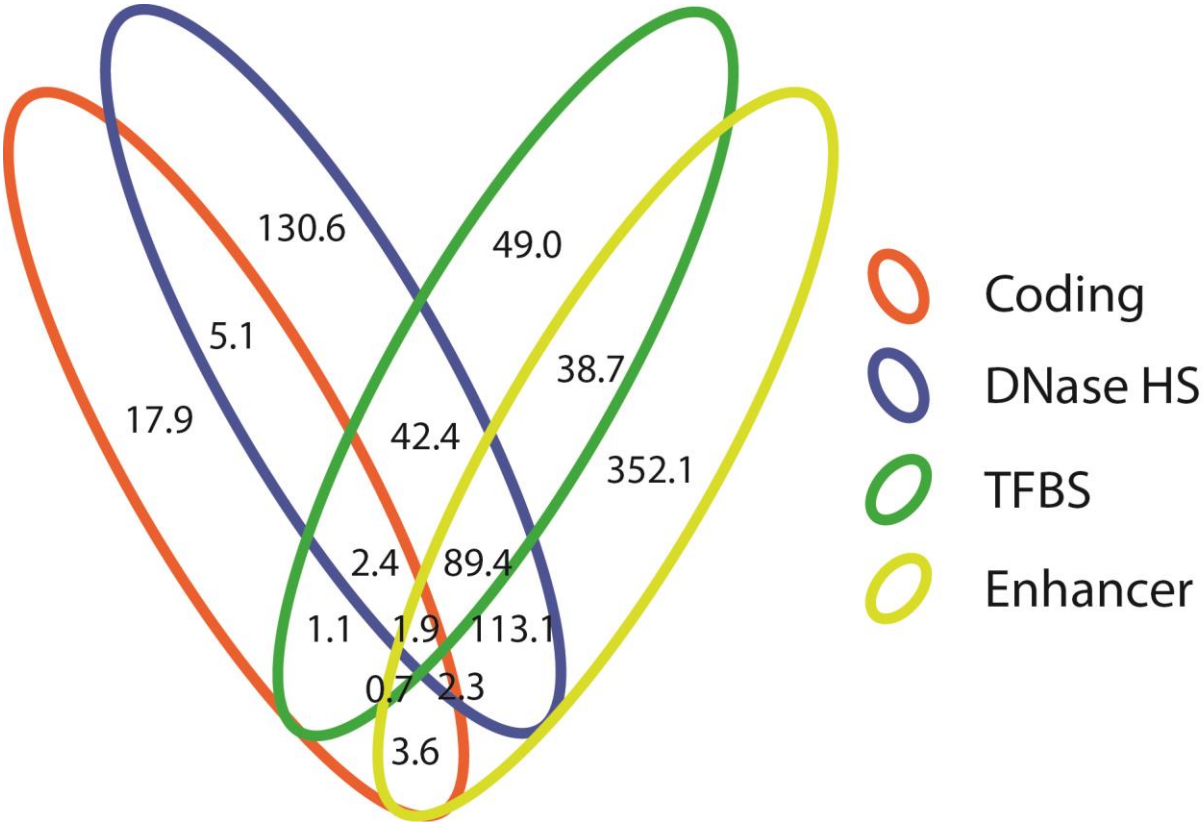

Figure 4.11: The overlap between different human functional annotations in megabases. The considerable overlap between some annotations has the consequence that evidence of sequence constraint on one type of annotation may instead be attributable to a different annotation that covers the same inter-gap segment.

4.4.6 7.1–9.2% of human genomes are constrained at present

If I make the simplifying assumption that the exponential decay model of functional sequence applies outside of the range of divergences I examined, then by extrapolating back to zero divergence I can estimate the total proportion of human genomes that is under present-day purifying selection with respect to indels. I performed this extrapolation across different annotation sets (**Table 4.1**). Although there is some variation in these estimates, I quote the estimate derived separately across multiple different annotation categories, namely coding sequence, DNase HSs, TFBSs, enhancers, un-annotated sequences, and other sequences (the latter consisting of promoter, UTR and lncRNA sequences). This is because this estimate

allows the rate of turnover to vary across each annotation type, and thus is likely to be more accurate than estimates that assume a single rate of turnover across the whole genome, or the whole noncoding genome. I therefore estimated that 8.2% (253Mb), 7.1–9.2% (220–286Mb), of the human genome is under contemporaneous purifying selection with respect to indels.

Table 4.1: The quantities of constrained sequence at present estimated by different methods. The annotations are mutually exclusive sets as in Figure 4.10.

Annotation sets that extrapolations are derived from	Proportion of the human genome constrained at present (%; 95% confidence interval)
Genome	6.1 (5.3–7.0)
Coding, noncoding	6.5 (5.4–7.6)
Coding, DNase HSs, TFBS, Enhancer, Promoter/UTR/lncRNA, Un-annotated	8.2 (7.1–9.2)

4.4.7 Variation in constraint and turnover between different annotations sets

The results show the genome-wide variation in the constraint and turnover of different categories of human functional element. However, there is no one definitive set of annotations of each element type: for example the TFBSs called will depend on the TF and cell-type examined in the ChiP-seq experiment. To explore this I examined the constraint and turnover across a high-confidence set of TFBSs (Arbiza et al. 2013), and a set of lncRNAs called by ENCODE rather than the set (Hangauer et al. 2013) I used in the main analysis.

The high-confidence TFBSs were a sub-set of the ENCODE sets, defined by Arbiza et al. (2013) with a pipeline consisting of 3 steps: (1) TFBSs peak calls from the ENCODE experiments were obtained, discarding peak calls from time-course experiments, control experiments, cell types that were treated with chemicals, and peaks residing on the sex chromosomes. (2) de novo motif discovery was then applied to the peak calls with MEME (Bailey and Elkan 1994) to identify a single high quality motif for each TF. (3) This consensus motif was then used to search the entire set of TFBS peaks for corresponding binding sites, and highly degenerate positions at the edges of motifs were removed. After

removing TFs that had less than 500 predicted binding sites, the final resulting set of TFBSs was across 78 different TFs.

For these high-confidence TFBS, I found a significantly higher level of constraint ($p = 6.6\times10^{-4}$ for differences in present-day constraint), with over three times the proportion of bases inferred to be subject to purifying selection compared to the larger ENCODE TFBS set. I estimated that 28.6–43.8% (depending on divergence) of high-confidence TFBS bases are under purifying selection with respect to indels (**Figure 4.12**), which is quite consistent with an estimate that approximately 32% of bases are subject to purifying selection with respect to substitutions (Arbiza et al. 2013). I also found a lower rate of turnover for the high-confidence TFBSs than the ENCODE TFBSs ($p = 0.07$; computing p-value as in **Figure 4.9A**). A lower rate of turnover for the high-confidence TFBSs is unsurprising given the higher levels of constraint for this set.

Comparing the large set of tens of thousands of lncRNAs predicted from RNA-sequencing data that I used in the main analysis (Hangauer et al. 2013) with a smaller set of lncRNAs produced by ENCODE, I found higher levels of constraint for the ENCODE set. The levels of turnover for the original lncRNA set were the highest of all the genomic annotations, but the ENCODE lncRNAs appear to turnover even faster (**Figure 4.13**), although not significantly so at a standard confidence level ($p = 0.31$). This provides evidence that lncRNAs may be the most rapidly turning over of all the basic functional element types. However, given the small size of the ENCODE data set, 7.5Mb compared to 92Mb for the other set, the data set is rather noisy, as reflected in the wide confidence interval for the rate of turnover (**Figure 4.13B**).

These results show that the levels of constraint and rates of turnover vary significantly among different annotation sets for the same types of functional elements, and thus that my results cannot be completely generalised to make categorical statements concerning a particular type

of functional element. This is likely to be the case for all studies until annotation sets become much more standardised, for example to the extent that they are for protein coding gene sets in the human genome. However, these results do support the notion that more robustly defined annotation sets show higher levels of constraint than larger less stringently defined sets. Therefore, this shows that conservation metrics are informative proxies for functionality, and that conservation scores are likely to be a good tool to help prioritise the importance of candidate loci implicated in a particular biological process, as is already implemented in packages such as PolyPhen-2 (Adzhubei et al. 2010).

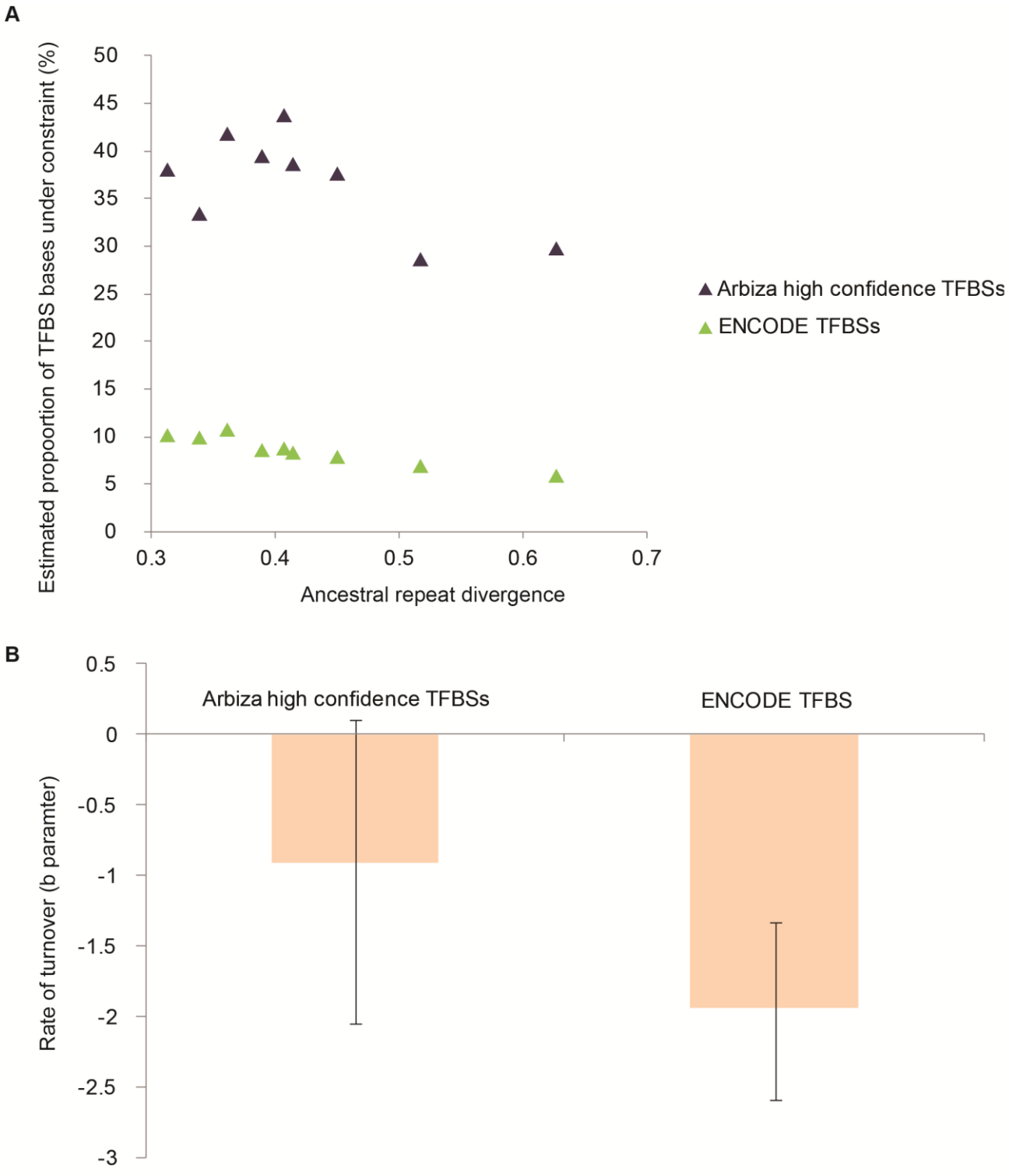

Figure 4.12: The conservation and turnover of ENCODE TFBSs and a set from Arbiza et al. (2013). A. The proportion of TFBS bases identified as constrained by NIM1 plotted against the divergence. B. The estimated rates of turnover for the two different TFBS data sets.

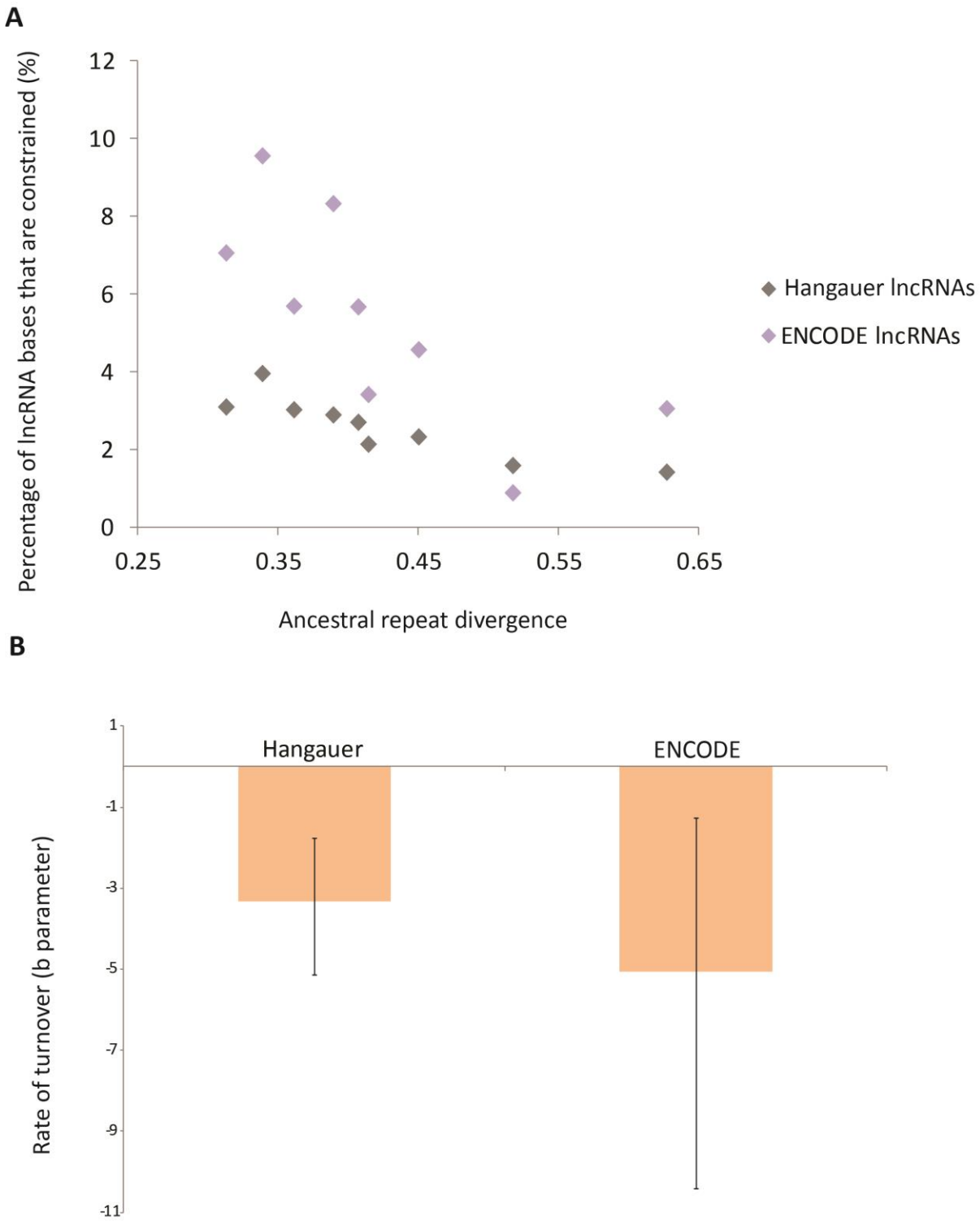

Figure 4.13: The conservation and turnover of ENCODE lncRNAs and a set from Hangauer (2013). A. The proportion of lncRNA bases identified as constrained by NIM1 plotted against the divergence. B. The estimated rates of turnover of the two different lncRNA data sets.

4.5 Discussion

The fraction of the human genome sequence in which mutations are preferentially purged owing to their deleterious effect has remained contentious ever since the first estimate was made (Waterston et al. 2002). At that time it was not well appreciated that the amount of human constrained sequence that is also constrained in mouse is a minority of all human constrained sequence, owing to the relatively rapid gain and loss of functional sequence in their two lineages since their last common ancestor.

I found that sequence with experimental evidence for biochemical activity, but lacking evidence for pan-mammalian conservation, is enriched with NIM1-constrained sequence, and I argue that this is incompatible with the notion of a technical artefact causing the signature of turnover. Extensive simulations described in **Chapter 3** indicating that estimates of constrained sequence are consistent across the divergence range I investigate support this conclusion. My estimate that 7.1–9.2% of human genomes is subject to contemporaneous selective constraint considerably exceeds previous estimates and falls short of others (Ponting and Hardison 2011; Ponting et al. 2011). I have demonstrated that the indel method as adopted by Meader et al. (2010) led to estimates for specific species pairs that were inflated, in large part owing to inaccuracies in whole genome alignments, errors which also will have affected the accuracy of other estimates based on comparisons between species' genomes. In principle, polymorphism data within a single species could be used instead to estimate the proportion of the genome subject to negative selection, but this approach would not indicate how much functional sequence turns over during evolution. Furthermore, while polymorphism data have been used to this end, this approach is technically highly challenging, and results have so far been controversial (Ward and Kellis 2012; Green and Ewing 2013; Ward and Kellis 2013). Further estimates (Thomas et al. 2003; Siepel et al. 2005; Lunter et al. 2006) are lower because they, by design, were not sensitive to lineage-

specific constrained sequence. Nevertheless, my current estimates have associated caveats. While my results showed that turnover is a real and substantial effect, simulations show that NIM1 underestimates the true amount of mutually constrained sequence to an extent that shows some dependence on the divergence. Although simulations and theory indicate that point estimates of constraint remain conservative, the possibility of an upward bias in the inferred *rate* of turnover cannot be excluded, which in turn could lead to upwardly biased extrapolations. In addition, the assumptions of the turnover model, in particular that all elements within a class are subject to the same rate of turnover, clearly are only approximately valid. These potential sources of error are not reflected in the confidence estimates (**Table 4.1**).

As expected, turnover has occurred least in protein coding sequence, and instead has been most concentrated on noncoding sequence (**Figure 4.3, Figure 4.10**). For example, of the 43.5Mb of sequence annotated by the ENCODE project as being within a human TFBS peak and that I found to be constrained (19.3% of the total size of ENCODE TFBS peaks), only a third (30.6%; 13.3Mb) was identified by the NIM1 as being constrained in both human and mouse. A slightly higher proportion (45.6%; 19.8Mb) was constrained in human and dog, presumably reflecting the divergence between these species. These estimates are in good agreement with previous experimental findings, for instance 23–41% of TF binding events have been found to be conserved across human, dog and mouse for four liver TFs (Odom et al. 2007), while for two additional liver TFs, 7–14% of TF binding events are shared between human and mouse, and 15–20% between human and dog (Schmidt et al. 2010).

The phenomenon of turnover is well supported by both anecdotal evidence (Ludwig et al. 2000; Odom et al. 2007; Schmidt et al. 2010) and by broader studies of particular classes of elements, mostly TFBSs and enhancer elements (Dermitzakis and Clark 2002; Moses et al. 2006; Doniger and Fay 2007).

LncRNAs, the class of functional element I inferred to turnover fastest, have also been previously shown to evolve rapidly. Amongst a large set of approximately 10,000 human lncRNAs, the majority were found to be either primate-specific or so rapidly evolving that no homology was detected beyond primates, although 19% of lncRNA families were inferred to be conserved over more than 90My (Necsulea et al. 2014), approximating human – mouse divergence. Additionally, only 5.1% of zebra fish lncRNAs have an identifiable mammalian ortholog (Ulitsky et al. 2011), and lncRNAs show very rapid turnover at the level of the transcript (Kutter et al. 2012). The contribution of the present study was to quantify this phenomenon both genome-wide and across functional categories.

What my approach cannot clarify is to what extent the observed turnover at the sequence level amounts to different sequences encoding equivalent function (Ludwig et al. 2000; Dermitzakis and Clark 2002), or species-specific functional change (Doniger and Fay 2007; Lowe et al. 2011; Ward and Kellis 2012). Lineage-specific functional sequence can be functionally redundant sequence that has equivalent roles in different lineages, implying a model of compensatory evolution where the plasticity of regulatory networks allows them to be rewired at the sequence level without resulting in overt phenotypic changes. Under this scenario, conservation of function can be maintained without sequence conservation (Dermitzakis and Clark 2002). Several lines of evidence, from both anecdotal (Ludwig et al. 2000) and broader (Dermitzakis and Clark 2002; Doniger and Fay 2007) studies of TFBSs, indicate that a large fraction of sequence changes involving TFBSs preserve function. For example, some deeply conserved transcription factors have species-specific binding sites in the vicinity of orthologous genes (Odom et al. 2007; Schmidt et al. 2010) implying that despite their sequence divergence, the different DNA binding sites confer equivalent functions (on orthologous genes) in different lineages. Comprehensive studies of human and mouse embryonic heart enhancers found these to be weakly conserved (Blow et al. 2010;

May et al. 2012), despite human enhancer sequences largely driving expected tissue-specific expression in mouse embryonic heart tissue (May et al. 2012). Another study found that two mammalian hypothalamic enhancers have no homolog across non-mammalian vertebrates, yet are still able to drive specific expression patterns in zebrafish neurons (Domene et al. 2013). These findings are consistent with gene expression evolution being shaped predominantly by stabilizing selection on the expression level (Brawand et al. 2011), which on the sequence level may involve an interplay between drift and weak positive selection (Chaix et al. 2008).

However, not all TFBS turnover events are neutral or nearly neutral on the level of gene expression, and the fraction of such events that change gene expression may be substantial (Doniger and Fay 2007). More generally, lineage-specific sequence is clearly a likely substrate for lineage-specific biology (Lowe et al. 2011; Ward and Kellis 2012). Such positively selected lineage-specific functional sequence presumably underlies lineage-specific traits, such as the unique cognitive abilities of great apes or the enhanced chemosensation of some rodents. However, this sequence need not be abundant; just 2Mb of sequence could produce 250,000 eight base pair binding-sites, equating to more than ten sites per protein coding gene, conceivably enough to underlie lineage-specific traits (Green and Ewing 2013). Furthermore, adaptations to pre-existing functional sequence remain an alternative plausible mode for creating species-specific change (Ames and Lovell 2011).

Nevertheless, the sheer ubiquity of sequence turnover, and the clear potential for substantial regulatory change resulting from it, suggests that many aspects of noncoding human biology will not be fully recapitulated by orthologous sequence in eutherian model organisms, including mouse. Thus, my findings could provide a quantitative basis for assessing the relevance of model organisms to specific questions of human biology. So, even though this research implies that many model organisms, even mammalian ones, will be of limited use to

inform human noncoding biology via homology, these model organisms can still be invaluable for understanding general biological mechanisms, including those that involve noncoding sequences.

Chapter 5: Patterns of sequence constraint and turnover are similar across both avian and mammalian lineages

5.1 Abstract

Using evolutionary constraint as a proxy for functionality I estimated that approximately 8% of mammalian genomes is functional, a far larger proportion than the protein coding component, but still implying that the vast majority of the genome is functionless. Moreover, the quantity of functional sequence that is shared between a given pair of mammalian species is strongly negatively correlated with the divergence between the two species, consistent with the notion that functional sequence turns over rapidly as it is lost and gained over mammalian evolution. Here I examine the constraint and turnover of sequence in avian genomes, and compare my results to those I previously observed in mammalian genomes. I estimate the quantity of constrained sequence ($\alpha_{selIndel}$) between pairs of avian genomes, and examine how this varies with genomic features such as the protein coding complement, G+C content, and chromosome size. I estimate that 10.3–16.8% of avian genomes are subject to present-day purifying selection. This is a larger proportion than estimated for the mammalian genome, and in terms of absolute quantities of constrained sequence, suggests that mammalian genomes may contain more functional sequence than avian ones, although this is unclear with the current data. The trends of turnover are similar between birds and mammals. This demonstrates that the rapid turnover of functional sequence is not a specific feature of the mammalian lineage and is likely to be a ubiquitous aspect of vertebrate evolution due to regulatory networks being rewired over relatively short evolutionary time-scales.

5.2 Introduction

There is increasing evidence that the turnover of functional sequence is a pervasive biological phenomenon in mammals (**Chapter 4**). I understand turnover to mean the loss or gain of purifying selection at a particular locus of the genome, when changes in the physical or genetic environment, or mutations at the locus itself, cause the locus to switch from being functional to being non-functional or vice versa. Comparative genomic analyses have shown that the quantity of constrained sequence is strongly negatively correlated with divergence between the pairs of mammalian species (Smith et al. 2004; Meader et al. 2010), implying that there is an abundance of lineage-specific constrained sequence that has arisen since the divergence of eutherian mammals from earliest their common ancestor.

Experimental studies have demonstrated that specific classes of functional element turnover rapidly. For example, transcription factor binding sites (TFBSs) of highly conserved TFs are often species-specific (Odom et al. 2007; Schmidt et al. 2010), and 81% of lncRNA families from a large catalogue of human lncRNAs were found to be primate-specific (Necsulea et al. 2014). In **Chapter 4**, I provided extensive support for the notion that protein coding sequence is widely conserved, while the turnover of functional noncoding sequence is rapid. In addition, I compared how turnover varies between different categories of annotated elements, predominantly using data from the ENCODE project (Dunham et al. 2012).

While the turnover of functional sequence has been extensively documented in the mammalian lineage, evidence of functional turnover has also been found in other phylogenetic groups including in birds, flies, yeast, plants, and bacteria (Mustonen and Lassig 2005; Moses et al. 2006; Doniger and Fay 2007; Kunstner et al. 2011a; Hupalo and Kern 2013), implying that functional turnover is widespread across the tree of life. The avian lineage is an important lineage in which to examine patterns of molecular evolution due to their close phylogenetic relationship to mammals and the economic importance of birds.

Furthermore, avian genomes are much smaller than mammalian genomes, so it is not clear that the genome-wide patterns of evolution of functional sequence in birds will be found to parallel those observed in mammals. Avian genomes are only approximately 1Gb in size, compared to mammalian genomes that are typically around 3Gb, and non-avian reptilian genomes that are intermediate in size at about 2Gb (Alfoldi et al. 2011). The relatively small size of avian genomes is due to a contraction of their genome from the ancestral state, since the common ancestor of birds and mammals probably had a large genome similar to the size of extant mammalian genomes (Organ et al. 2007). This genome size contraction in birds could purely reflect a streamlining of the non-functional genome, or the contraction could also occur across the functional genome. Under the first scenario avian and mammalian genomes would be expected to contain similar quantities of functional sequence, while under the second scenario avian genomes would consist of less functional sequence. Identifying patterns of turnover in avian genomes would facilitate meaningful interpretation of the mammalian results as it gives a contextual reference point from which to understand the trends observed in mammals.

Avian genomes have a karyotype characterised by large macrochromosomes and small microchromosomes; some studies choose to create an additional intermediate category of moderately sized chromosomes (International Chicken Genome Sequencing Consortium 2004). This avian karyotype is predicted to have evolved 150–200My by the fission of larger chromosomes (Burt 2002). These chromosomal classifications are correlated with many genomic features: chromosome size is negatively correlated with gene density, recombination rate, and G+C content, CpG island density (Burt 2002), substitution rates, and evolutionary constraint in protein coding regions (Axelsson et al. 2005). This makes avian genomes suitable for examining fundamental biological patterns of molecular evolution, including potentially recombination associated processes such as biased-gene conversion and Hill-

Robertson interference (Hill and Robertson 1966; Duret and Galtier 2009).

Only one previous study has examined trends of constrained sequence turnover across the avian lineage. Kunstner et al. (2011a) found sequences were subject to purifying selection in the untranslated regions (UTRs) flanking avian genes, and that more closely related avian species share more constrained flanking sequence than more divergent species. This could be explained by the turnover of UTR sequences across avian evolution. However, the study only demonstrated that there is a larger quantity of constrained UTR sequence between zebra finch, European crow, pied flycatcher and blue tit (all songbirds), than there is between zebra finch and the more distantly related chicken. Therefore, it is unclear if the turnover trend holds true across avian evolution. Furthermore, their results could be specific to UTRs and therefore not reflect a genome-wide trend.

Here I examine patterns of constraint and turnover across avian evolution from the ancestral galliform genome, as represented by chicken and turkey genomes that shared a common ancestor about 45My ago (Pereira and Baker 2006), to the divergence of the galliformes from the passerine song birds approximately 100My ago (White et al. 2011), with the passerine lineage being represented by the Darwin's and zebra finches. I also utilise whole genome data from a budgerigar and two penguins to provide further power for the analyses. A phylogenetic tree of the relationships among the different bird species used for the analyses is shown in **Figure 5.1**.

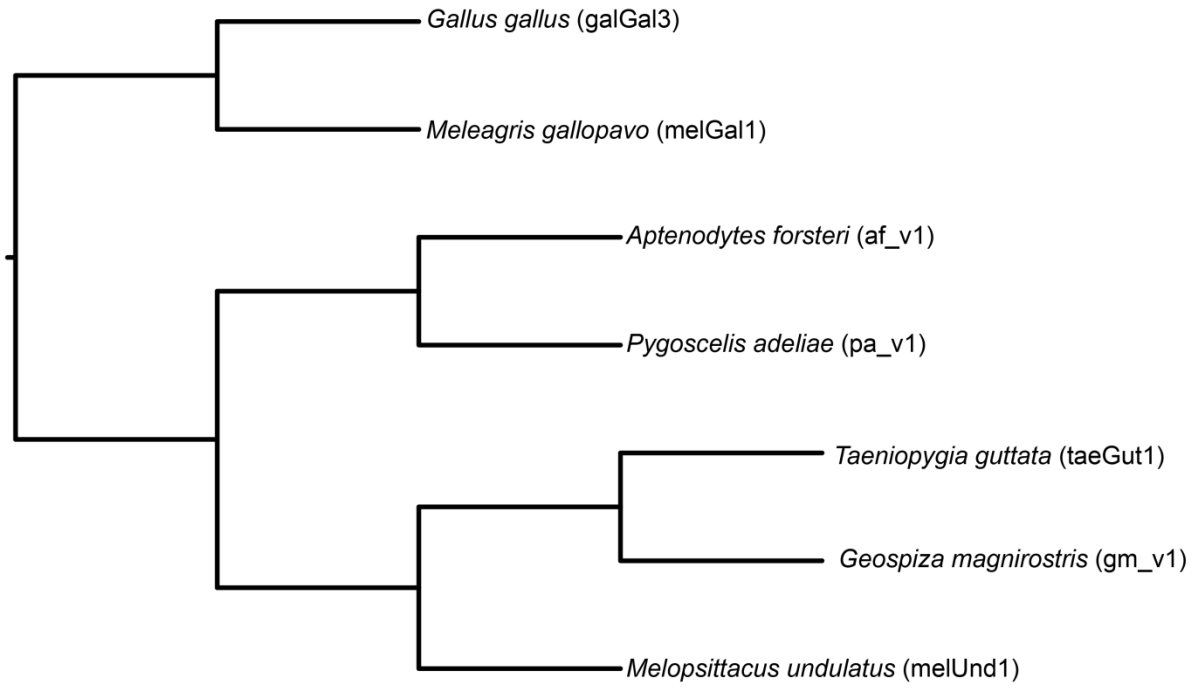

Figure 5.1: Phylogenetic tree topology depicting the relationships among different avian species. The symbols refer to the genome assemblies used as described in Table 2.1. The branch lengths are not to scale. The tree is visualised with the figtree software (Rambaut 2008).

I applied the Neutral Indel Model 1 (NIM1) to estimate the quantity of sequence constrained with respect to indels (a quantity termed $\alpha_{selIndel}$) between different pairs of avian genomes. NIM1 requires as an input chained and netted LASTZ whole genome pairwise alignments that I processed further with a trimming step to remove poor quality aligned sequence (alignment parameter details are given in **Table 2.2** and **Table 2.3**). The NIM1 estimates $\alpha_{selIndel}$ by quantifying the excess of longer inter-gap segments (IGSs; un-gapped alignment blocks) inferred from the alignment, compared to those predicted under neutral evolution. The neutral background is modeled as the lengths of medium sized IGSs, which conform to a geometric distribution (Lunter et al. 2006). This is the same methodology as described extensively in **Chapter 3** and utilised in **Chapter 4**, an approach that builds on two previous studies (Lunter et al. 2006; Meader et al. 2010). I found that patterns of sequence constraint

are quite concordant between avian and mammalian genomes, but that mammalian genomes may contain more functional sequence than avian genomes. I also found that the trends of turnover of constrained sequence are similar between these two lineages. Using gene annotations and chromosomal information for the chicken genome, the patterns of constraint and turnover were stratified to examine how they correlate with different genomic properties.

5.3 Materials and methods

5.3.1 Protein coding sequences

Protein coding regions from the chicken (galGal3) genome were obtained from Ensembl release 61. These coding sequences were identified with a pipeline that combines information across genomic DNA, cDNA and EST data to build transcript and gene models (Curwen et al. 2004). Prior to conducting downstream analyses the coordinates of protein coding sequences were merged to create single-coverage non overlapping regions.

5.3.2 Defining chromosome type

To divide chicken genome into macrochromosomes, intermediate sized chromosomes, and microchromosomes, I followed categories used previously, partitioning the chicken genome into 5 macrochromosomes (GGA1–5), 5 intermediate chromosomes (GGA6–10), 28 microchromosomes (GGA11–38), and 2 sex chromosomes (Z and W) (International Chicken Genome Sequencing Consortium 2004).

5.3.3 Substitution rates for neutral sequences

The divergence of ancestral repeat (AR) sequences was estimated as described in **Chapter 2** by fitting the HYK85 substitution model to the aligned trimmed sequences using PAML (Yang 2007). The estimates of the lineage-specific synonymous substitution rates (dS) were made by applying the PAML M2a Maximum-likelihood branch model to a stringently defined and filtered set of one-one orthologous protein sequence constructed across diverse amniotic species (the details of the ortholog set are given in **Chapter 6)**. The M2a is a branch

model with two free parameters that estimates the substitution rate and permits dN/dS ratios exceeding one (Yang et al. 2005; Yang 2007).

5.4 Results

5.4.1 Estimates of $\alpha_{selIndel}$ between diverse avian genome pairs

I provided genome-wide estimates of $\alpha_{selIndel}$ for a variety of avian species pairs using the NIM1. I found that between 65.8Mb and 133.5Mb of avian genomes is mutually constrained between each pair of genome sequences (**Table 5.1**). This equates to approximately 5.7% to 11.5% of avian genomes (estimated using the mean avian genome size of 1159Mb) and 11.3% to 14.9% of the aligned sequence. Note that the quantity of constrained sequence as a proportion of the aligned sequence is relatively similar across all species pairs, as is also observed for mammalian species. Sequence constraint estimates are much higher than the quantity of protein coding sequence (24.6Mb predicted for the chicken genome) indicating that the majority of functional sequence in birds is noncoding, as has been inferred for mammals (Meader et al. 2010).

Table 5.1: Estimates of $\alpha_{selIndel}$ and quantities of aligned sequence for avian species pairs' genomes. The details of the genome assemblies and alignments are given in Table 2.1, Table 2.2, and Table 2.3.

Species pair	Estimate of $\alpha_{selIndel}$ (Mb; 95% confidence interval)	Aligned sequence (Mb)	$\alpha_{selIndel}$ as percentage of aligned (%)
galGal3 – melGal1	133.5 (131.4 – 135.4)	895.9	14.9
galGal3 – pa_v1	109.7 (119.4 – 120.9)	827.0	13.3
galGal3 – af_v1	107.6 (102.8 – 104.1)	822.9	13.1
galGal3 – taeGut1	83.8 (83.4 – 84.2)	664.7	12.6
galGal3 – gm_v1	65.8 (65.4 – 66.2)	580.4	11.3
taeGut1– gm_v1	124.8 (123.1 – 126.5)	925.6	13.5
taeGut1 – melUnd1	104.5 (103.9 – 105.0)	896.8	11.7
taeGut1 – pa_v1	120.1 (119.4 – 120.9)	940.9	12.8
taeGut1 – af_v1	103.5 (102.8 – 104.1)	816.0	12.6

5.4.2 Sequence constraint positively correlated with G+C content

I found that the quantity of constrained sequence estimated by the NIM1 increases with G+C content across all the alignments (**Figure 5.2**), which can also be observed directly from the IGS histograms (**Appendix Figures A.17–A.25**). This is consistent with expectations since genomic regions with a higher gene density and more protein coding sequence tend to be more GC rich. However, the trend is likely to be more complex than this since G+C content is also associated with mutation rates, gene densities, recombination rates, gene expression levels, and protein structures (Mouchiroud et al. 1991; Fullerton et al. 2001; D'Onofrio et al. 2002; Hardison et al. 2003).

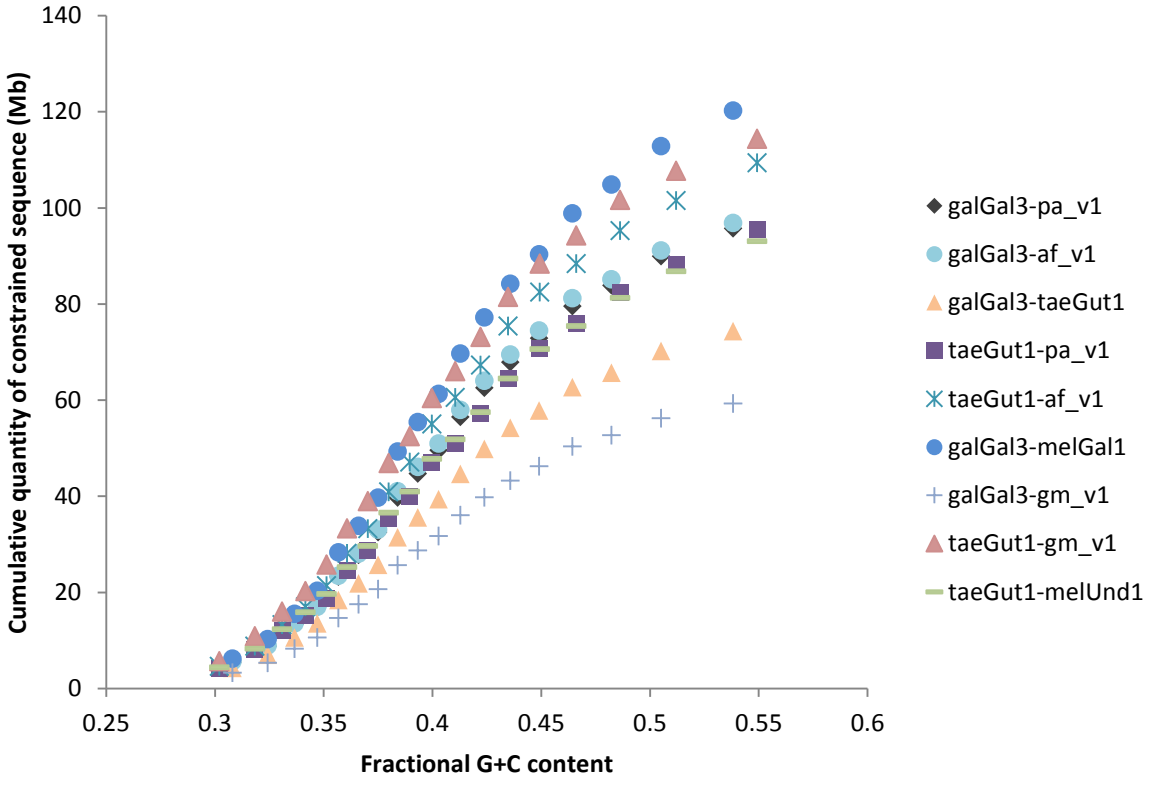

Figure 5.2: The correlation between the estimated quantity of constrained sequence by NIM1 and the G+C content of first mentioned species across equally populated GC-bins.

5.4.3 Turnover of functional sequence in avian genomes

I found a strong negative correlation between the estimates of $\alpha_{selIndel}$ for avian species pairs and the nucleotide divergence between their genome sequences, using ancestral repeats (ARs; aligned transposable element derived sequence) as the proxy for the neutral substitution rate (**Figure 5.3**). This implies that functional sequence turns over rapidly across avian evolution as it is lost and gained. Fitting a birth-death exponential model of sequence turnover to the data as described in **Chapter 2** and **Chapter 4**, I estimated the rate of turnover of functional sequence as b = 1.24 (0.39–1.80; 95% confidence interval). This equates to a half-life, $d_{1/2}$ = 0.56 (0.38–1.79), meaning that 0.56 neutral sites are expected to be mutated during the time it takes half of the functional sequence in the genome to turnover. Assuming a substitution rate of 1.91×10^{-9} per site per year (Nam et al. 2010), this $d_{1/2}$ value can in turn be converted to a half-life divergence time of 147My (99–469My). Note that this substitution rate is estimated from chicken 4-fold degenerate sites, and this estimate is lower than that predicted for the passerine lineage (Nam et al. 2010).

Since the methodology I use here to identify constrained sequence is the same as the approach I used in **Chapter 4** for examining patterns across mammalian evolution, the results are directly comparable between the two lineages. Despite the smaller size of avian compared to mammalian genomes, the estimates of $\alpha_{selIndel}$ and patterns of turnover are visually similar to the trends observed for mammals (**Figure 5.3**). Fitting the model of turnover to the mammalian data, the rate of turnover is estimated at b = 2.05 (1.67–2.45), or $d_{1/2}$ = 0.34 (0.28–0.42). Conducting a formal comparison using a likelihood ratio test, with the length of the NIM1 95% confidence interval to calculate the weight for each data point, I find that the differences in the rates of turnover are not significantly different between the mammalian and avian data (p = 0.81).

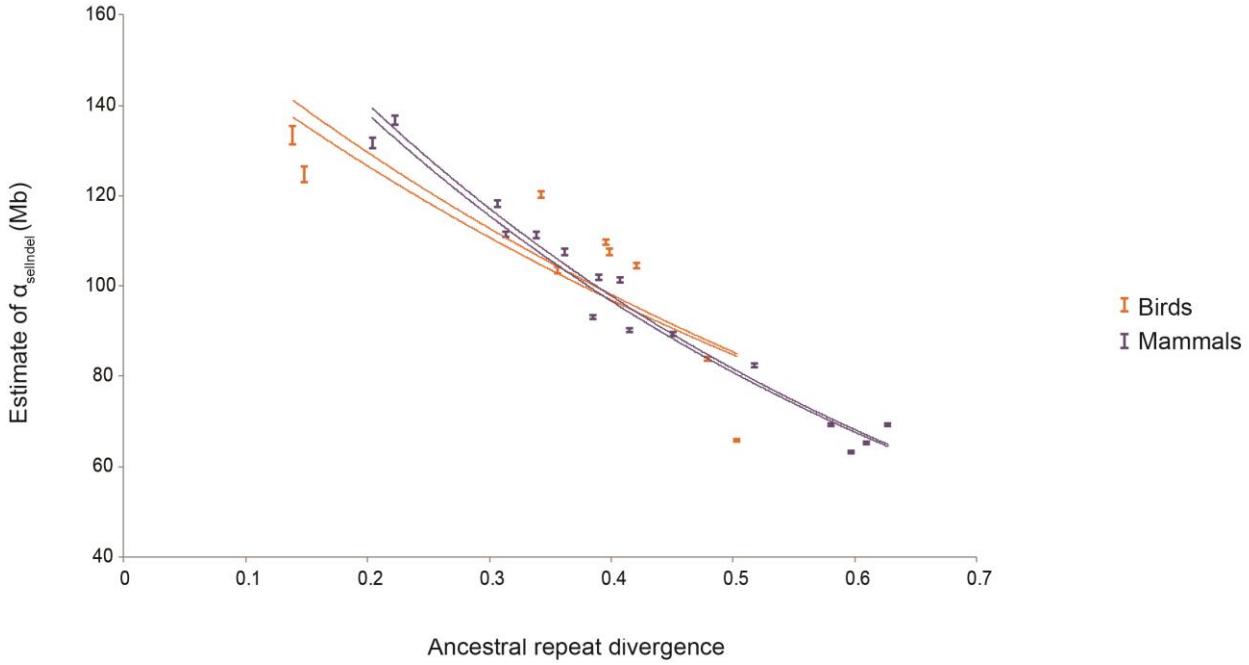

Figure 5.3: Turnover of functional sequence in avian (orange) compared to mammalian (purple) genomes. The estimates of $\alpha_{selIndel}$ are shown against the ancestral repeat sequence divergence for 9 avian and 16 mammalian species pairs. The larger spread of the avian estimates is likely due to the greater levels of incompleteness for bird assemblies.

5.4.4 Comparing turnover of noncoding and coding avian functional sequence

Using protein coding sequence defined by Ensembl in the chicken genome, I compared the quantities of constrained sequence and rates of turnover between chicken coding and noncoding sequences. As expected, protein coding sequence shows much higher levels of constraint than noncoding sequence, with approximately ten-fold higher levels of constraint in coding sequence (**Figure 5.4**). This is concordant with the trends for the mammalian data shown in **Chapter 4**, and consistent with expectations, since it is widely accepted that protein coding sequence is highly conserved compared to noncoding sequence (Waterston et al. 2002).

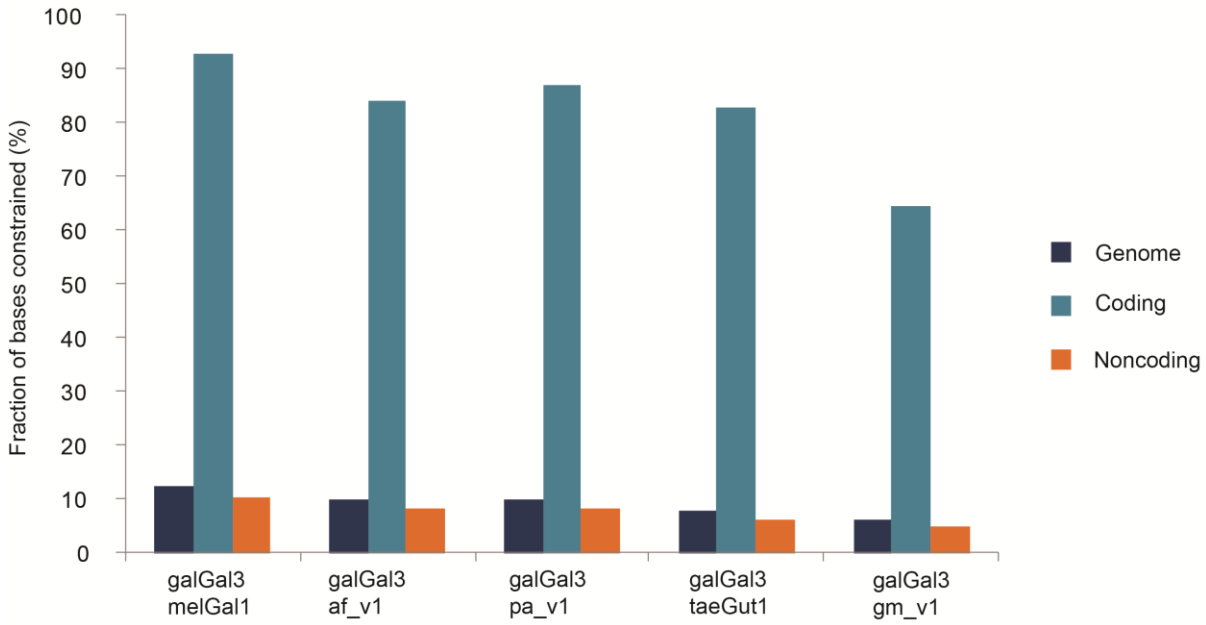

Figure 5.4: The fraction of genomic, coding and noncoding bases estimated to be constrained by NIM1 between the chicken genome and the genomes of other avian species. Coding sequences show far higher levels of conservation than noncoding sequence.

Furthermore, I found that constrained noncoding sequence is turning over rapidly, while constrained protein sequence appears more stably preserved (**Figure 5.5**). I estimate that noncoding sequence turns over with b = 1.60 (0.09–3.03), which is a $d_{1/2}$ of 0.43 (0.23–7.70), or a half-life time of 113My (60–2016My). By contrast, the rate of turnover of coding sequence was found to be lower with b = 0.61 (0.50–1.59), $d_{1/2}$ = 1.14 (0.44–1.39), and half-life time = 298My (115–364My). Note that the large confidence intervals, particularly for the noncoding data, reflect the uncertainty due to the paucity of data. Probably due to this lack of power, a likelihood ratio test comparison does not produce a significant p-value when comparing the rate of turnover of constrained coding and noncoding sequence (p = 0.56). Additionally, these quantitative estimates should be treated with caution due to the incomplete nature of the chicken protein coding gene set, and the relatively poor quality of

the Darwin's finch genome, which may artificially increase the inferred rate of turnover for both noncoding and noncoding sequences. The Darwin's finch genome is presented and extensively discussed in **Chapter 6**.

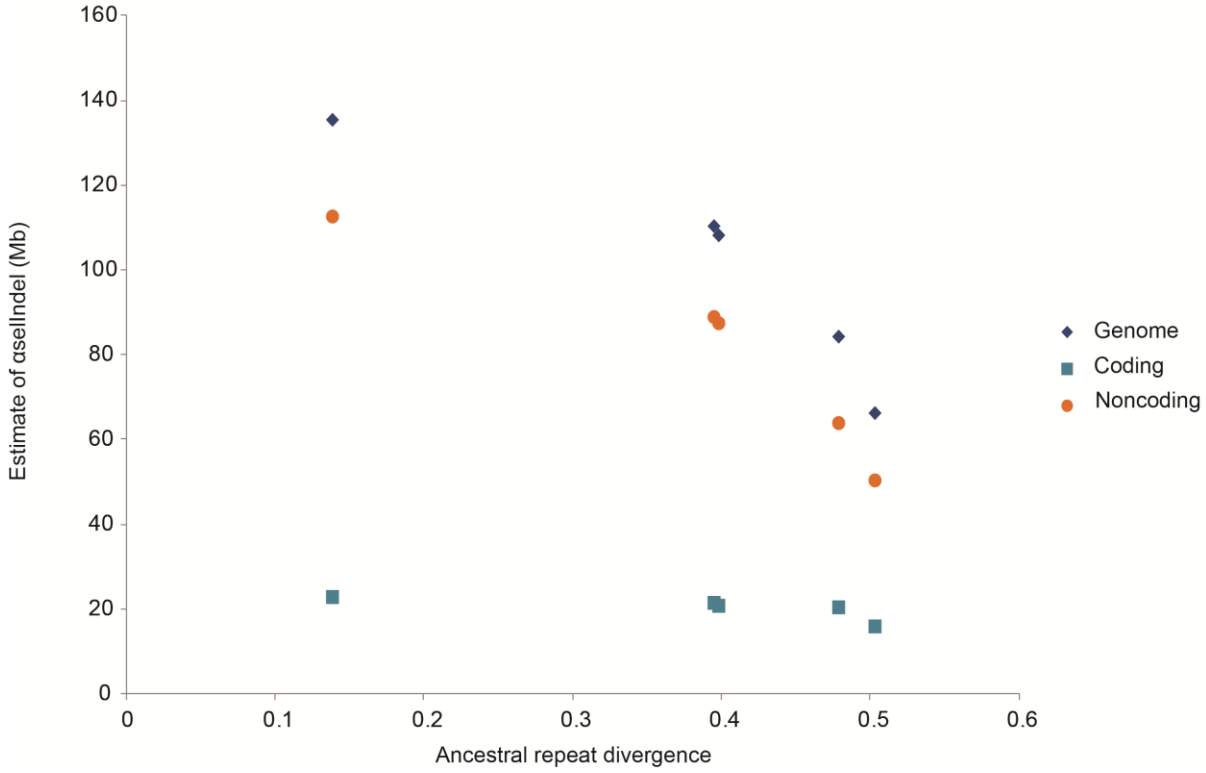

Figure 5.5: The turnover of chicken sequence sub-divided into the coding and noncoding components. The estimates of $\alpha_{selIndel}$ are shown against the ancestral repeat divergence. Noncoding sequence may show more rapid turnover than coding sequence, but the analysis is low on power.

5.4.5 10.3 –16.8% of avian genomes predicted to be functional at the present

From the fitted model of sequence turnover for the whole genome estimates of $\alpha_{selIndel}$ across the different species pairs, I extrapolated back to a divergence of zero to estimate the quantity of sequence subject to present-day purifying selection, and thus the total quantity of sequence predicted to be functional now. This quantity is estimated at 13.4% (10.3–16.9%) or approximately 155Mb (119–196Mb). This is considerably more proportionally than the 7.7% (6.6–8.9%) estimated for mammalian genomes with the same approach, but less in absolute terms, since the estimates for mammalian genomes equate to approximately 223Mb (191–258Mb). With the current data, it is uncertain if these absolute quantities are significantly different or not, but the results could suggest that mammalian genomes may contain more functional sequence than avian genomes. However, it is clear that a greater proportion of avian genomes is functional than mammalian genomes.

5.4.6 Smaller chromosomes harboured proportionally more functional sequence

I examined how estimates of $\alpha_{selIndel}$ vary across the chicken chromosomes of different sizes. I found for all species pairs that microchomosomes proportionally contain the most constrained sequence, followed by the intermediate sized chromosomes, and then the macrochromosomes (**Figure 5.6**). I thus found there to be a negative correlation between the density of functional elements and chromosome size. This pattern could be explained by the increased gene density on microchromosomes, since typically avian microchoromomes consist of approximately a quarter of the total genomic sequence, but encode half of the genes (Burt 2002). I expect genic regions to be enriched for functional elements and therefore microchromosomes to be enriched for constrained sequence.

However, the relationship between chromosome size and constrained sequence is not likely to be this simple, because recombination rates are correlated with chromosome size. The increased recombination rate on the microchromosomes means that they are subject to

reduced Hill-Robertson interference (Hill and Robertson 1966). The Hill-Robertson effect occurs as genomic regions with reduced recombination rates (such as the macrochromosomes) have increased linkage between the sites. This increased linkage means that selection acts on larger blocks because individually selected loci are less likely to become disassociated from their linkage partners. This causes a local reduction in the effective population size (N_e) of the genomic region. This in turn reduces the efficacy of selection over this region since natural selection acts not on the strength of selection (s) alone, but on the product of this and the effective population size, $N_e s$ (Kimura 1983). This is because in small populations the stochastic forces of genetic drift play a more important role in altering allele frequencies. So, this theory predicts that regions with higher recombination rates (which can be proxied by chromosome size) are expected to more efficiently purge their deleterious variants, and thus microchromsomes are expected to be under more effective purifying selection than macrochromosomes.

The Z chromosome harbors proportionally less constrained sequence than the autosomes (**Figure 5.6A**). This could be explained by the relatively low N_e of the Z chromosome, which reduces the efficacy of purifying selection.

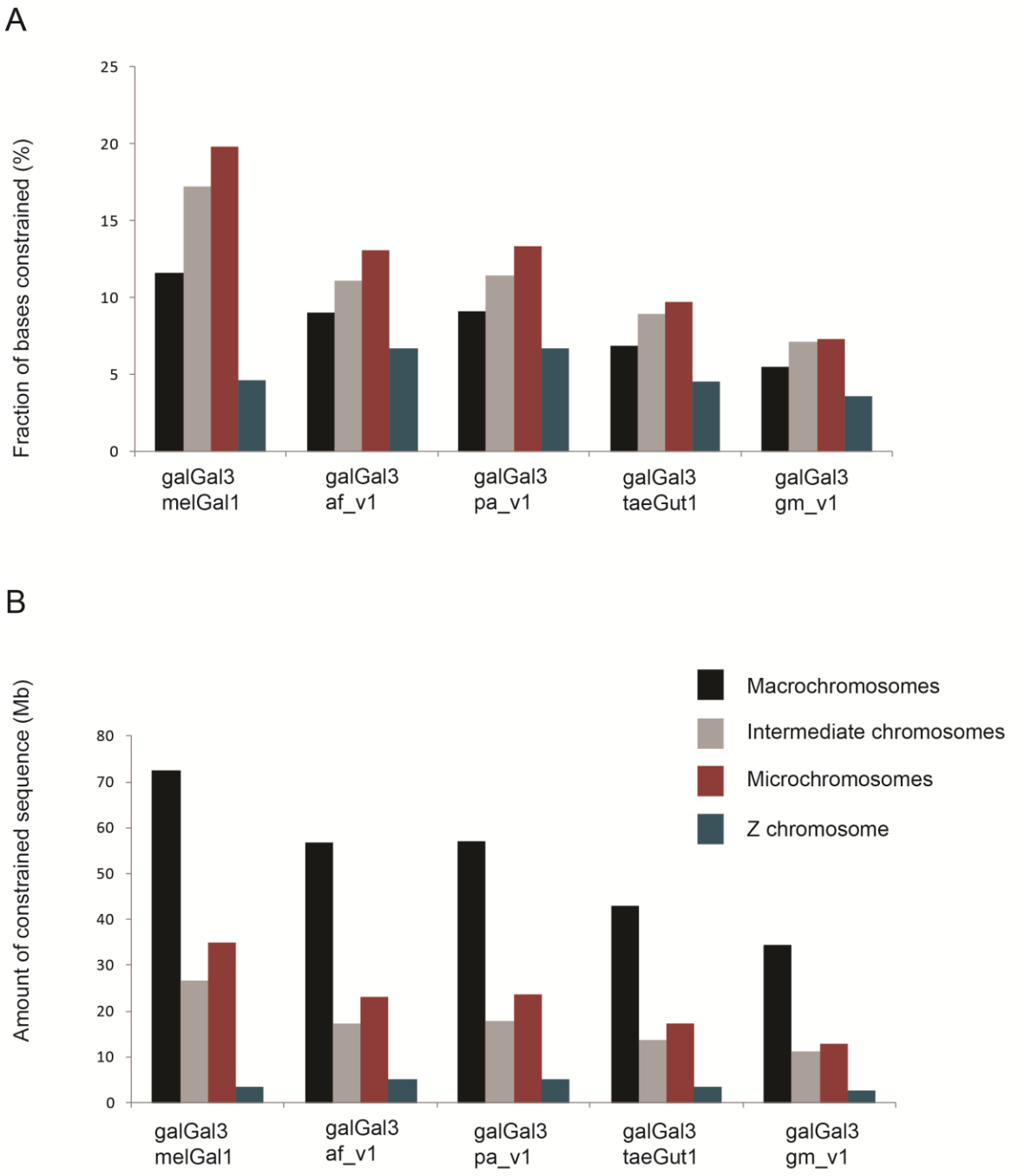

Figure 5.6: The percentage A., and absolute quantity B., of chicken bases estimated to be constrained by NIM1 for alignments between chicken and different avian species, shown across the macrochromosomes, intermediate sized chromosomes, microchromosomes, and Z chromosome. Smaller chromosomes show proportionally higher levels of conservation than large chromosomes.

5.4.7 Avian ancestral repeats evolve faster than synonymous sites

I have used ancestral repeats (ARs) as the putatively neutral sequence from which to estimate divergence, as this makes the estimates directly comparable to those for the mammalian data presented in **Chapter 4**. For the mammalian data ARs showed substitution rates that were very similar to synonymous mutations (dS; where the amino acid sequence is unchanged by a mutation), so the choice of neutral standard was unimportant (**Figure 4.2**). However, a previous study inferred there is significant constraint on synonymous sites in bird genomes, since they evolve significantly slower than ARs (Kunstner et al. 2011b). Consistent with this, I found that AR divergence estimates are typically around 25% higher than estimates of dS for the same species pair (**Figure 5.7A**). Therefore, in avian genomes the choice of sequence divergence measure will impact on the turnover rate estimates and in particular the extrapolation to estimate the total quantity of sequence subject to purifying selection at the present day (**Figure 5.7B**). Nevertheless, since ARs evolve faster than synonymous sites they appear to be candidates for neutrally evolving sequence. However, I cannot categorically rule out other possible explanations for this substitution rate difference, such as rampant positive selection across ARs, or some unknown mutational bias increasing the substitution rate within ARs. Furthermore, ARs make up less than 10% of avian genomes compared to 40–50% of mammalian genomes (Lander et al. 2001; Waterston et al. 2002; International Chicken Genome Sequencing Consortium 2004; Warren et al. 2010), which mean that mammalian genomes contain approximately 12-fold more ARs than avian genomes. This relative scarcity of ARs in avian genomes means that there is diminished power to examine patterns of AR evolution.

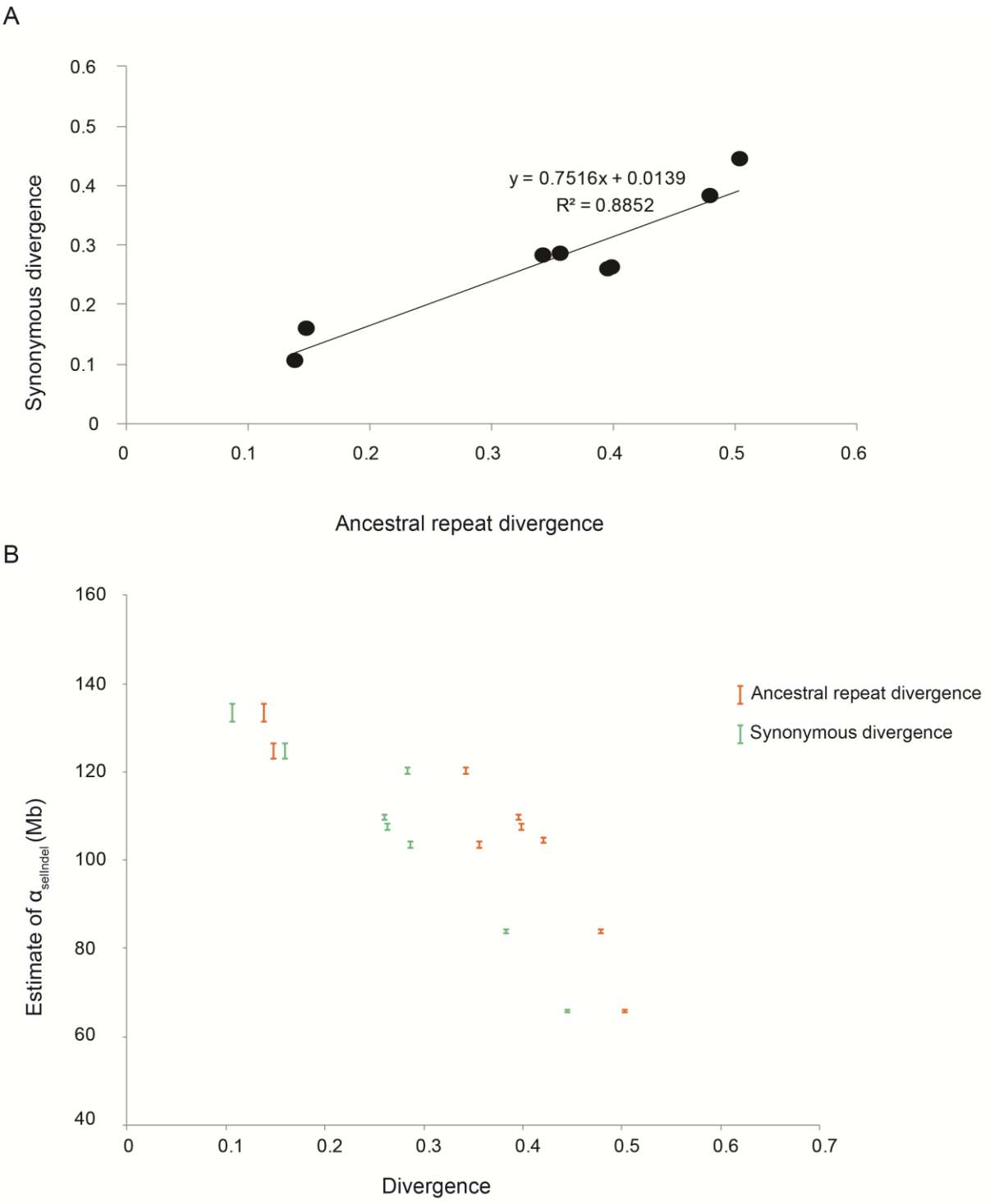

Figure 5.7: The difference in substitution rate between ancestral repeats (ARs) and synonymous sites, two site types that have previously been deemed suitable proxies for estimating the rate of evolution in neutral sequences. A. Although the two proxies correlate well, B. ARs consistently showed faster rates of evolution than dS.

5.4.8 Impact of various genome assemblies on estimates of $\alpha_{selIndel}$

I demonstrated in **Chapter 3** that the choice of genome assembly and alignment build can significantly impact on the $\alpha_{selIndel}$ estimator, and showed that the introduction of a log-odds alignment trimming step makes the estimates relatively robust to such technical variation. However, since many fewer avian genomes have been sequenced compared to mammalian genomes, I needed to use some lower quality genome assemblies during the analysis, and hence thought it necessary to explore the impact of this further on the analyses.

The first avian genome assemblies were those for the chicken and zebra finch, which were sequenced with traditional Sanger sequencing (International Chicken Genome Sequencing Consortium 2004; Warren et al. 2010). These assemblies are of superior quality to the subsequently generated turkey assembly, which was the first avian genome sequenced with the aid of next-generation sequencing (NGS) technologies, in this case 454-sequencing from Roche Life sciences (Dalloul et al. 2010). Since then, a number of avian genomes have been sequenced with either 454-sequencing (such as the *G. magnirostris* genome), Illumina based sequencing (such as the adelie and emperor penguin genomes), or hybrid assemblies using both technologies (such as the budgerigar genome). Although NGS assemblies are much faster and cheaper to produce than conventional Sanger sequencing based assemblies, they also tend to be more fragmented, contain more sequencing errors, and be less complete.

Recently, new versions have been released of both the chicken and zebra finch genome assemblies. Estimates of the quantity of constrained and aligned sequence are very similar between the different trimmed chicken – zebra finch alignments (**Table 5.2**), implying that assembly quality is not an issue for these Sanger sequenced genomes. To improve alignment quality, I only generated alignments where one of these two genome sequences was used as the target sequence. Nevertheless, the quantity of aligned sequence to the *G. magnirostris* genome still appears low, and the points involving this genome are slight outliers to the trend

of turnover (**Table 5.1, Figure 5.3**). This is almost certainly due to the low quality of the assembly since it has a low scaffold N50 value and a high number of ambiguous 'N' bases (**Table 2.1**). Additionally, the assembly was estimated to cover just 89% of the euchromatic genome and 75% of the complete genome (**Chapter 6)**.

Table 5.2: The estimates of $\alpha_{selIndel}$ and the quantities of aligned sequence for different chicken – zebra finch alignments based on different genome assemblies are concordant.

Genome assemblies	Estimate of $\alpha_{selIndel}$ (Mb; 95 % confidence interval)	Amount of aligned sequence (Mb)
galGal3 – taeGut1	83.8 (83.4 – 84.2)	664.7
galGal4 – taeGut2	83.9 (83.4– 84.3)	636.1

5.5 Discussion

I found that estimates of $\alpha_{selIndel}$ by the NIM1 between a pair of avian genome sequences are strongly negatively correlated with the divergence, as measured by either the AR divergence or dS. This is consistent with the notion that functional sequence turns over rapidly across avian evolution. The turnover of functional sequence is estimated to have a half-life of $d_{1/2}$ = 0.56 AR substitution units, and noncoding sequence shows higher levels of turnover than coding sequence, although the difference is not statistically significant.

These patterns of constraint and turnover in birds are similar to those observed across mammalian evolution using the same approach. However, my results suggest that mammalian genomes may contain more functional sequence than avian genomes, although this is not clear with the current analytical power. It appears that mammalian and avian genomes contain functional sequence that turns over at similar rates in both lineages. This is not evident *a priori*, since avian genomes have a different structural makeup to mammalian genomes.

My findings suggest that mammalian genomes may contain more functional sequence than the avian genomes. The quantity of functional sequence in a genome could reflect our naïve

conceptions of organismal complexity, as suggested by evidence that there is more constrained sequence in mammalian genomes than in the genomes of fish, fruit flies, or nematode worms (Meader 2010; Meader et al. 2010). Traditionally, mammals have been viewed as cognitively superior to birds, but such anthropocentric notions appear ill-founded since birds have 'ape-like intelligence', including the ability to use tools and sophisticated spatial memory (Emery 2006). Moreover, such advanced cognition is actually not a general feature of either animal class. Rather than this putative additional functional sequence in mammals underlying further organismal complexity, I tentatively suggest that there could have been a contraction in the functional repertoire of avian genomes. Such a contraction in the functional element composition might have been as a consequence of the generally high effective population sizes in birds, which increases the efficacy of purifying selection and so might have reduced the functional redundancy within avian genomes compared to their mammalian counterparts. An alterative explanation is that the inferred deficit of functional sequence in avian genomes could be due to the relatively poor quality of some avian genome assemblies, which makes it difficult to align sequences and identify constrained regions.

I find that the rates of turnover are not significantly different between the mammalian and avian lineage, despite the different structural compositions of their genomes, with avian genomes being smaller, karyotypically different, and bearing far fewer transposable elements. This suggests that the gain and loss of functional sequence may have proceeded in a 'molecular clock' manner. This is consistent with the idea that there could be fundamental universal laws of genome evolution. To take an analogous example, it has been argued that patterns of gene family evolution follow a power law distribution, and can be modeled by a birth-death-and-innovation model that incorporates just the duplication, deletion and modification of genes (Koonin 2011). It is plausible that the turnover of functional sequence could follow a birth-death model that is universal, at least across diverse amniotes.

Chapter 6: Evolutionary analyses of a Darwin's finch genome reveal examples of positive selection

6.1 Abstract

A classical example of repeated speciation coupled with ecological diversification is the evolution of 14 closely related species of Darwin's (Galápagos) finches (Thraupidae, Passeriformes). Their adaptive radiation in the Galápagos archipelago took place in the last 2–3My and some of the molecular mechanisms that led to their diversification are now being elucidated. Here I report evolutionary analyses of genome of the large ground finch, *Geospiza magnirostris*. 13,291 protein-coding genes were predicted from a 991.0Mb *G. magnirostris* genome assembly. Gene orthology relationships were then defined and whole genome alignments constructed between the *G. magnirostris* and other vertebrate genomes. Genic evolutionary rate comparisons indicate that similar selective pressures acted along the *G. magnirostris* and zebra finch lineages, which is surprising given that effective population size values are very different between the two lineages. The neutral sequence divergence, as approximated by the synonymous substitution rate, is estimated to be similar between the ancestral passerine and galliform lineages. 21 otherwise highly conserved genes were identified that each show evidence for positive selection on amino acid changes in the Darwin's finch lineage. Two of these genes (*IGF2R* and *POU1F1*) have been implicated in beak morphology changes in Darwin's finches. Five of 47 genes showing evidence of positive selection in early passerine evolution have cilia related functions, and may be examples of adaptively evolving reproductive proteins. These results provide insights into past evolutionary processes that have shaped *G. magnirostris* genes and its genome.

6.2 Introduction

Since their collection by Charles Darwin and fellow members of the HMS Beagle expedition from the Galápagos Islands in 1835 and their introduction to science, Darwin's finches have been subjected to intense research. Many biology textbooks use Darwin's finches (formerly known as Galápagos finches) to illustrate a variety of topics in evolutionary theory, including speciation, natural selection, and niche partitioning (Freeman and Herron 2003; Barton 2007; Futuyma 2009). Darwin's finches continue to be a very valuable source of biological discovery. Several unique characteristics of this clade have allowed multiple important recent breakthroughs in our understanding of changes in island biodiversity, mechanisms of repeated speciation coupled with ecological diversification, evolution of cognitive behaviours, principles of beak/jaw biomechanics as well as the underlying developmental genetic mechanisms in generating morphological diversity (Grant and Grant 1989; Abzhanov 2010).

Recent molecular phylogenetic reconstructions suggest that the adaptive radiation of Darwin's finches in the Galápagos archipelago took place in the last 2–3My, following their evolution from a finch-like tanager ancestral species (**Figure 6.1**) that probably arrived on the islands from Central or South America (Sato et al. 2001; Burns et al. 2002; Petren et al. 2005). Nuclear microsatellite and mitochondrial DNA have undergone limited diversification, partly because the Galápagos history of the finches has been relatively short, and partly because of introgressive hybridization (Grant and Grant 2008; Grant and Grant 2010). Morphological evolution in this group of birds is a fast and ongoing process that has been documented over the years in many publications on their population-level ecology, morphology and behaviour (Grant and Grant 1989).

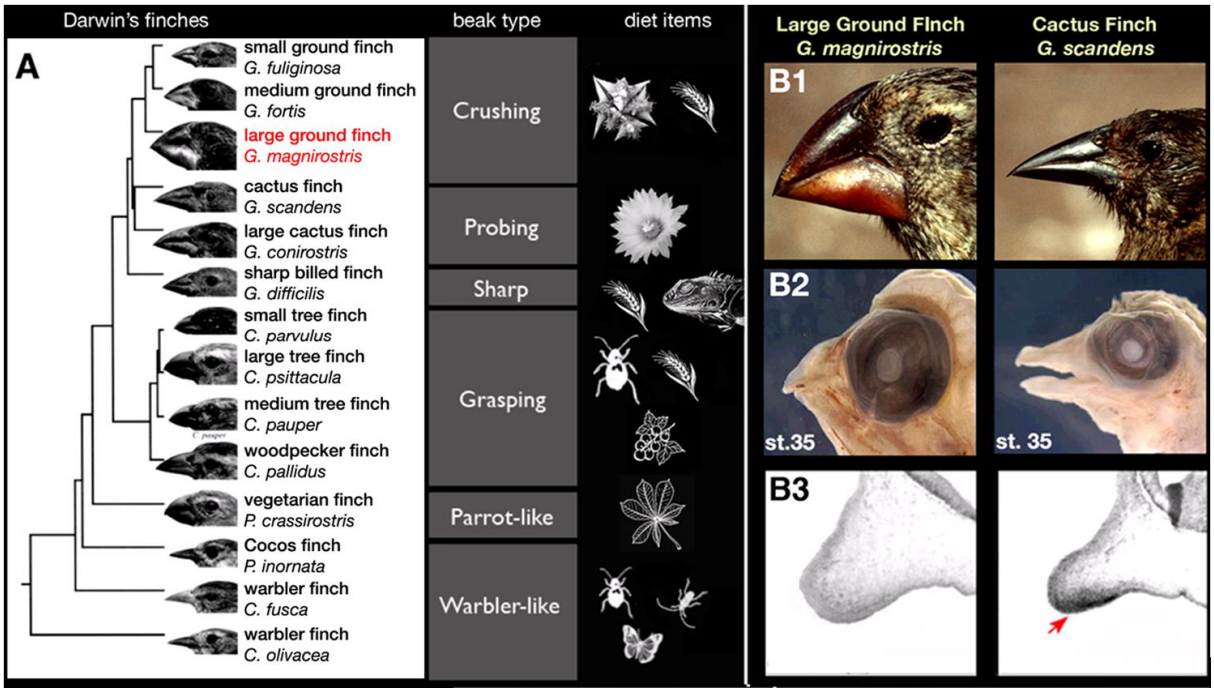

Figure 6.1: Evolutionary mechanisms for beak shape diversity in Darwin's finches (Thraupidae, Passeriformes). A. The molecular phylogeny of 14 species of Darwin's finches shows a range of beak shapes in this group of birds. These species have beaks of different shapes that allow them to feed on many different diets: insects, seeds, berries, and young leaves. The phylogeny was taken from Grant and Grant (1989). B1. Large ground finch (left) has a very deep and broad bill adapted to crack hard and large seeds, while the cactus finch (right) has an elongated and pointy beak for probing cactus flowers and fruits. B2. *Geospiza* finch bills develop their distinct shapes during embryogenesis and are apparent upon hatching; mid-development stage 35 embryos are shown from Abzhanov et al. (2004). B3. The cactus finch-specific expression of CaM was validated by in situ hybridization after it was identified as a candidate by a microarray screen (Abzhanov et al. 2006).

Beak size and shape, as well as body size, are the principal phenotypic traits that have diversified in Darwin's finches (Grant and Grant 1989). The most studied group within the Darwin's finches is the monophyletic genus *Geospiza*, which includes three distinct bill shapes: the basal sharp-billed finch *G. difficilis* has a small and symmetrical beak used to feed on a mixed diet of insects and seeds; cactus finches *G. scandens* and *G. conirostris* feature an elongated and pointed bill suitable for probing cactus flowers and fruit; and ground finches possess deep and broad bills adapted for cracking seeds (Grant and Grant 1989). Among the ground finches, which include small, medium and large species, the large ground finch *G. magnirostris* has the most modified beak that it uses to crack (and then consume) large and hard seeds (**Figure 6.1**). Importantly, beak shapes develop during early embryogenesis and finch hatchlings show species-specific features. Recent molecular analysis has shown that the ground finch bill morphology correlates with a developmentally earlier and broader gene expression of *Bone morphogenetic protein 4* (*Bmp4*), especially in the large ground finch. Functional experiments mimicking such changes in *Bmp4* expression using laboratory chicken embryos are consistent with its role in this *Geospiza* beak trait (Abzhanov et al. 2004). Similar experiments elucidated the roles of three further developmental factors, *Transforming Growth Factor beta Receptor Type II* (*TGFβRII*), *beta-Catenin* (*βCat*) and *Dickkopf-3* (*Dkk3*), at later stages of beak development that help in forming the bill shapes that are unique to ground finches (Mallarino et al. 2011). Other analyses revealed an important role of change in *Calmodulin* (*CaM*) expression pattern for the development of elongated bills of cactus finches (Abzhanov et al. 2006).

In 2008 a project was initiated to sequence the genomes of some of the Darwin's finches, inspired by the 200th anniversary of the birth of Charles Darwin that occurred the following year. A consortium of evolutionary biologists, Darwin's finch biologists and genome sequencing technologists was formed and funds raised to pay for sequencing that was

subsequently conducted at Roche-454. There was interest to perform a whole genome analysis of the large ground finch *G. magnirostris* because of the evolutionary importance of the entire clade of Darwin's finches to the fields of ecology and evolutionary biology, the potential of genomic analysis for uncovering the genetic basis of key phenotypic traits and the then scarcity of genomic studies of birds (especially when compared to mammals). The species was chosen because it arose relatively recently and it has one of the most adapted and distinctive bill shapes. The embryonic individual chosen for genome sequencing was sampled from a population from the small and well isolated island of Genovesa which exhibit the largest bills of all existing Darwin's finches, and have an estimated effective population size of 75–150 individuals (Grant and Grant 1989).

The field of evolutionary and comparative genomics should benefit broadly from analyses of an additional species of passerine. *G. magnirostris* diverged from the first passerine bird to be sequenced, the zebra finch (*Taeniopygia guttata*) (Warren et al. 2010), approximately 25My ago (Cracraft and Barker 2009), which is comparable to the divergence time separating mouse and rat (Arnason et al. 2008). I describe here the genome analyses from the initial sequencing, which includes the genome assembly, genome quality assessment, transposable element prediction, base composition analyses, gene prediction, orthologue and paralogue assignment, and further downstream analyses. I investigate a variety of evolutionary processes, but focus on whether episodes of positive selection have occurred along the *G. magnirostris* terminal lineage and more broadly over passerine evolution. I find evidence of positive selection on two genes, *POU1F1* (POU domain, class 1, transcription factor 1; also known as Pit1, growth hormone factor 1) and *IGF2R* (insulin-like growth factor 2 receptor), that have been implicated in Darwin's finch beak development. I validate that these mutations are true biological events using sequence data from other Darwin's finch species. I also examine patterns of selection over the songbird lineage since the divergence of passerines

from the galliformes. I find 47 predicted passerine-specific positively selected genes, and these are highly enriched in genes with cilium-related functions. Through protein sequence analysis, I infer that five of these positively selected genes may be examples of adaptively evolving reproductive proteins.

I led the analysis of the genome sequence: I performed the evolutionary rate and positive selection analyses, conducted the enrichment tests, did the genome quality assessments, contributed to the gene prediction and orthologue/paralogue assignment analyses, assisted with the transposable element and G+C content analyses, and drafted the now published manuscript (Rands et al. 2013) that forms the basis of this Chapter. I acknowledge here the contributions of others to the genome project; the individuals are affiliated with the University of Oxford unless specified otherwise. Aaron Darling (University of California Davis) led construction of the genome assembly. Matt Fujita (University of Texas at Arlington) carried out the G+C content and contributed towards transposable element analyses. Lesheng Kong helped with the gene predictions and orthologue/paralogue assignments. Matt Webster (Uppsala University) analysed substitution patterns in the positively selected genes and contributed helpful discussions on the isochore analysis. Celine Clabaut (University of California Davis) collected and processed finch material for sequencing. Richard Emes (University of Nottingham) assisted with the application of the test for positive selection. Andreas Heger assisted with technical aspects of gene predictions and orthologue/paralogue assignment. Luis Sánchez-Pulido helped with the protein sequence analysis. Michael Brent Hawkins (Harvard University) cloned and assisted in the analysis of candidate genes. Michael Eisen (University of California Berkeley) helped initiate the project. Clotilde Teiling (454 Life Sciences) helped with the genome sequencing. Jason Affourtit (454 Life Sciences) participated in experimental design, and technical consultation and coordination of the library construction and sequencing. Benjamin Boese (454 Life

Sciences) contributed to the genome sequencing. Peter Grant and Rosemary Grant (both Princeton University) wrote the background for the manuscript. Jonathan Eisen (University of California Davis) initiated the project. Arhat Abzhanov (Harvard University) participated in design and coordination of the study and helped write the manuscript background. Chris Ponting coordinated the analyses and helped draft the manuscript.

6.3 Materials and methods

6.3.1 DNA isolation, library construction and sequencing, and genome assembly

These steps were performed by others as described in Rands et al. (2013).

6.3.2 Whole genome alignments

As described in **Chapter 2**, whole genome pairwise alignments were constructed with LASTZ, and were then subsequently chained and netted using various UCSC utilities (Kent et al. 2003). The target genome sequences (chicken, turkey, or zebra finch) when not placed on specific chromosomes were discounted when calculating amounts of aligned sequence; such amounts are thus likely to be conservative estimates. These unplaced sequences were ignored because some sequence in the zebra finch genome assembly is artificially present in two copies, both in assembled chromosomes and in sequence not placed on chromosomes.

Using MULTIZ (Blanchette et al. 2004), I combined zebra finch – *G. magnirostris* whole genome alignments with whole genome alignments between zebra finch – *G. fortis* obtained from UCSC. This resulted in the generation of multiple sequence alignments across zebra finch, *G. magnirostris*, and *G. fortis*.

6.3.3 Repetitive element prediction

A library of de novo repetitive elements was constructed from the *G. magnirostris* assembly with RepeatScout (Price et al. 2005). RepeatScout takes a consensus seed based approach to identify repeat families. The programme starts by identifying high frequency kmers and then extends them progressively to form a consensus sequence. Alignments are subsequently

inferred between the consensus and matching occurrences within the genome. This RepeatScout defined library of template repeats was then used as an input for RepeatMasker to define the transposable element derived sequences in the genome sequence.

6.3.4 G+C Content analyses

G+C content windows were made by first constructing large non-overlapping windows and then partitioning these again into smaller windows, discarding windows with greater than 20% missing data, following an approach described previously (Fujita et al. 2011). The equilibrium GC content to which GC is evolving over time (GC*) was calculated as GC* = u/(u+v) where u is the rate of GC → AT mutations and v the rate of AT → GC mutations (Meunier and Duret 2004; Axelsson et al. 2012).

6.3.5 Gene predictions and orthologue/paralogue assignment

Gene predictions from the *G. magnirostris* genome assembly were made with a computational pipeline, Gpipe, using protein-coding genes from human, chicken, zebra finch as templates (Heger and Ponting 2008). There are four steps, as illustrated in **Figure 6.2**. (1) During pre-processing, the input template transcripts were clustered together into genes and the longest unspliced transcript was designated as the representative transcript, while the others are variant transcripts. (2) The representative template transcripts were then masked for low complexity regions with SEG and then aligned against the *G. magnirostris* genome with exonerate (Slater and Birney 2005). This resulted in a list of template transcripts paired with matching genomic regions. Note that there were often multiple matches for a given transcript. (3) The representative template amino acid sequences were then aligned to their paired genomic region with the genewise mode of Exonerate to produce transcript predictions from the genomic DNA. The best-matching likely orthologous pairs were processed first, and paralogues then defined as those that intersect with the previously covered regions. The variant transcripts were then aligned in an equivalent manner to the same regions where the

associated representative transcript matched previously. The orthologue and paralogue classes were refined further by creating priority lists that rank matches based on their sequence identity, ensuring the top ranking matches are processed first. (4) Finally, there is a post-processing step that filters the transcript predictions and removes further redundancies. This filtering is carried out based on a variety of criteria: the alignment coverage of the template, any frameshifts or stop codons in the amino acid sequence, and the exon structure conservation (Heger and Ponting 2007).

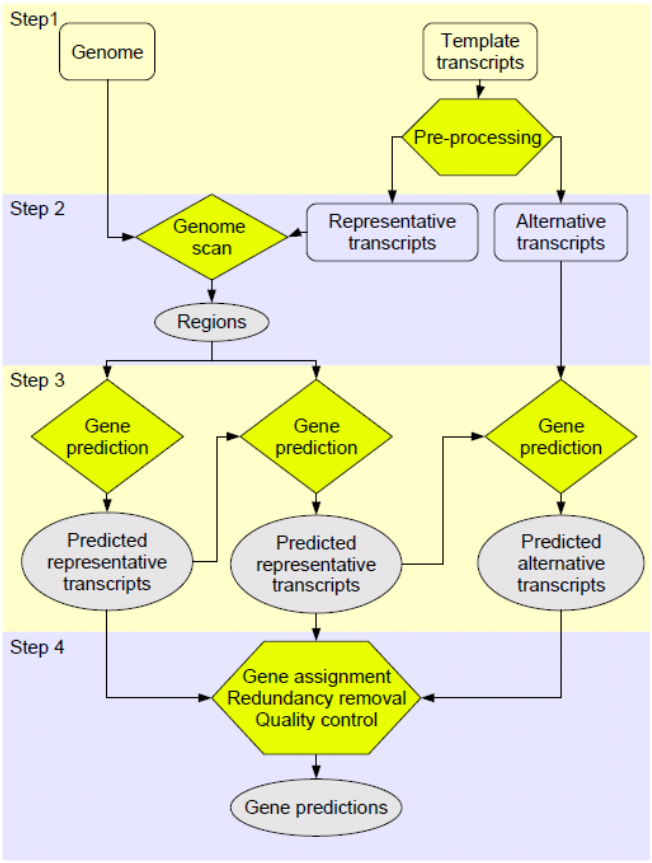

Figure 6.2: Details of Gpipe, the transcript and gene prediction pipeline. The figure was taken from Heger and Ponting (2007).

Gene sets were constructed with Gpipe in a similar manner for the genomes of the *Pygoscelis adeliae and Aptenodytes forsteri* penguins (assembly identifiers pa_v1 and af_v1; **Table 2.1**), but predictions were merged with those made by the Beijing Genomics Institute. Gene sets

were downloaded from Ensembl release 61 (February 2011) for all other five species: *Gallus gallus* (galGal3), *Meleagris gallopavo* (melGal1), *Taeniopygia guttata* (taeGut1), *Homo sapiens* (hg19), *Mus musculus* (mm9), and *Anolis carolinensis* (anoCar2) (**Table 2.1** for assembly details). Orthologues and paralogues were subsequently assigned using OPTIC (Heger and Ponting 2008). This consists of four steps: (1) orthologues are assigned between pairs of genomes using PhyOP (Goodstadt and Ponting 2006) based on a distance metric derived from BLASTP alignments, (2) pairwise orthologues are grouped into clusters, (3) sequences within a cluster are aligned using MUSCLE (Edgar 2004), and (4) phylogenetic tree topologies are estimated using TreeBeST (Vilella et al. 2009) with clusters being split into orthologous groups using the pufferfish *Tetraodon* as the outgroup.

The completeness of these gene sets was examined in two ways. First, the number of simple 1:1 orthologues between human and zebra finch was compared to the number between human and *G. magnirostris*. Second, the number of genes was calculated with orthologues predicted in *G. magnirostris* from a set of metazoan single copy genes from Creevey et al. (2011). 15 genes were excluded from the analysis, since were retired from the current Ensembl release.

From the OPTIC ortholog sets, a refined orthologue set was constructed of simple 1:1 orthologues shared across human, mouse, chicken, turkey, *G. magnirostris*, zebra finch, and the *Anolis* lizard. False positive predictions of positive selection will be more frequent in poorly aligned or sequence error-prone sequence (Mallick et al. 2009). Multiple sequence alignments (MSAs) of protein-coding sequence were thus very stringently filtered to remove poorly aligned regions using SEG, GBLOCKS, GUIDANCE (Talavera and Castresana 2007; Penn et al. 2010), and further approaches that I describe below. Strict GBLOCKS settings were used (minimum number of sequences for a conserved position=5, minimum number of sequences for a flanking position=6, maximum number of contiguous nonconserved positions=6, minimum length of block=10), only alignment columns with a GUIDANCE

score of 1 were kept, and no gaps were allowed (**Figure 6.3**). All codons containing a base with a phred quality score of 30 or less, which equates to a 0.1% probability of the base being falsely called, were also excluded. Alignment columns in 15bp windows were removed when these windows contained greater than five substitutions between aligned *G. magnirostris* and zebra finch. Such runs of substitutions may represent sequence or alignment errors. Further alignment columns that lie within seven codons of previously filtered sequence were also removed, since otherwise such codons are enriched in predicted positively predicted sites (**Figure 6.4**). The final step was crucial since it reduced the signal of positive selection approximately 30-fold. Conceptually this can be understood as aggressive filtering led to the concatenation of isolated islands of aligned sequences, causing the appearance of well aligned sequence, where in fact non-homologous pseudo-contiguous regions were being created. Finally, I discarded all genes whose remaining alignment columns numbered fewer than 10% of their predicted numbers of codons, or were less than 100 codons in length. This procedure resulted in a set of "strict" 1:1 orthologues containing 1,452 genes.

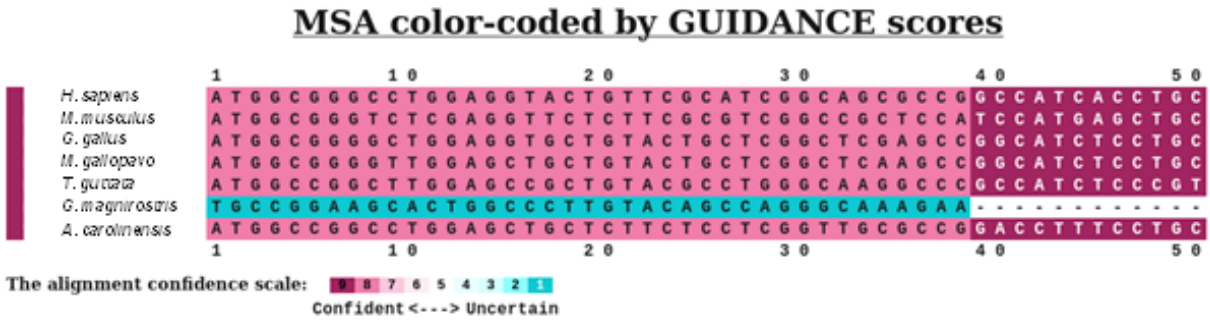

Figure 6.3: An example of a multiple sequence alignment viewed with GUIDANCE. The software was able to detect the lineage-specific poor alignment from *G. magnirostris*, unlike GBLOCKS that only considers information from whole alignment columns.

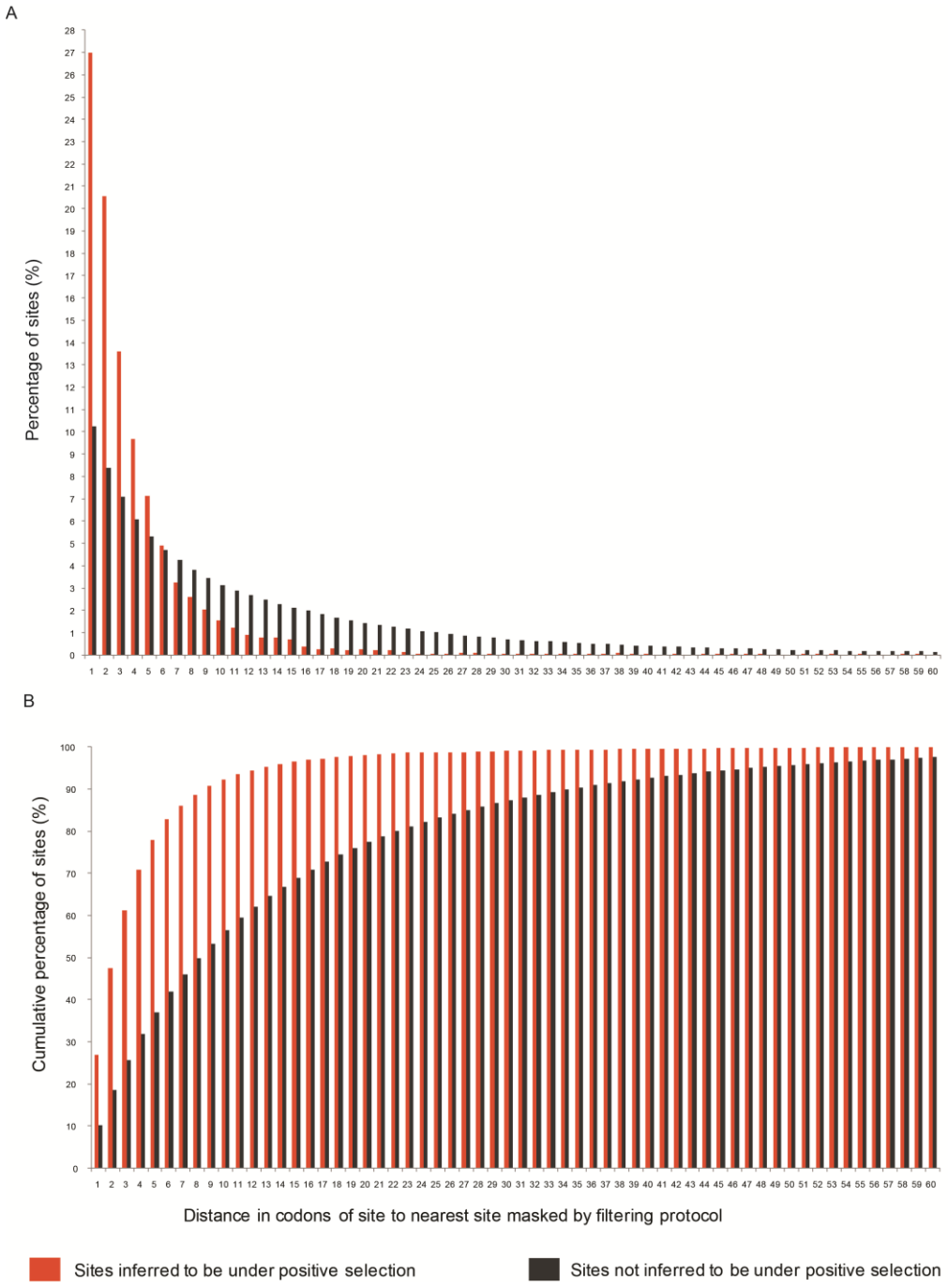

Figure 6.4: Distance of previously inferred positively selected sites from sequence already removed, A. shows the percentage, and B. the cumulative percentage of sites. Under the original implementation, positively selected sites were over represented adjacent to previously filtered alignment columns. Therefore, codon sites were further removed if they lay within seven codons of previously filtered sequence before applying the evolution rate analyses and positive selection tests to produce the results presented.

6.3.6 Evolutionary rate and positive selection analyses

dS, dN , and dN/dS values were inferred from the filtered MSAs by applying the PAML M2a Maximum-likelihood branch model (Yang 1997; Yang 2007). The branch lengths were then calculated by taking the median values across all genes in the strict orthologue set.

The filtered MSAs and guide trees were also provided as input for the branch-site test for positive selection of Zhang et al. (2005). The test identifies genes with particular codons showing evidence of positive selection by comparing a null model, where dN/dS (ω) is never allowed to exceed one (so only negative or neutral evolution is considered), to an alternative model in which some sites on the *G. magnirostris* lineage are allowed to have ω greater than one (implying positive selection). The test was run twice, and only cases where the two tests converged to within log-likelihood values at or within 0.01 were taken forward for downstream analysis. Subsequently, a likelihood ratio test (LRT) was used to compare the null and alternative model, and a Chi-squared test applied to compare the significance of the LRT scores. The number of positively selected sites in genes inferred to have evolved under positive selection was estimated using a Bayes Empirical Bayes approach, which specifies prior parameters and then integrates over the uncertainties. This approach has better performance than a Naive Empirical Bayes methodology (Yang et al. 2005).

It has been suggested that the branch site test of Zhang et al. (2005) is not statistically robust when the number of substitutions in the MSAs is small (Nozawa et al. 2009). However, this criticism is largely based on the study of Bakewell et al. (2007) who apply a branch site model across three very closely related primate species. Additionally, it has been suggested that branch-site methods are susceptible to high false positive error when branches assumed to have dN/dS values less than one are in fact evolving rapidly (Suzuki 2008). However, the validity of these criticisms has been challenged (Yang et al. 2009; Fletcher and Yang 2010; Yang and dos Reis 2011). The application of the test here across seven diverse amniotes

should be robust, since the large number of species, considerable divergence between many species pairs, and the fact that only filtered sequences greater than 100 codons long were tested, mean that there are relatively large numbers of substitutions in each alignment.

6.2.7 Enrichment analysis

Gene Ontology (GO) annotations for chicken genes were downloaded from http://www.geneontology.org/ (Ashburner et al. 2000). GO terms were interpolated to ensure that for each GO term assigned to a gene, all "parental" terms of the GO term were also assigned to that gene. For each GO term, the number of positively selected and non-positively selected sites in genes assigned with that GO term was calculated. A hypergeometric test was then applied in R (R Core Team 2012) to calculate a P-value for each GO term that represents the probability that the number of positively selected sites observed to be associated with a GO term (or greater number than this) would be seen by chance if positively selected sites were distributed randomly across the genes. The hypergeometric distribution is a discrete probability distribution that describes the probability of receiving X 'successes' (positively selected codon sites in this case) in n draws out of a 'population' (total number of codon sites) of size N. There is no replacement between draws, which is appropriate since a single site cannot be called as being positively selected multiple times.

A Bonferroni correction was then applied to account for multiple testing (Bonferroni 1936), producing the adjusted P-value that is quoted in the text. The Bonferroni correction is quite conservative since it simply divides the original P-value by the number of tests conducted.

6.3.8 Homology prediction

Homologues of human *CCDC147* were predicted using profile-based iterative searches with the HMMer3 (Eddy 2009), and later the more sensitive HMMer2 (Eddy 1996), algorithms. The algorithms are based on profile hidden Markov models, which take a log-odds likelihood approach to scoring alignment following Karlin/Altschul theory. The models searched for

significant sequence similarity between the *CCDC147* sequence and protein sequences in the UniRef50 database (Wu et al. 2006). Sequences with significant *E*-value similarity to *CCDC147* where kept, and the *G. magnirostris* and zebra finch (and later *G. fortis*) *CCDC147* predicted sequences were added to multiple sequence alignments that were aligned using the progressive greedy alignment programme T-Coffee (Notredame et al. 2000). Alignments were inspected manually, and lower quality aligned sequences removed, before a phylogenetic tree of the relationship between the sequences was inferred using a Neighbor-joining tree methodology (Sonnhammer and Hollich 2005).

6.4 Results

6.4.1 A *G. magnirostris* genome assembly

A DNA sample was taken from a *G. magnirostris* individual embryo collected by Celine Clabaut during a field trip to the island of Genovesa (Galápagos) in 2009. Sequencing was performed using the Roche 454 technology with both long read and mate-pairs libraries, and then assembled using Roche's algorithm Newbler, as described in the **Materials and methods**. The resulting assembly contains 991.0Mb across 12,958 scaffolds with a scaffold N50 of 382Kb and a median read coverage of 6.5-fold.

Completeness of the *G. magnirostris* genome assembly was estimated using two approaches. First, I determined the amount of euchromatic sequence that aligns between zebra finch and chicken, but that does not align to *G. magnirostris*. Since chicken is an outgroup to both zebra finch and *G. magnirostris*, one can assume that most sequence present in both the zebra finch and chicken genome assemblies will also be present in the *G. magnirostris* assembly, with rare exceptions where lineage-specific deletions have occurred along the Darwin's finch lineage. Thus, the 122Mb of chicken sequence aligned to zebra finch that is absent from the *G. magnirostris* assembly provides an estimate of the *G. magnirostris* euchromatic genome assembly's incompleteness. Second, the assembly consists of approximately 7.529Gb of

sequence data, and the depth of coverage for reads on assembled contigs peaks at 6.0. Consequently, under a simplifying assumption that all regions of the genome are equally represented in libraries and among successful sequencing runs, an estimate of the true genome size is 7.529 / 6.0 or 1.25Gb. In summary, the *G. magnirostris* genome assembly is estimated to cover approximately 89% of the euchromatic genome or approximately 76% of the complete genome. The estimated 1.25Gb size of the *G. magnirostris* genome is similar to the mean avian genome size of 1.38Gb (Gregory et al. 2007).

Assembly sequence quality was assessed first by examining whether GT-AG dinucleotide splice sites in 6,188 chicken genes, each with a single orthologue in zebra finch and *G. magnirostris*, exhibited apparently substituted nucleotides in aligned *G. magnirostris* sequence. Of 168,849 nucleotides 151 (0.31%) showed sequence changes, providing an estimate of the assembly's nucleotide substitution errors. Although this is higher than error rates inferred in other sequenced avian genomes, such as the 0.05% rate estimated for zebra finch (Warren et al. 2010), it is likely to overestimate the true error rate, because some substitutions will reflect mis-alignments or genuine point mutations. In a second approach, I counted the number of insertions or deletions (indels) that are present in the three-way alignment of zebrafinch with *G. magnirostris* and a *G. fortis* sequence that was recently released with GenBank entry: AKZB00000000.1 (Zhang et al. 2012). If one conservatively assumes that there have been no *G. magnirostris* lineage-specific indels then the upper-bound estimate for the indel error is 1.98 indels per Kb of aligned sequence. These errors will have led to a lowering of the number of protein-coding gene models that were predicted for *G. magnirostris.*

These approaches took advantage of whole genome alignments constructed for *G. magnirostris* and chicken, zebra finch and turkey. 57% of the *G. magnirostris* assembly aligned to chicken and 58% to turkey (**Table 6.1**), which is similar to the 58% and 56% of the

zebra finch assembly that aligned to chicken and turkey, respectively (Dalloul et al. 2010). A large proportion (83%) of the Darwin's finch genome could be aligned to zebra finch (**Table 6.1**), consistent with their more recent ancestry than with chicken or turkey, which are both galliformes.

Table 6.1: Amount of sequence aligned between *G. magnirostris* and genome assemblies from other avian species.

Species Pair	Assembly size (Mb)		Aligned sequence (Mb)	Percentage of the *G. magnirostris* genome aligned (%)
	Target species	Query species		
G. gallus – *G. magnirostris*	1037	991	569	57
M. gallopavo – *G. magnirostris*	1046	991	578	58
T. guttata – *G. magnirostris*	1058	991	823	83

6.4.2 Recently acquired repetitive elements absent from the assembly

One expects this *G. magnirostris* genome assembly to be most incomplete within highly repetitive sequence. Use of either a library of transposable element sequences constructed from the *G. magnirostris* genome, using RepeatScout (Price et al. 2005), or a zebra finch repeat library resulted in the identification of 3.3% or 4.1% of the assembly as being repetitive, respectively. This proportion is over two-fold lower than observed for zebra finch or chicken genomes (International Chicken Genome Sequencing Consortium 2004; Warren et al. 2010), and it is clear that there is a deficit of closely-related transposable elements present in the *G. magnirostris* assembly (**Figure 6.5**). Highly repetitive sequence in the *G. magnirostris* genome is thus likely to be disproportionately missing from the assembly.

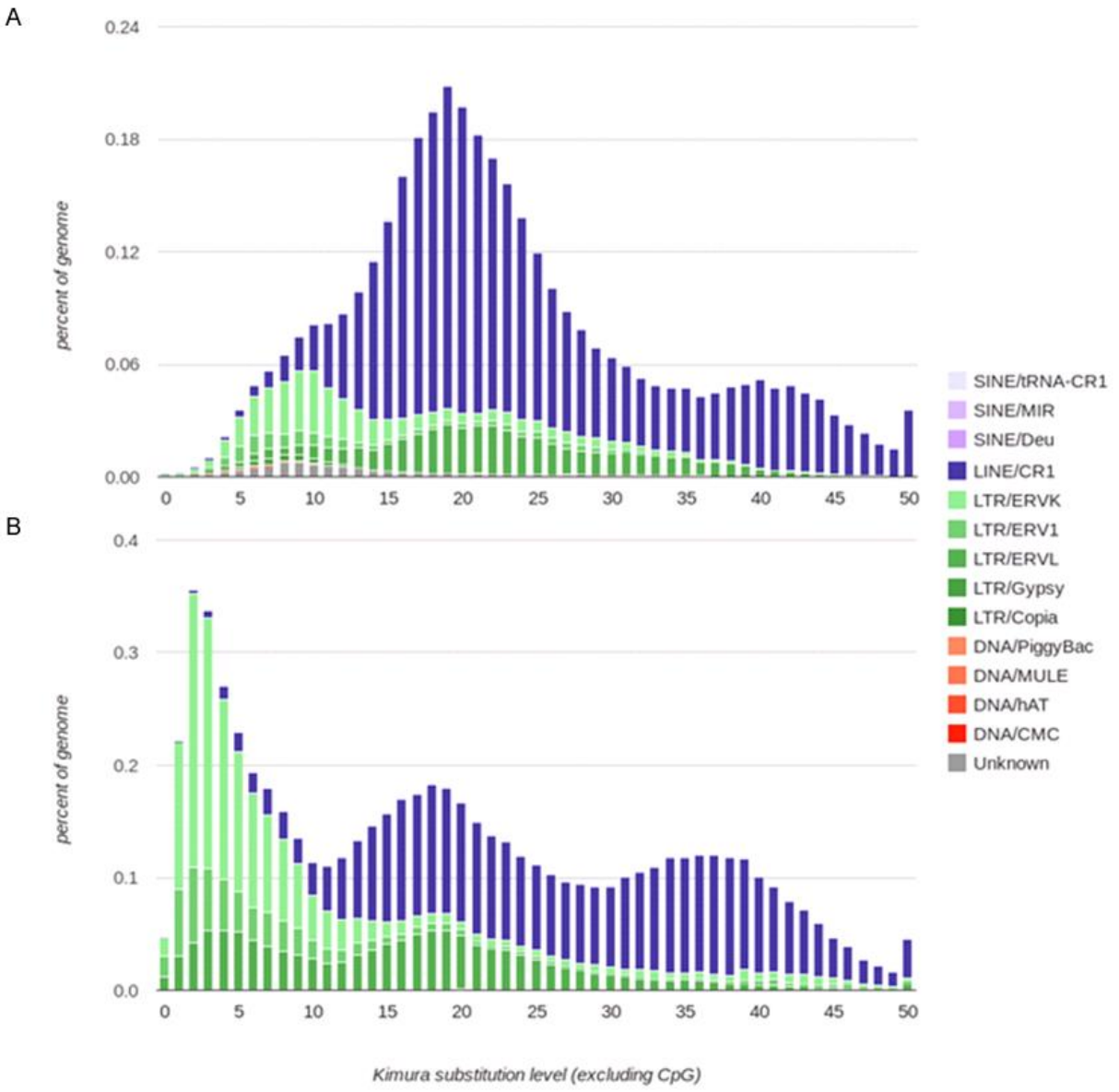

Figure 6.5: Histograms showing the divergence of transposable element (TE) sequences relative to their consensus sequences for A. *G. magnirostris* TEs and B. zebra finch TEs. Those that are more diverged are more likely to be older. The paucity of lowly diverged TEs in the *G. magnirostris* genome assembly indicates that it is likely to be most incomplete within repetitive sequence. The figures were generated using scripts from Juan Caballero available at https://github.com/caballero/RepeatLandscape.

6.4.3 Base composition patterns in *G. magnirostris* similar to other avian genomes

The *G. magnirostris* genome assembly has a G+C proportion of 40.1%, which is similar to all other evaluated amniote genomes. Medium-sized scaffolds (sizes between 2398bp and 46677bp) were more G+C-rich (44.6%) than small or large scaffolds (41.2% and 39.8%, respectively). Visual inspection of the *G. magnirostris* genome reveals that it exhibits substantial spatial heterogeneity in its base composition; similarly to all other amniotic genomes, but unlike that of the *Anolis* lizard (Fujita et al. 2011), genic G+C content of genomic regions has remained relatively constant (**Figure 6.6**).

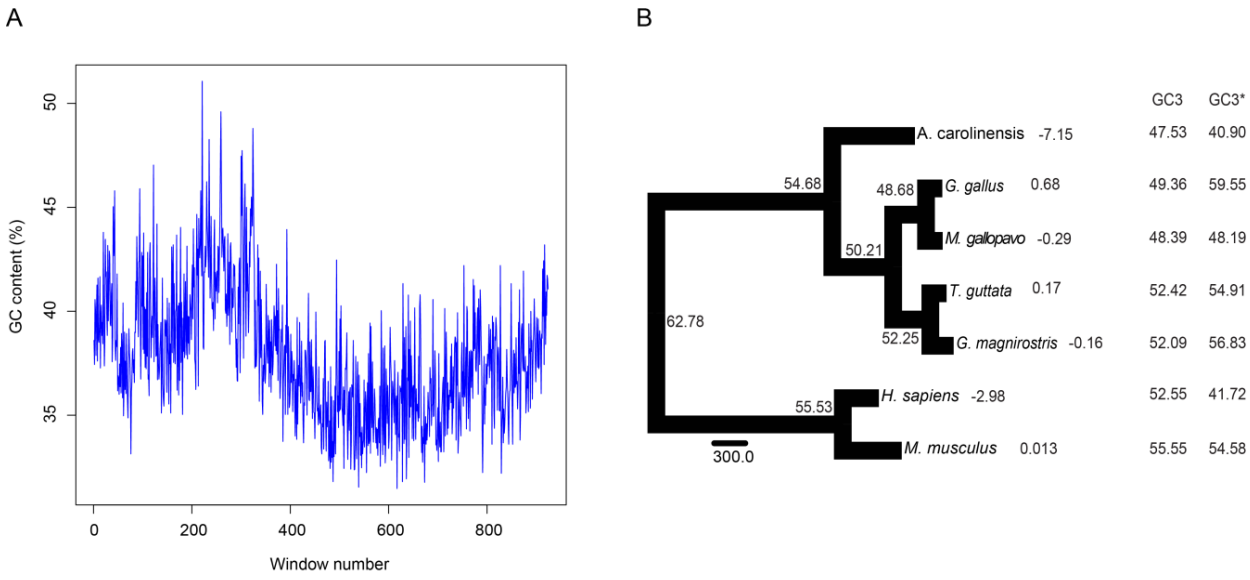

Figure 6.6: GC content distribution in *G. magnirostris*. A. shows the variation of GC content in 3Kb windows along scaffold 10304, the largest scaffold in the assembly. B. shows the third codon position GC content (GC3) and the equilibrium GC3 (GC3*) percentages in different vertebrate lineages. The predicted increase in G+C content along the Darwin's finch lineage is consistent with the maintenance of GC-rich isochores.

However, due to the fragmented nature of the genome assembly it was far from trivial to determine that the base composition of the genome was similarly heterogeneous to other sequenced avian genomes. In fact preliminary whole genome analyses suggested the assembly had a relatively homogeneous base composition, which is intermediate between the other sequenced birds and the anolis genome, and this initially did not appear to be attributable to the incompleteness of the genome assembly since the trend appeared to hold true for aligned sequence (**Figure 6.7**, **Figure 6.8**).

However, it was puzzling that these results did not appear consistent with the predictions based on GC3 content and visual inspection of aligned regions (**Figure 6.9**). In a further analysis the variation in G+C content was examined across relatively short contiguous segments of 2–10kb in length. Crucially, these windows were only constructed across contiguous sequence, unlike previous approaches. The result of this is that aligned G+C content appears no more homogeneous compared to the other birds, and thus the isochore structure in Darwin's finch appears similar to other sequenced endotherms. The previous results for aligned sequence can therefore be explained as G+C content was examined in windows that cross alignment blocks and scaffold boundaries, and this amalgamation flattens the base composition landscape. This is because G+C content does vary across the genome, but this variation is not detected if G+C content is calculated over non-contiguous sequence.

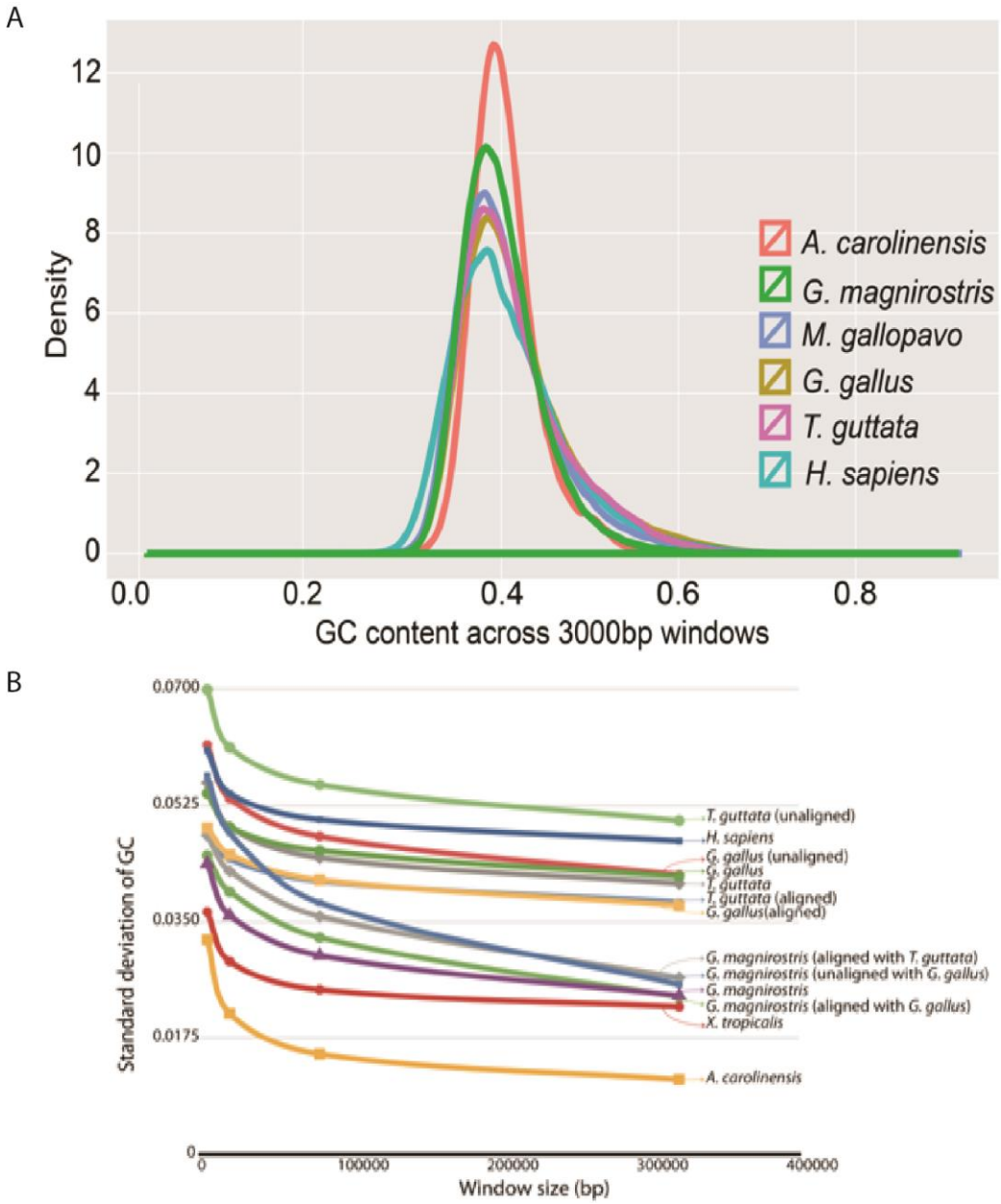

Figure 6.7: Apparent base composition homogeneity in the Darwin's finch genome. A. G+C content appears to be more homogeneous than the other bird species and human as shown by the more peaked distribution. B. G+C content appears less variable in Darwin's finch, even for sequence aligned the zebra finch and chicken genomes.

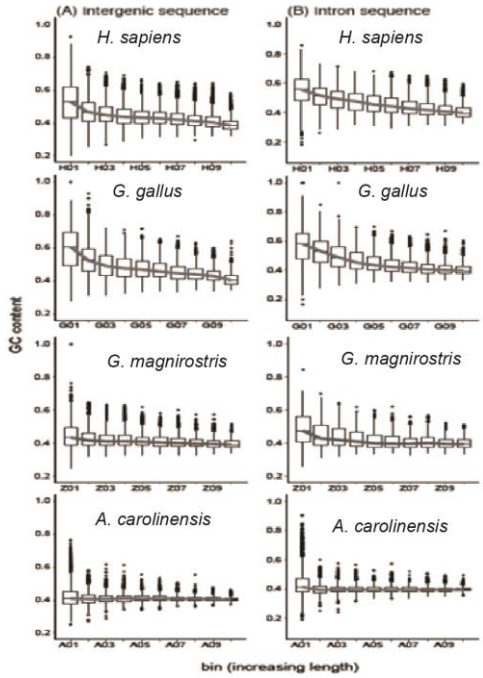

Figure 6.8: G+C content in noncoding sequences. G+C content is negatively correlated with increasing bin size, but this trend appears to be less extreme for *G. magnirostris* than for human and chicken.

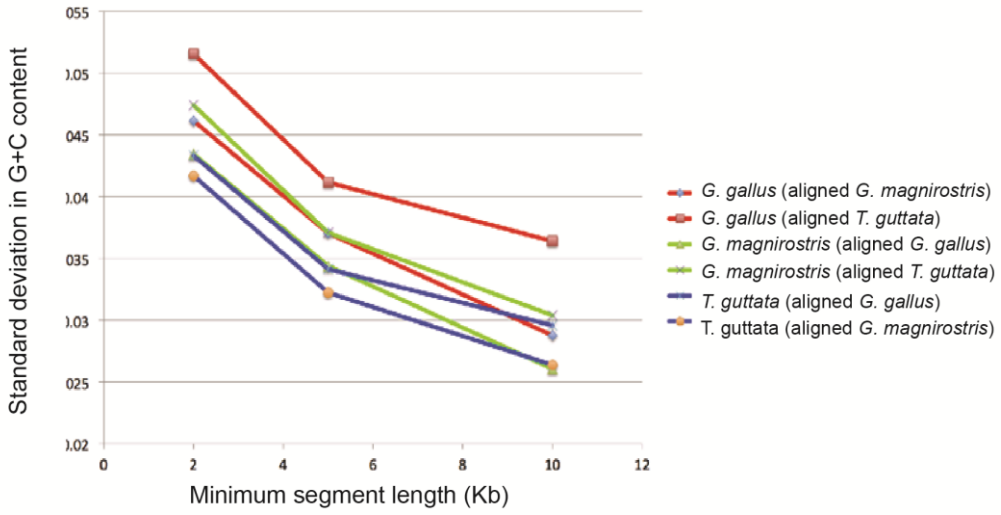

Figure 6.9: The variation in G+C content across contiguous aligned segments between different avian species. The variation in *G. magnirostris* is similar to that observed in chicken and zebra finch.

6.4.4 *G. magnirostris* predicted genes and orthologues

13,291 protein-coding genes were predicted in the *G. magnirostris* genome assembly. Protein-coding sequences were aligned from three amniote species (human, chicken, and zebra finch) to the *G. magnirostris* genome assembly, and overlapping transcript predictions were reconciled using the Gpipe pipeline (Heger and Ponting 2007). To analyse the evolution of *G. magnirostris* protein-coding genes, the orthologues and paralogues among *G. magnirostris* and seven other Euteleostomi (human, mouse, chicken, turkey, zebra finch, *Anolis* lizard and tetraodon) were assigned using the OPTIC pipeline (Heger and Ponting 2007; Heger and Ponting 2008). I then produced a high quality set of 1,452 simple orthologue sets (genes that have been spared from duplication or deletion in the bird, reptile and mammalian lineages since their last common ancestor) among the seven amniote species. These 1,452 gene sets represent a stringent set of evolutionarily conserved "core" protein-coding genes in vertebrates.

Examining the completeness of these gene sets, it was noted that there were 10,222 simple 1:1 orthologue sets between human and zebra finch, while there were only 7,416 simple 1:1 orthologue sets between human and *G. magnirostris*. The smaller gene orthologue set between human and *G. magnirostris* could imply that 27% of genes are missing from the gene set, and thus the gene set could be 73% complete. A similar proportion (71%) of 1,109 metazoan single copy orthologues curated by Creevey et al. (2011) have orthologues among our predicted *G. magnirostris* genes. Our approaches ensure that each gene in these orthologue sets has at least one transcript that covers at least 80% of the human, chicken or zebrafinch template transcript. One should note that these gene set completeness estimates are lower-bound estimates for assembly completeness since this orthology analysis will exclude some partially, imperfectly or fragmentary predicted *G. magnirostris* gene models.

6.4.5 Evolutionary rates across Darwin's finch, passerines, and galliformes

Evolutionary rates (dS, dN, and dN/dS values) were estimated for the filtered alignments for the 1,452 sets of orthologues for seven amniote species (**Figure 6.10**). The median dS value for the *G. magnirostris* lineage (0.051) is over 15-fold larger than our predicted nucleotide error rate (0.31%; see above), which indicates that sequencing errors will have little effect on most of our comparative genomic analyses. The estimated median dS value between zebra finch and *G. magnirostris* (dS = 0.093) is similar to that for chicken and turkey. Divergence of chicken and turkey lineages occurred approximately two-fold earlier, estimated at 44–59My ago from mitochondrial and *cyt b* DNA sequences using a Bayesian framework informed by fossil data (Pereira and Baker 2006), than the presumed zebra finch and *G. magnirostris* lineages split, which was approximately 25My ago. This implies that neutral evolution was approximately two-times faster in the zebra finch and *G. magnirostris* lineages than in the chicken and turkey lineages, which is consistent with previous findings (Nabholz et al. 2011). A similarly elevated neutral evolutionary rate observed for the rodent lineage has been ascribed to their shorter generation times and their greater rate of DNA replication errors during germ cell division (Li et al. 1996). The generation time of chicken is approximately two years (Keightley and Eyre-Walker 2000), shorter than that of extant generation time of approximately three to five years for *Geospiza* species (Price et al. 1984). Nevertheless, the relatively rapid rate of neutral evolution for the zebra finch or *G. magnirostris* lineages would be consistent with historic generation times, over the last 25 million years, for their ancestral species being much shorter than for extant ones.

The lineage-specific median dN/dS value is slightly smaller for *Geospiza* than it is for zebra finch (**Figure 6.10**). Since smaller dN/dS values are expected for lineages with larger effective population sizes, N_e (Ellegren 2009), the result is surprising, because the very low N_e values of 38–60 of extant *Geospiza* species (Grant and Grant 1992) are much smaller than

the current effective population size of zebra finch (25,000–7,000,000) (Balakrishnan and Edwards 2009). A possible explanation for the lower *Geospiza* specific dN/dS values is that purifying selection may have been stronger in Darwin's finches than Zebra finches due to environmental factors. The unusual archipelago environment of the Galápagos Islands could have led to stronger selective pressures than those experienced on the mainland.

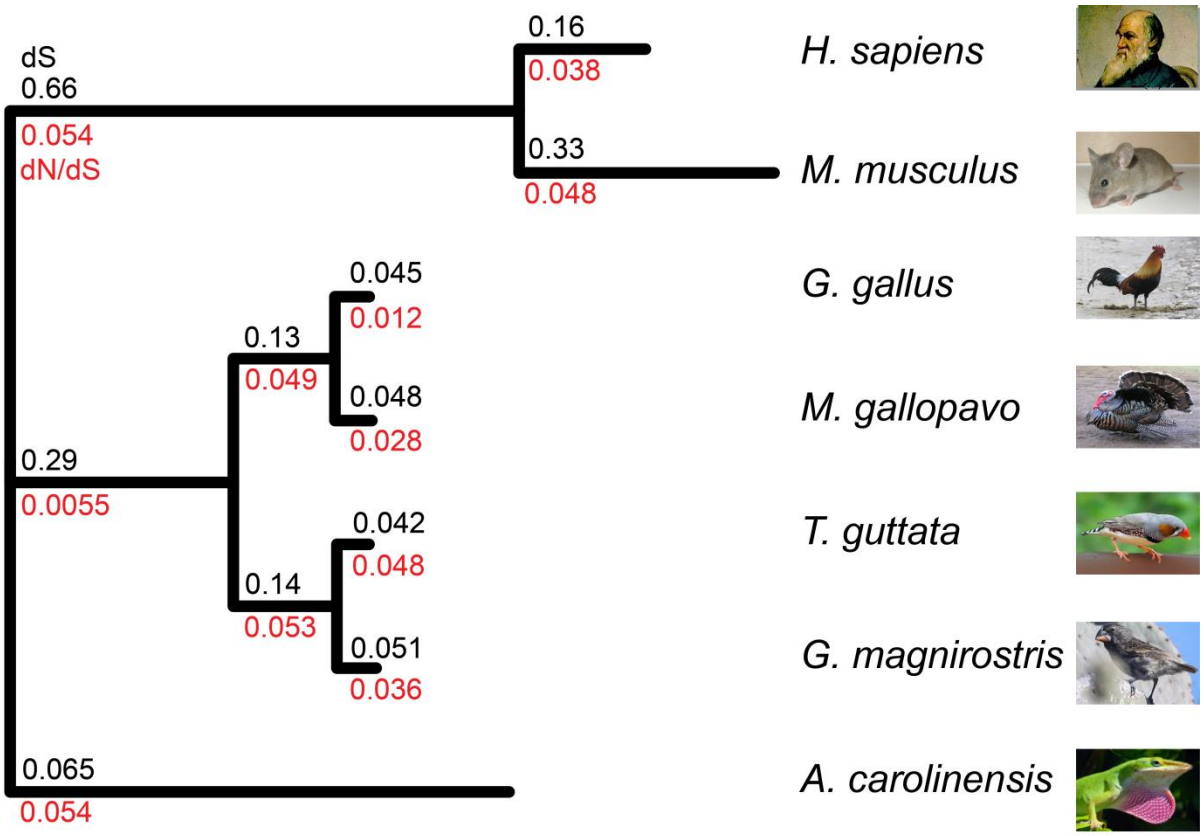

Figure 6.10: Phylogeny of seven amniotic species. Branch lengths are proportional to dS; the degree of constraint (dN/dS) for each terminal lineage is also indicated (values shown in red). Evolutionary rates (dS and dN/dS) are median values deriving from 1,452 alignments of simple one-to-one orthologues present in each species.

6.4.6 Positive selection in Darwin's finches and across the passerine lineage

For each of the 1,452 sets of orthologs I next inferred amino acid sites that evolved under positive selection along the *G. magnirostris* lineage, and each of the other three avian lineages. For this I used the branch-sites method (Zhang et al. 2005) and a Bayes Empirical Bayes approach (Yang et al. 2005) to predict sites that evolved under positive selection (those with a posterior probability greater than 95% of falling in a site class where dN/dS = ω >1 along a defined branch; **Figure 6.11**). This procedure resulted in predicting 21, 16, 24 and 51 positively-selected genes (PSGs) in *G. magnirostris*, zebra finch, chicken and turkey lineages, respectively (**Figure 6.11**). This is far fewer than reported previously in avian genomes (Nam et al. 2010), which likely reflects the lower number of genes that I analysed, the fact these genes are from a more widely conserved orthologue set, and the stringent filters on aligned sites that I needed to employ to discard potentially misaligned or poor quality sequence. Three of the *G. magnirostris* PSGs (Ubiquitin carboxyl-terminal hydrolase; Ubiquitin carboxyl-terminal hydrolase 47; and *IGF2R*) may have been subject to GC-biased gene conversion (Duret and Galtier 2009) as indicated from their relatively high numbers of AT→GC substitutions (**Table 6.2**).

Genes that are predicted to have been under positive selection in the *G. magnirostris* lineage have elevated values of dN/dS in that lineage, but not the zebra finch lineage, and vice versa (**Figure 6.11**). Of the 21 *G. magnirostris* PSGs (**Table 6.2**), three were identified as PSGs in other avian lineages: xanthine dehydrogenase (*XDH*), perhaps as a result of its role in the innate immune system (Vorbach et al. 2003), mitochondrial ATP binding cassette (ABC) transporter, *ABCB10*, which is essential for erythropoiesis (Hyde et al. 2012) and nebulin (*NEB*), which encodes a large muscle protein (Pappas et al. 2011).

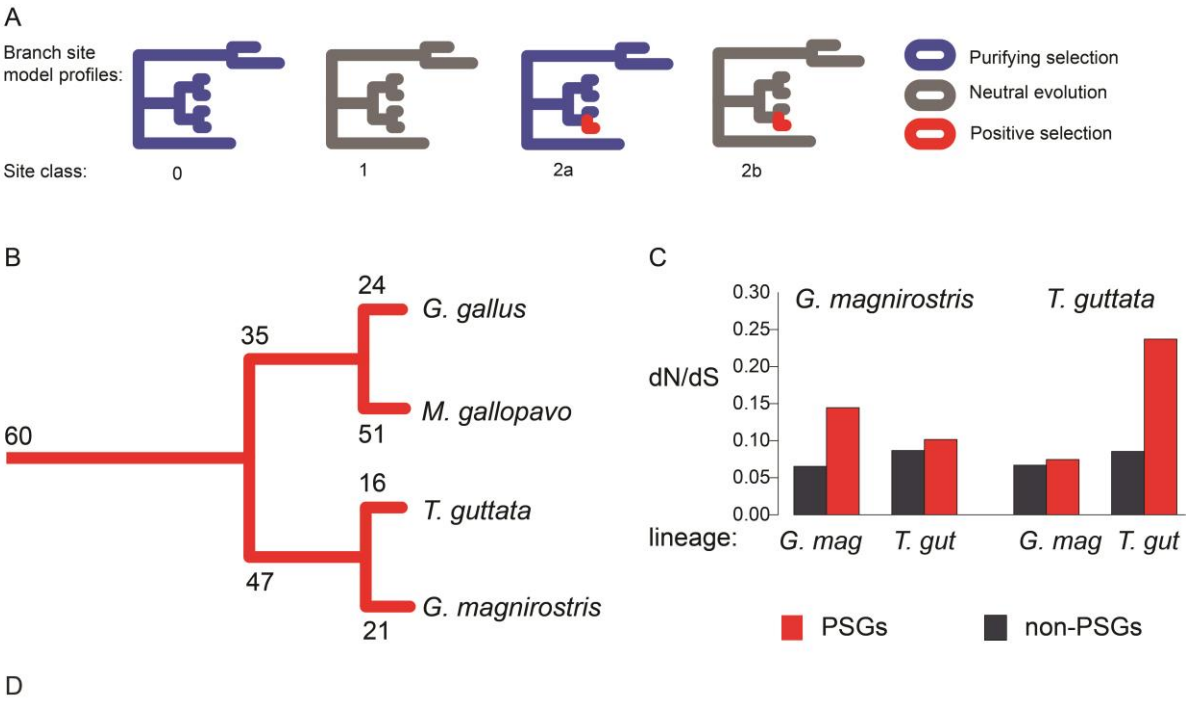

Figure 6.11: Positive selection analyses. A. Branch-site test models of Zhang et al. (2005). The schematic represents the positive selection model. Under the null model, sites fall into site classes 0 or 1 only. The two models are compared using a likelihood ratio test. B. The number of positively selected genes identified on *G. magnirostris*, zebra finch, passerine, chicken, turkey, galliform, and avian branches. C. Average levels of dN/dS for the *G. magnirostris* or zebra finch lineages for *G. magnirostris* and zebra finch positively selected genes (PSGs) and for non-PSGs inferred by parsimony.

Table 6.2: Base composition properties of *G. magnirostris* positively selected genes. The genes in bold show a high rate of AT→GC changes.

Short gene name	Ensembl gene ID of chicken 1:1 ortholog (omitting ENSGALG prefix)	Length of filtered aligned sequence (bp)	GC content proportion	Number of GC→ AT changes	Number of AT→ GC changes	*G. magnirostris* equilibrium GC content (GC*)
FKBP6	00000000837	340	0.539	3	4	0.609
MFF	00000003079	513	0.517	2	2	0.517
ASB6	00000004378	362	0.624	2	1	0.453
SART3	00000004887	924	0.524	2	1	0.270
UBP47	**00000005569**	**1262**	**0.448**	**2**	**20**	**0.890**
TRAF7	00000005767	662	0.568	5	4	0.513
XDH	00000008701	2353	0.491	14	19	0.567
E1BY77	**00000008909**	**785**	**0.497**	**2**	**9**	**0.816**
F1N8A7	00000010043	622	0.469	0	3	N/A
P2RY1	00000010357	763	0.482	9	1	0.0939
F1NDU4	00000011096	424	0.453	2	1	0.293
PRKAG3	00000011360	657	0.556	3	3	0.556
ANO10	00000011513	694	0.433	7	5	0.353
IGF2R	**00000011621**	**2501**	**0.446**	**8**	**27**	**0.731**
F1NIP9	00000012138	357	0.336	0	1	N/A
LRR1	00000012230	278	0.607	0	4	N/A
C7orf25	00000012333	737	0.482	5	12	0.691
Q9DEH4	00000012495	5369	0.509	26	37	0.596
ARSK	00000014672	488	0.391	4	5	0.445
F1NR67 (POU1F1)	00000015495	411	0.533	2	5	0.740
E1BV11	00000016811	389	0.388	3	2	0.297

Two *G. magnirostris* PSGs are of particular note: *POU1F1* (POU domain, class 1, transcription factor 1; also known as Pit1, growth hormone factor 1) and *IGF2R* (insulin-like growth factor 2 receptor). These genes' putatively adaptive amino acid substitutions were confirmed using sequence data from two other Darwin's finch species, *G. fortis* (medium ground finch) (Zhang et al. 2012) and from *G. difficilis* (sharp-beaked ground finch) (**Figure 6.12**). Disruption of either gene in the mouse is known to result in craniofacial abnormalities (Snell 1929; Wang et al. 1994) and *POU1F1*, despite its description as a pituitary-specific transcription factor in mammals (Ingraham et al. 1988), is differentially expressed in the developing beaks of ducks, quails and chickens (Brugmann et al. 2010). There is a functional link between these two genes since *POU1F1* regulates prolactin and growth hormone genes in mammals and birds (Weatherly et al. 2001), and decreased growth hormone results in a decrease in activity of the insulin/IGF-1 signalling pathway (Hsieh et al. 2002). In mouse bone, growth hormone is known to regulate many genes of the *insulin*/*IGF-1* or *Wnt* signaling pathways, as well as *Bmp4* (Govoni et al. 2006) whose gene expression change is linked to bill morphology in *G. magnirostris* (Abzhanov et al. 2004). Moreover, a key member of the *IGF* pathway (IGF binding protein, a molecule that controls ligand-receptor interaction) was identified in Darwin's finches as one of the top differentially expressed candidate genes in a microarray screen in species with divergent beak shapes (Abzhanov et al. 2006). Positive selection acting on *POU1F1* and *IGF2R* may thus have contributed to the evolution of beak morphology in the *G. magnirostris* lineage.

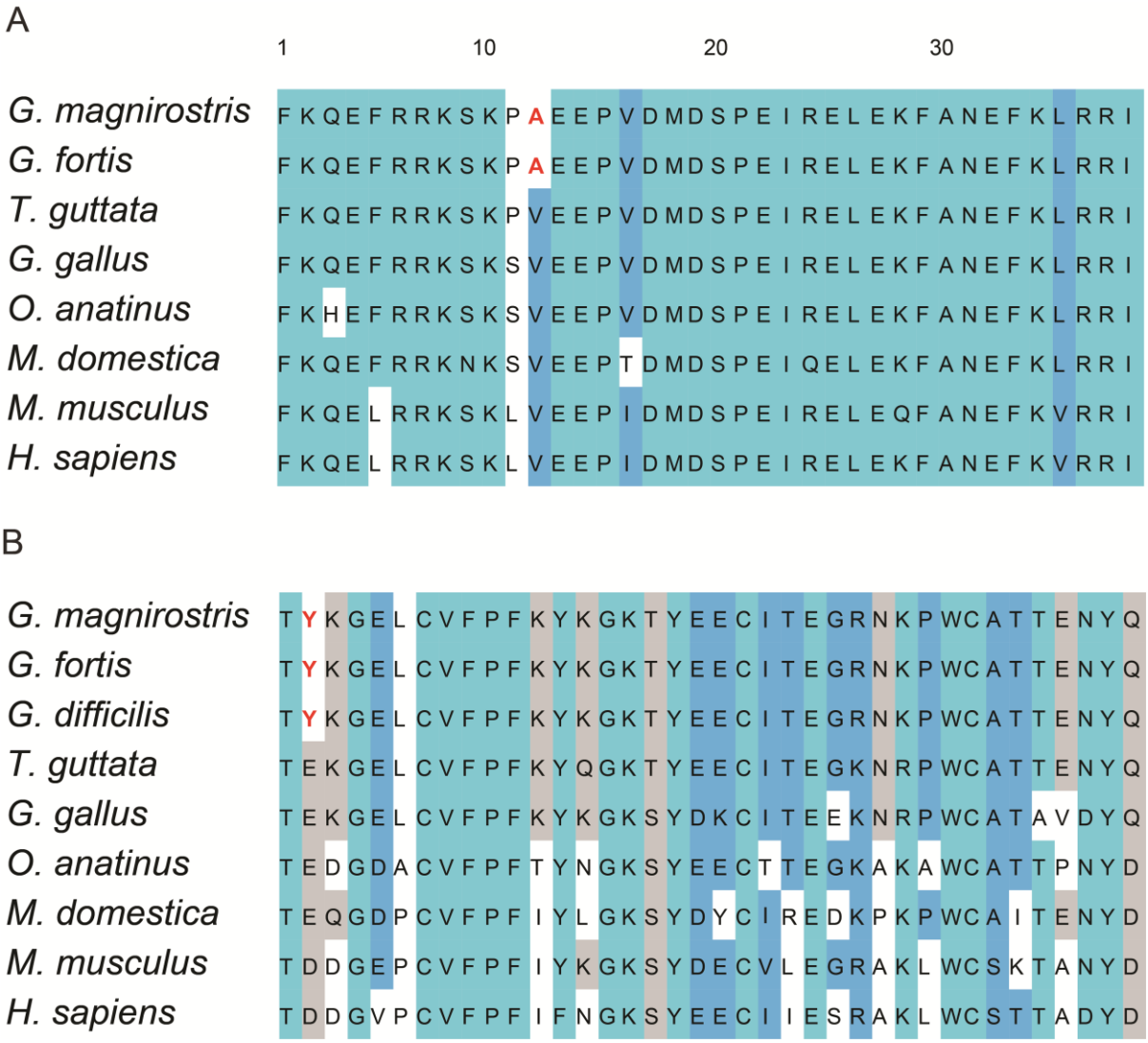

Figure 12: Alignment showing the candidate *Geospiza* positively selected codon sites (highlighted in red) in A. *POU1F1* and B. *IGF2R*. Alignment visualised with the belvu software (Sonnhammer and Hollich 2005).

I also predicted 47 genes to have undergone positive selection on the passerine branch prior to the split of the zebra finch and *G. magnirostris* lineages (**Table 6.3**). I performed an enrichment analysis to test whether any Gene Ontology (GO) terms (Ashburner et al. 2000) were overrepresented among genes with positively selected sites along the passerine branch compared to other genes in the stringent 1:1 ortholog set of 1,452 genes. This test identified 'cilium' (GO:0005929) as the most significantly enriched term ($p = 8.1\times10^{-20}$; **Table 6.4**). This term is annotated to three passerine PSGs: coiled-coil domain containing 40 (*CCDC40*), axonemal dynein intermediate chain 2 (*DNAI2*), and cytoplasmic dynein 2 light intermediate chain 1 (*DYNC2LI1*). *DNAI2* protein is a component of respiratory ciliary axonemes and sperm flagella, and human *DNAI2* mutations are associated with respiratory tract dysfunction and infertility (Loges et al. 2008). *DYNC2LI1* is present in the mammalian ciliary axoneme (Perrone et al. 2003). Two further passerine PSGs, namely coiled-coil domain containing 147 (*CCDC147*) and its paralogous gene, coiled-coil domain containing 146 (*CCDC146*), are likely to possess functions related to cilia and spermatazoan flagella (see below), although this is not reflected in current GO annotations.

Table 6.3: Positively selected genes along the passerine branch. P-values of less than 0.01 are highlighted in bold.

Short gene name	Ensembl gene ID of chicken 1:1 ortholog (omitting ENSGAL prefix)	P-value that gene is under positive selection
RNF207	00000000710	0.015
FOXRED1	00000001033	0.013
E1C8X7	00000001226	0.040
SAMHD1	00000001231	**0.0055**
F1NVV4	00000002466	0.039
F1P582	00000002490	0.038
GMPPB	00000002500	0.016
E1BQB6	00000002891	0.011
PMS2	00000003430	0.030
SGMS1	00000003701	**0.0077**
TTYH3	00000004310	0.043
DNAI2	00000004495	**0.0072**
E1BQH2	00000004813	0.043
F1N867	00000005693	0.047
SLC52A3	00000006194	0.018
LRRC33	00000006402	0.037
C12H3orf19	00000006459	0.045
SETD5	00000006571	0.033
CCDC40	00000007042	**0.0000017**
FAM46D	00000007157	0.018
CLEC16A	00000007167	0.030
PDIA1	00000007233	**0.0053**
Q5ZJK3	00000007651	0.019
ARMC9	00000007691	**0.0082**
PSME4	00000008163	0.046
TRIM36	00000008188	0.035
CCDC146	00000008309	0.013
CCDC147	00000008417	**0.00096**
ATL2	00000008515	**0.000013**
XDH	00000008701	0.011
E1BRU9	00000008813	0.014
LBR	00000009305	**0.0098**
LRRC34	00000009402	**0.00010**
RAD51B	00000009491	0.038
DYNC2LI1	00000009954	**0.0019**
AKR1A1	00000010244	0.043
SLC4A4	00000011604	0.029
Q7ZSX8	00000011715	0.018
DTX3L	00000012075	0.040

Short gene name	Ensembl gene ID of chicken 1:1 ortholog (omitting ENSGAL prefix)	P-value that gene is under positive selection
CHST11	00000012698	0.018
FUCA2	00000013773	**0.000023**
Q5ZL94	00000015916	0.022
DCA13	00000016073	0.046
Q5ZLT2	00000016323	**0.0033**
F1NHR2	00000016558	**0.0026**
ATM	00000017159	**0.0046**
GPR162	00000022926	0.030

Table 6.4: The Gene Ontology (GO) enrichments for positively selected genes along the passerine branch.

GO term	GO term description	Enrichment P-Value
GO:0005929	cilium	8.14×10^{-20}
GO:0071844	cellular component assembly at cellular level	3.63×10^{-9}
GO:0022607	cellular component assembly	8.07×10^{-8}
GO:0048858	cell projection morphogenesis	1.55×10^{-7}
GO:0044085	cellular component biogenesis	2.48×10^{-7}
GO:0044463	cell projection part	3.17×10^{-7}
GO:0032990	cell part morphogenesis	6.66×10^{-7}
GO:0042995	cell projection	1.25×10^{-6}
GO:0030030	cell projection organization	2.25×10^{-5}
GO:0000902	cell morphogenesis	9.19×10^{-4}
GO:0048646	anatomical structure formation involved in morphogenesis	1.35×10^{-3}
GO:0007017	microtubule-based process	9.23×10^{-3}
GO:0043623	cellular protein complex assembly	1.77×10^{-72}

CCDC147 is of particular interest as it has evolved unusually rapidly along the passerine branch (**Figure 6.13, Figure 6.14**). It is predicted to harbour 40% more positively selected sites than any other gene inferred for any branch, making it the most pervasively positive selected of all the genes I tested. 27 codon sites in *CCDC147* that are shared by *G. magnirostris* and zebra finch were identified as having been subject to positive selection (posterior probability of >95%), and all 27 of these codon site changes were validated using *G. fortis* sequence data with GenBank entry: AKZB00000000.1 (Zhang et al. 2012). It is likely that vertebrate *CCDC147* and *CCDC146* homologues encode spermatazoan flagella proteins because its *Chlamydomonas reinhardtii* homologue MBO2 (Carvalho et al. 2001; Tam and Lefebvre 2002) is a flagellar protein, and its fruitfly homologues are involved in fertility: *ORY* maps to the *ks-1* fertility factor region, *CG5882* homozygous mutants are sterile (Mummery-Widmer et al. 2009), and *CG6059* is specifically expressed in the testis. In addition, human *CCDC147* shows the strongest differential expression in the testis, http://www.ebi.ac.uk/arrayexpress/experiments/E-GEOD-7307 (Kapushesky et al. 2012). The positive selection I infer across five passerine genes (*CCDC40*, *DNAI2*, *DYNC2LI1*, *CCDC146* and, most pervasively, *CCDC147*) thus could have been a consequence of sperm competition, a potent selective pressure that has been implicated previously in adaptive changes (Swanson and Vacquier 2002). Proportionally fewer spermatozoa arrive at the site of fertilisation in zebra finches than in chickens (Birkhead et al. 1993), implying that sperm competition could have been particularly intense in early passerine evolution.

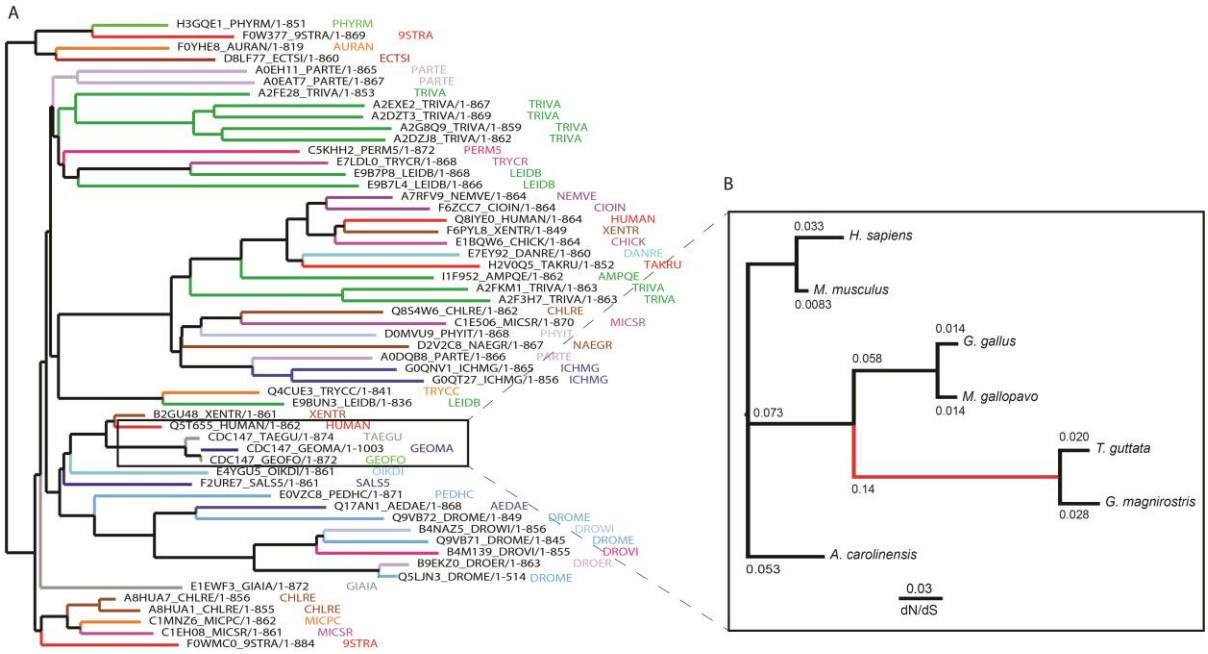

Figure 6.13: Gene trees showing the evolution of *CCDC147*. A. Phylogenetic reconstruction of gene tree from alignment with the Belvu software (Sonnhammer and Hollich 2005). The tree is constructed from the multiple sequence alignment with a Neighbour-joining tree approach. B. Lineage-specific dN/dS values estimated for the *CCDC147* gene across aminotes. The long passerine branch highlighted in red is inferred to have experienced many events of positive selection.

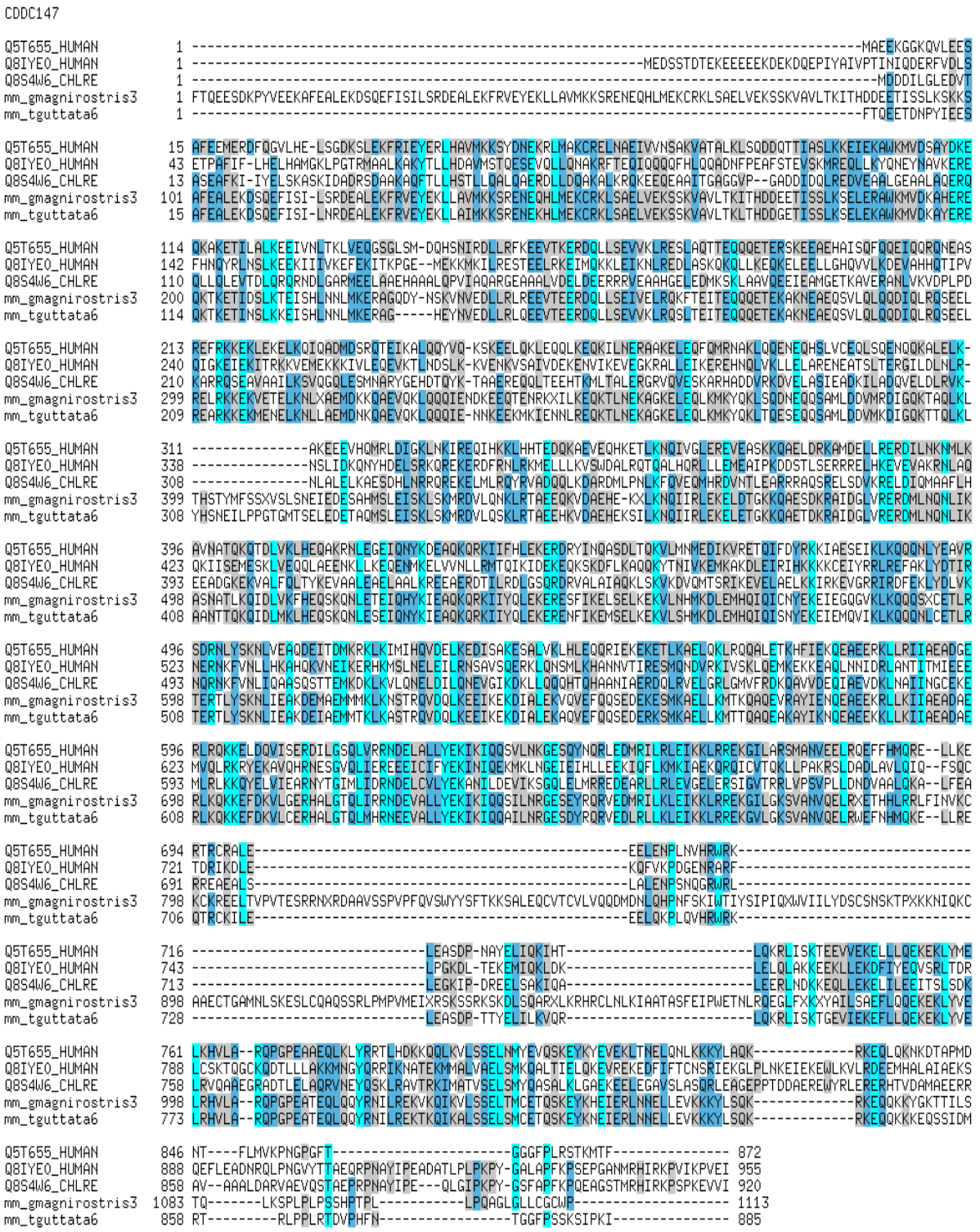

Figure 6.14: Alignment of *G. magnriostris* and zebra finch CCDC147 between human CCDC147, human CCDC146, and algae MBO2. The alignment was visualised with the Belvu software (Sonnhammer and Hollich 2005).

6.5 Discussion

I have described the sequencing, assembly, and analysis of a Darwin's finch genome. While I found the genome assembly to be of relatively low quality, and this presented a number of technical challenges, subsequent downstream analyses still provided informative results about Darwin's finch and more general avian biology. I found the rate of protein evolution to be similar between *G. magnirostris* and zebra finch as measured by dS, but dN/dS ratios are slightly lower in *G. magnirostris*, which is surprising given their current very small population size. I identified a number of candidate positively selected genes specific to the Darwin's finch lineage with particular amino acid residues that show statistical evidence of positive selection. Two of these genes, *POU1F1* and *IGF2R*, are of special interest since they may be involved in beak development. Experiments that misexpress *POU1F1* or *IGF2R* variants during avian craniofacial development will be required to further investigate this hypothesis.

This first genome sequence of a Darwin's finch has utility beyond the purview of Darwin's finch biology. Avian species are currently under-sampled as a taxonomic group compared with mammals. Moreover, the passerine order contains over half of all bird species, which equates to approximately 5,000 identified species, almost as many as the total number of mammalian species (Mayr 1946; Wilson and Reeder 2005). However, passerine genomes are underrepresented and the range of genome-scale resources presented here should facilitate further research into the evolution of this unusual group of passerine birds. My identification of positively selected genes on the passerine branch not mentioned in previous studies that used only the zebra finch genome sequence (Nam et al. 2010; Warren et al. 2010), demonstrates the extra power this additional passerine sequence provides for investigating wider avian biology.

Chapter 7: Conclusions and future perspectives

I have developed and utilised evolutionary and comparative genomic methods as tools to inform vertebrate functional biology. By classifying genome sequences into regions that are subject to natural selection, either purifying or positive selection, I have identified putatively functional sequence in mammalian and avian genomes. Furthermore, I have characterised this candidate functional sequence based on its genomic location using catalogues of previously defined biological and biochemical annotations such as genes, proteins, untranslated regions (UTRs), transcription factor binding sites (TFBSs), enhancers, promoters, DNase1 hypersensitivity sites, long noncoding RNAs, transposable elements, and gene ontology annotations (Ashburner et al. 2000; Curwen et al. 2004; Dunham et al. 2012).

In **Chapter 3**, I developed and refined approaches for estimating the quantity of sequence mutually constrained with respect to insertions and deletions (indels) between two sequences (a quantity I term $\alpha_{selIndel}$), building on previous work (Lunter et al. 2006; Meader et al. 2010). I explained several theoretical nuances of a neutral indel model (NIM1) and refined the way in which NIM1 calculates the bounds of $\alpha_{selIndel}$ estimates. I described a new neutral indel model (NIM2) that estimates $\alpha_{selIndel}$ with a likelihood approach. I introduced a log-odds approach to trim whole genome alignments of poorly aligned sequence and show that this substantially improves the robustness of $\alpha_{selIndel}$ estimates to variation in assembly quality and alignment build. I also performed extensive genome simulations that provided validation for the approaches over a wide range of parameterisations. These methods make a significant advancement over previous approaches for estimating $\alpha_{selIndel}$.

I applied these improved approaches for estimating $\alpha_{selIndel}$ to human and other mammalian genomes in **Chapter 4**. I found that estimates of $\alpha_{selIndel}$ are strongly negatively correlated with the divergence between the mammalian species pair, consistent with previous

observations (Smith et al. 2004; Meader et al. 2010). This correlation implies that functional sequence turns over rapidly as it is lost and gained over relatively short evolutionary time periods (Ponting et al. 2011). Extrapolating the trend of turnover back to a divergence of zero, I estimated that 7.1–9.2% of human genomes is subject to purifying selection now and thus likely to be currently functional. Comparing my estimate to previous estimates of the quantities of constrained sequence in the human genome, it is higher than some estimates that miss lineage-specific constrained sequence and lower than others that may be inflated by technical artefacts (Ponting and Hardison 2011).

Furthermore, my estimate that approximately 8% of the human genome has evolved under negative selection is around ten-fold lower than the quantity of sequence covered by the ENCODE defined elements (Dunham et al. 2012; Ecker et al. 2012; Pennisi 2012). Such observations have led to animated discussions over what constitutes functionality in genome sequences (Eddy 2012; Doolittle 2013; Eddy 2013; Graur et al. 2013). However, I feel that this is a rather sensationalist debate, and it should be acknowledged that functionality in biology is an intuitive but vague concept, with no single definition suitable across all contexts. The large discrepancies between the estimates of the proportion of the human genome that is functional made with evolutionary and biochemical approaches is simply a reflection of the fact that these metrics are measuring very different quantities. Rather than quibble over which metric is most meaningful, it seems more constructive to harness these different approaches together to provide complementary information on the importance and nature of DNA sequences.

Examining the relationship between $\alpha_{selIndel}$ estimates and species divergence stratifying the human data by the type of biological or biochemical element, I found that coding sequence shows high levels of constraint and low rates of turnover, while the opposite is observed for noncoding sequences. The rate of turnover for coding sequence is estimated such that half of

functional coding sequence is lost and re-gained over 1300My (950–2250My), while this half-life for functional noncoding sequence is estimated at just 127My (116–139My). Examining the turnover across constrained human biochemical elements, mainly using those defined by the ENCODE consortium (Dunham et al. 2012), I found that long noncoding RNAs (lncRNAs), enhancer and promoters show evidence of particularly rapid turnover, while DNase I hypersensitive sites have been relatively stable, although these conclusions are sensitive to the annotation set chosen as representative of a particular element type.

This identification and characterisation of the turnover of functional sequence is not merely a point of academic interest, because it has implications for the use of model organisms. Given the high rate of turnover observed over mammalian evolution, it seems reasonable to suggest that model organisms relatively divergent from humans, including mouse, will be of limited utility to inform human biology via sequence homology. While protein sequence analysis has already provided many insights into the function of human proteins, including those implicated in diseases (Sanchez-Pulido et al. 2002; Sanchez-Pulido et al. 2003; Sanchez-Pulido and Ponting 2011), such homology analysis have been much less informative for noncoding sequences. My work provides the basis for a quantitative framework that could be used to help assess the value of model organisms in different contexts. For example, since human lncRNAs show the highest degree of turnover of all the element types I examine, I predict that these will be among the most difficult to examine with model organisms. Note that I am only describing the limited scope of sequence homology analyses, as clearly even highly divergent model organisms can be useful to establish general biological mechanisms of relevance to humans. However, the vast majority of medically relevant studies conducted thus far, at least in non-vertebrate model organisms, have tended to come from homology analyses of protein coding sequences (Botstein et al. 1997; Kaletta and Hengartner 2006).

In **Chapter 5**, I examined patterns of sequence constraint and turnover across the avian lineage using the same approach for estimating $\alpha_{selIndel}$ as applied to mammalian genomes. One previous study examined patterns of turnover in birds using UTR sequences (Kunstner et al. 2011a), but this is the first systematic genome-wide study of the turnover of functional sequence across avian evolution. I found the degree of sequence constraint on avian sequences was positively correlated with G+C content, protein content, and negatively correlated with chromosome size. Furthermore, I infer that avian functional sequence turns over such that the half-life divergence ($d_{1/2}$) is estimated at 0.56 (0.38–1.79) in units corresponding to one expected neutral substitution per site. This estimate is similar to estimates for the mammalian lineage, so it appears that the turnover of functional sequence has proceeded at a comparable rate in the two lineages. Extrapolations suggested that 10.3–16.8% of avian genomes are functional at present, a considerably larger proportion, but possibly a smaller absolute amount, than observed in mammalian genomes.

I provided a comparison of the comparative genomic results between mammals and birds; this meta-comparative genomic approach could be an important area for evolutionary studies in the future. This is a dimension beyond simply comparing patterns between genomes sequences, since it is comparing patterns across entire lineages. Although such comparisons come with their inevitable difficulties and complexities, the plethora of relatively cheap next generation sequencing data becoming available (Shendure and Ji 2008) means that such approaches could be common-place in the future. Meta-comparative genomic studies could inform the study of the genomic basis of interactions between organisms. For example, host-pathogen evolution has been described as undergoing a co-evolutionary arms race, where the pathogen evolves a way of targeting the host, then the host's immune system needs to adapt in response to this, and then the parasite evolves further attributes to overcome the host's response (Nesse and Williams 1994; Gandon et al. 2008). This 'cat-and-mouse game' is

expected to continue indefinitely. At the genomic level, this cycle implies the rapid turnover of functional sequence in both the host and pathogen lineages, which could be molecularly characterised in many biological systems. This is of medical relevance, since it could help us to better comprehend antibiotic resistance. Understanding organismal interactions in terms of functional sequence turnover would be informative in not just antagonistic symbioses, but also positive interactions where both interacting partners involved receive benefits. Such investigation could provide tangible benefits, since all organisms are probably involved in a mutualism at some point in their lives (Bronstein 1994), and mutualisms are essential for many ecosystem services, such as crop pollination.

I examined patterns of selection and genome evolution in a model organism for understanding adaptation and speciation in **Chapter 6**, where I performed analyses on a Darwin's finch genome from *Geospiza magnirostris* (the large ground finch). I was involved in many of the genome analyses, including the genome quality assessments, transposable element prediction, G+C content analyses, gene prediction, and ortholog/paralogue assignment. However, I focus on the results from the evolutionary rate analyses. I infer that purifying selection acted equally or more efficiently over protein sequences in the Darwin's finches compared to those of zebra finch, an unexpected result given that Darwin's finches have a very small effective population size. Applying a branch-site test (Yang et al. 2005; Zhang et al. 2005) I find evidence of positive selection at specific amino acid residues for 21 genes in the Darwin's finch lineage for 21, and two of these genes, *POU1F1* and *IGF2R*, may be involved in beak development. I also identify 47 positively selected genes across the ancestral passerine lineage and infer through gene ontology enrichment and protein sequence analyses that cilia-related functions are overrepresented among these genes; and I suggest that these genes may be evolving adaptively in response to reproductive pressures.

This switch to examining patterns of positive selection, rather than negative selection, is not simply examining the opposite side of neutral evolution. In fact, purifying and positive selection are inextricably linked due to the dynamic turnover of functional sequence. I precisely defined turnover as the gain or loss of function at a particular locus of the genome, as changes in the physical and genetic environment, and mutations in the sequence at the locus itself, cause the locus to switch from being functional to being non-functional or vice versa. I defined positive selection as the propagation of advantageous mutations through a population due to natural selection. When turnover occurs due to mutations at the locus, these gain-of-function mutations are positively selected. Following the fixation of these positively selected mutations, the locus will become subject to purifying selection to maintain the newly acquired function.

So, rapid turnover does tend to imply high levels of positive selection. However, generally studies focussing on the turnover of functional sequence have not made reference to positive selection and vice-versa (Nielsen et al. 2007; Ponting and Hardison 2011; Ponting et al. 2011; Vitti et al. 2013), so these two strands of research are proceeding in parallel despite the fact that they are clearly intertwined. I can think of two reasons why this may be the case. First, historically positive selection analyses tend to have focussed on coding regions for practical ease, often using dN/dS or McDonald-Kreitman style tests (Eyre-Walker 2006), while turnover studies have tended to focus on noncoding sequences, particularly TFBSs (Dermitzakis and Clark 2002; Moses et al. 2006). Since positive selection in protein coding sequences is often expected to entail altering the function of already functional sequence, it would not classify as turnover by my definition since the sequence is not passing through a non-functional state. Second, positive selection is tied to the concept of adaptive evolution, because positive selection implies beneficial changes that lead to novel traits and lineage-specific biology. By contrast, turnover of functional sequence is more often thought of as

occurring by a model of compensatory evolution, where a degree of functional redundancy in biological systems allows the turnover of functional sequence to simply rewire functional networks without necessarily creating functional novelty in any meaningful phenotypic sense. The key question is whether the lineage-specific functional sequence that both positive selection and turnover predicts has either equivalent or disparate functions in the different lineages. Answering this question will facilitate the fusion of these two currently overly-independent concepts.

References

Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, Kang HM, Marth GT, McVean GA. 2012. An integrated map of genetic variation from 1,092 human genomes. *Nature* 491(7422): 56-65.

Abzhanov A. 2010. Darwin's Galapagos finches in modern biology. *Philos Trans R Soc Lond B Biol Sci* 365(1543): 1001-1007.

Abzhanov A, Kuo WP, Hartmann C, Grant BR, Grant PR, Tabin CJ. 2006. The calmodulin pathway and evolution of elongated beak morphology in Darwin's finches. *Nature* 442(7102): 563-567.

Abzhanov A, Protas M, Grant BR, Grant PR, Tabin CJ. 2004. Bmp4 and morphological variation of beaks in Darwin's finches. *Science* 305(5689): 1462-1465.

Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR. 2010. A method and server for predicting damaging missense mutations. *Nature methods* 7(4): 248-249.

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

Allison AC. 1956. The sickle-cell and haemoglobin C genes in some African populations. *Annals of human genetics* 21(1): 67-89.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J Mol Biol* 215(3): 403-410.

Ames RM, Lovell SC. 2011. Diversification at transcription factor binding sites within a species and the implications for environmental adaptation. *Mol Biol Evol* 28(12): 3331-3344.

Arbiza L, Gronau I, Aksoy BA, Hubisz MJ, Gulko B, Keinan A, Siepel A. 2013. Genome-wide inference of natural selection on human transcription factor binding sites. *Nat Genet* 45(7): 723-729.

Arnason U, Adegoke JA, Gullberg A, Harley EH, Janke A, Kullberg M. 2008. Mitogenomic relationships of placental mammals and molecular estimates of their divergences. *Gene* 421(1-2): 37-51.

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT et al. 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* 25(1): 25-29.

Asthana S, Roytberg M, Stamatoyannopoulos J, Sunyaev S. 2007. Analysis of sequence conservation at nucleotide resolution. *PLoS Comput Biol* 3(12): e254.

Axelsson E, Webster MT, Ratnakumar A, Ponting CP, Lindblad-Toh K. 2012. Death of PRDM9 coincides with stabilization of the recombination landscape in the dog genome. *Genome Res* 22(1): 51-63.

Axelsson E, Hultin-Rosenberg L, Brandström M, Zwahlén M, Clayton D, Ellegren H. Natural selection in avian protein-coding genes expressed in brain. 2008. *Mol Ecol* 17: 3008-3017

Axelsson E, Webster MT, Smith NG, Burt DW, Ellegren H. 2005. Comparison of the chicken and turkey genomes reveals a higher rate of nucleotide divergence on microchromosomes than macrochromosomes. *Genome Res* 15(1): 120-125.

Bailey TL, Elkan C. 1994. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proceedings / International Conference on Intelligent Systems for Molecular Biology* 2: 28-36.

Bakewell MA, Shi P, Zhang J. 2007. More genes underwent positive selection in chimpanzee evolution than in human evolution. *Proc Natl Acad Sci U S A* 104(18): 7489-7494.

Balakrishnan CN, Edwards SV. 2009. Nucleotide variation, linkage disequilibrium and founder-facilitated speciation in wild populations of the zebra finch (Taeniopygia guttata). *Genetics* 181(2): 645-660.

Balmer JE, Blomhoff R. 2009. Evolution of transcription factor binding sites in mammalian gene regulatory regions: handling counterintuitive results. *J Mol Evol* 68(6): 654-664.

Barton NH. 2007. *Evolution*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y.

Bersaglieri T, Sabeti PC, Patterson N, Vanderploeg T, Schaffner SF, Drake JA, Rhodes M, Reich DE, Hirschhorn JN. 2004. Genetic signatures of strong recent positive selection at the lactase gene. *Am J Hum Genet* 74(6): 1111-1120.

Birkhead TR, Pellatt EJ, Fletcher F. 1993. Selection and utilization of spermatozoa in the reproductive tract of the female zebra finch Taeniopygia guttata. *J Reprod Fertil* 99(2): 593-600.

Blanchette M, Kent WJ, Riemer C, Elnitski L, Smit AF, Roskin KM, Baertsch R, Rosenbloom K, Clawson H, Green ED et al. 2004. Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res* 14(4): 708-715.

Blow MJ, McCulley DJ, Li Z, Zhang T, Akiyama JA, Holt A, Plajzer-Frick I, Shoukry M, Wright C, Chen F et al. 2010. ChIP-Seq identification of weakly conserved heart enhancers. *Nat Genet* 42(9): 806-810.

Bonferroni CE. 1936. *Teoria statistica delle classi e calcolo delle probabilità*. Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze.

Botstein D, Chervitz SA, Cherry JM. 1997. Yeast as a model organism. *Science* 277(5330): 1259-1260.

Brandström M, Ellegren H. 2007. The genomic landscape of short insertion and deletion polymorphisms in the chicken (Gallus gallus) Genome: a high frequency of deletions in tandem duplicates. *Genetics* 176(3): 1691-1701.

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

Bray N, Pachter L. 2004. MAVID: constrained ancestral alignment of multiple sequences. *Genome Res* 14(4): 693-699.

Bronstein JL. 1994. Our Current Understanding of Mutualism. *Q Rev Biol* 69(1): 31-51.

Brugmann SA, Powder KE, Young NM, Goodnough LH, Hahn SM, James AW, Helms JA, Lovett M. 2010. Comparative gene expression analysis of avian embryonic facial structures reveals new candidates for human craniofacial disorders. *Hum Mol Genet* 19(5): 920-930.

Burns KJ, Hackett SJ, Klein NK. 2002. Phylogenetic relationships and morphological diversity in Darwin's finches and their relatives. *Evolution* 56(6): 1240-1252.

Burt DW. 2002. Origin and evolution of avian microchromosomes. *Cytogenet Genome Res* 96(1-4): 97-112.

Carvalho AB, Dobo BA, Vibranovski MD, Clark AG. 2001. Identification of five new genes on the Y chromosome of Drosophila melanogaster. *Proc Natl Acad Sci U S A* 98(23): 13225-13230.

Chaix R, Somel M, Kreil DP, Khaitovich P, Lunter GA. 2008. Evolution of primate gene expression: drift and corrective sweeps? *Genetics* 180(3): 1379-1389.

Charlesworth B, Morgan MT, Charlesworth D. 1993. The effect of deleterious mutations on neutral molecular variation. *Genetics* 134(4): 1289-1303.

Chiaromonte F, Weber RJ, Roskin KM, Diekhans M, Kent WJ, Haussler D. 2003. The share of human genomic DNA under selection estimated from human-mouse genomic alignments. *Cold Spring Harbor symposia on quantitative biology* 68: 245-254.

Chiaromonte F, Yap VB, Miller W. 2002. Scoring pairwise genomic sequence alignments. *Pacific Symposium on Biocomputing Pacific Symposium on Biocomputing*: 115-126.

Comte de Buffon. 1780. *Les époques de la nature*. De l'Imprimerie royale.

Cooper GM, Stone EA, Asimenos G, Green ED, Batzoglou S, Sidow A. 2005. Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res* 15(7): 901-913.

Cracraft J, Barker FK. 2009. *The Timetree of Life*. Oxford University Press.

Creevey CJ, Muller J, Doerks T, Thompson JD, Arendt D, Bork P. 2011. Identifying single copy orthologs in Metazoa. *PLoS Comput Biol* 7(12): e1002269.

Curwen V, Eyras E, Andrews TD, Clarke L, Mongin E, Searle SM, Clamp M. 2004. The Ensembl automatic gene annotation system. *Genome Res* 14(5): 942-950.

D'Onofrio G, Ghosh TC, Bernardi G. 2002. The base composition of the genes is correlated with the secondary structures of the encoded proteins. *Gene* 300(1-2): 179-187.

Dalloul RA, Long JA, Zimin AV, Aslam L, Beal K, Blomberg Le A, Bouffard P, Burt DW, Crasta O, Crooijmans RP et al. 2010. Multi-platform next-generation sequencing of the domestic turkey (Meleagris gallopavo): genome assembly and analysis. *PLoS Biol* 8(9).

Darwin C. 1859. *The origin of species by means of natural selection: or, the preservation of favored races in the struggle for life*. AL Burt.

Darwin C, Wallace A. 1858. On the tendency of species to form varieties; and on the perpetuation of varieties and species by natural means of selection. *Journal of the Proceedings of the Linnean Society of London Zoology* 3(9): 45-62.

Davydov EV, Goode DL, Sirota M, Cooper GM, Sidow A, Batzoglou S. 2010. Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS Comput Biol* 6(12): e1001025.

Dermitzakis ET, Clark AG. 2002. Evolution of transcription factor binding sites in Mammalian gene regulatory regions: conservation and turnover. *Mol Biol Evol* 19(7): 1114-1121.

Derrien T, Vaysse A, Andre C, Hitte C. 2012. Annotation of the domestic dog genome sequence: finding the missing genes. *Mammalian Genome : Oficial Journal of the International Mammalian Genome Society* 23(1-2): 124-131.

Dobzhansky T. 1973. Nothing in biology makes sense except in the light of evolution. *The American Biology Teacher* 35(3): 125-129.

Domene S, Bumaschny VF, de Souza FS, Franchini LF, Nasif S, Low MJ, Rubinstein M. 2013. Enhancer turnover and conserved regulatory function in vertebrate evolution. *Philos Trans R Soc Lond B Biol Sci* 368(1632): 20130027.

Doniger SW, Fay JC. 2007. Frequent gain and loss of functional transcription factor binding sites. *PLoS Comput Biol* 3(5): e99.

Doolittle WF. 2013. Is junk DNA bunk? A critique of ENCODE. *Proc Natl Acad Sci U S A*.

Dubcovsky J, Dvorak J. 2007. Genome plasticity a key factor in the success of polyploid wheat under domestication. *Science* 316(5833): 1862-1866.

Dunham I, Kundaje A, Aldred SF, Collins PJ, Davis CA, Doyle F, Epstein CB, Frietze S, Harrow J, Kaul R, et al. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* 489(7414): 57-74.

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

Ecker JR, Bickmore WA, Barroso I, Pritchard JK, Gilad Y, Segal E. 2012. Genomics: ENCODE explained. *Nature* 489(7414): 52-55.
Eddy SR. 1996. Hidden Markov models. *Curr Opin Struct Biol* 6(3): 361-365.
Eddy SR. 2009. A new generation of homology search tools based on probabilistic inference. *Genome Inform* 23(1): 205-211.
Eddy SR. 2012. The C-value paradox, junk DNA and ENCODE. *Current Biology :* 22(21): R898-899.
Eddy SR. 2013. The ENCODE project: missteps overshadowing a success. *Current Biology :* 23(7): R259-261.
Edgar RC. 2004. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* 5: 113.
Ehrlich M, Wang RY. 1981. 5-Methylcytosine in eukaryotic DNA. *Science* 212(4501): 1350-1357.
Ellegren H. 2009. A selection model of molecular evolution incorporating the effective population size. *Evolution* 63(2): 301-305.
Emery NJ. 2006. Cognitive ornithology: the evolution of avian intelligence. *Philos Trans R Soc Lond B Biol Sci* 361(1465): 23-43.
Eory L, Halligan DL, Keightley PD. 2010. Distributions of selectively constrained sites and deleterious mutation rates in the hominid and murid genomes. *Mol Biol Evol* 27(1): 177-192.
Eyre-Walker A. 2006. The genomic rate of adaptive evolution. *Trends in Ecology & Evolution* 21(10): 569-575.
Eyre-Walker A, Woolfit M, Phelps T. 2006. The distribution of fitness effects of new deleterious amino acid mutations in humans. *Genetics* 173(2): 891-900.
Felsenstein J, Churchill GA. 1996. A Hidden Markov Model approach to variation among sites in rate of evolution. *Mol Biol Evol* 13(1): 93-104.
Fisher R. 1930. *The Genetical Theory of Natural Selection*. Clarendon Press, Clarendon Press, Oxford.
Fletcher W, Yang Z. 2009. INDELible: a flexible simulator of biological sequence evolution. *Mol Biol Evol* 26(8): 1879-1888.
Fletcher W, Yang Z. 2010. The effect of insertions, deletions, and alignment errors on the branch-site test of positive selection. *Mol Biol Evol* 27(10): 2257-2267.
Freeman S, Herron J. 2003. *Evolutionary Analysis*. CramOutline&Highlight101.
Fujita MK, Edwards SV, Ponting CP. 2011. The Anolis lizard genome: an amniote genome without isochores. *Genome Biol Evol* 3: 974-984.
Fullerton SM, Bernardo Carvalho A, Clark AG. 2001. Local rates of recombination are positively correlated with GC content in the human genome. *Mol Biol Evol* 18(6): 1139-1142.
Futuyma DJ. 2009. *Evolution*. Sinauer Associates, Sunderland, Mass.
Gandon S, Buckling A, Decaestecker E, Day T. 2008. Host-parasite coevolution and patterns of adaptation across time and space. *Journal of Evolutionary Biology* 21(6): 1861-1866.
Garber M, Guttman M, Clamp M, Zody MC, Friedman N, Xie X. 2009. Identifying novel constrained elements by exploiting biased substitution patterns. *Bioinformatics* 25(12): i54-62.
Gavrilets S. 2000. Rapid evolution of reproductive barriers driven by sexual conflict. *Nature* 403(6772): 886-889.
Gerstein MB, Bruce C, Rozowsky JS, Zheng D, Du J, Korbel JO, Emanuelsson O, Zhang ZD, Weissman S, Snyder M. 2007. What is a gene, post-ENCODE? History and updated definition. *Genome Res* 17(6): 669-681.

Goodstadt L. 2010. Ruffus: a lightweight Python library for computational pipelines. *Bioinformatics* 26(21): 2778-2779.
Goodstadt L, Ponting CP. 2006. Phylogenetic reconstruction of orthology, paralogy, and conserved synteny for dog and human. *PLoS Comput Biol* 2(9): e133.
Gould SJ. 2002. *The Structure of Evolutionary Theory*. Harvard University Press, Harvard University.
Govoni KE, Lee SK, Chadwick RB, Yu H, Kasukawa Y, Baylink DJ, Mohan S. 2006. Whole genome microarray analysis of growth hormone-induced gene expression in bone: T-box3, a novel transcription factor, regulates osteoblast proliferation. *Am J Physiol Endocrinol Metab* 291(1): E128-136.
Grant BR, Grant PR. 1989. *Evolutionary dynamics of a natural population : the large cactus finch of the Galápagos*. University of Chicago Press, Chicago.
Grant PR, Grant BR. 1992. Hybridization of bird species. *Science* 256(5054): 193-197.
Grant PR, Grant BR. 2008. *How and Why Species Multiply: The Radiation of Darwin's finches*. Princeton University Press.
Grant PR, Grant BR. 2010. Conspecific versus heterospecific gene exchange between populations of Darwin's finches. *Philos Trans R Soc Lond B Biol Sci* 365(1543): 1065-1076.
Graur D, Zheng Y, Price N, Azevedo RB, Zufall RA, Elhaik E. 2013. On the immortality of television sets: "function" in the human genome according to the evolution-free gospel of ENCODE. *Genome Biol Evol* 5(3): 578-590.
Green P, Ewing B. 2013. Comment on "Evidence of abundant purifying selection in humans for recently acquired regulatory functions". *Science* 340(6133): 682; discussion 682.
Gregory TR, Nicol JA, Tamm H, Kullman B, Kullman K, Leitch IJ, Murray BG, Kapraun DF, Greilhuber J, Bennett MD. 2007. Eukaryotic genome size databases. *Nucleic Acids Res* 35(Database issue): D332-338.
Hangauer MJ, Vaughn IW, McManus MT. 2013. Pervasive Transcription of the Human Genome Produces Thousands of Previously Unidentified Long Intergenic Noncoding RNAs. *PLoS genetics* 9(6): e1003569.
Hardison RC, Roskin KM, Yang S, Diekhans M, Kent WJ, Weber R, Elnitski L, Li J, O'Connor M, Kolbe D et al. 2003. Covariation in frequencies of substitution, deletion, transposition, and recombination during eutherian evolution. *Genome Res* 13(1): 13-26.
Harmston N, Baresic A, Lenhard B. 2013. The mystery of extreme non-coding conservation. *Philos Trans R Soc Lond B Biol Sci* 368(1632): 20130021.
Hasegawa M, Kishino H, Yano T. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J Mol Evol* 22(2): 160-174.
Heger A, Ponting CP. 2007. Evolutionary rate analyses of orthologs and paralogs from 12 Drosophila genomes. *Genome Res* 17(12): 1837-1849.
Heger A, Ponting CP. 2008. OPTIC: orthologous and paralogous transcripts in clades. *Nucleic Acids Res* 36: D267-270.
Hill WG, Robertson A. 1966. The effect of linkage on limits to artificial selection. *Genetical Research* 8(3): 269-294.
Hsieh CC, DeFord JH, Flurkey K, Harrison DE, Papaconstantinou J. 2002. Effects of the Pit1 mutation on the insulin signaling pathway: implications on the longevity of the long-lived Snell dwarf mouse. *Mech Ageing Dev* 123(9): 1245-1255.
Hupalo D, Kern AD. 2013. Conservation and functional element discovery in 20 angiosperm plant genomes. *Mol Biol Evol* 30(7): 1729-1744.
Hyde BB, Liesa M, Elorza AA, Qiu W, Haigh SE, Richey L, Mikkola HK, Schlaeger TM, Shirihai OS. 2012. The mitochondrial transporter ABC-me (ABCB10), a downstream

target of GATA-1, is essential for erythropoiesis in vivo. *Cell Death Differ* 19(7): 1117-1126.
Ingraham HA, Chen RP, Mangalam HJ, Elsholtz HP, Flynn SE, Lin CR, Simmons DM, Swanson L, Rosenfeld MG. 1988. A tissue-specific transcription factor containing a homeodomain specifies a pituitary phenotype. *Cell* 55(3): 519-529.
International Chicken Genome Sequencing Consortium. 2004. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature* 432(7018): 695-716.
International Human Genome Sequencing Consortium. 2004. Finishing the euchromatic sequence of the human genome. *Nature* 431(7011): 931-945.
Jukes TH, Cantor CR. 1969. Evolution of protein molecules. In Munro (ed.). *Mammlian Protein Metabolism*. Academic Press, New York: pp. 21-132
Kaletta T, Hengartner MO. 2006. Finding function in novel targets: C. elegans as a model organism. *Nature Reviews Drug Discovery* 5(5): 387-398.
Kapushesky M, Adamusiak T, Burdett T, Culhane A, Farne A, Filippov A, Holloway E, Klebanov A, Kryvych N, Kurbatova N et al. 2012. Gene Expression Atlas update--a value-added database of microarray and sequencing-based functional genomics experiments. *Nucleic Acids Res* 40(Database issue): D1077-1081.
Keightley PD, Eyre-Walker A. 2000. Deleterious mutations and the evolution of sex. *Science* 290(5490): 331-333.
Kent WJ, Baertsch R, Hinrichs A, Miller W, Haussler D. 2003. Evolution's cauldron: duplication, deletion, and rearrangement in the mouse and human genomes. *Proc Natl Acad Sci U S A* 100(20): 11484-11489.
Kimura M. 1968. Evolutionary rate at the molecular level. *Nature* 217(5129): 624-626.
Kimura M. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J Mol Evol* 16(2): 111-120.
Kimura M. 1983. *The Neutral Theory of Molecular Evolution*. Cambridge University Press, Cambridge University Press, New York.
Kocher TD. 2004. Adaptive evolution and explosive speciation: the cichlid fish model. *Nat Rev Genet* 5(4): 288-298.
Kondrashov AS. 1995. Contamination of the genome by very slightly deleterious mutations: why have we not died 100 times over? *Journal of Theoretical Biology* 175(4): 583-594.
Koonin EV. 2005. Orthologs, paralogs, and evolutionary genomics. *Annual review of Genetics* 39: 309-338.
Koonin EV. 2011. Are there laws of genome evolution? *PLoS Comput Biol* 7(8): e1002173.
Kumar S, Subramanian S. 2002. Mutation rates in mammalian genomes. *Proc Natl Acad Sci U S A* 99(2): 803-808.
Künstner A, Nabholz B, Ellegren H. 2011a. Evolutionary constraint in flanking regions of avian genes. *Mol Biol Evol* 28(9): 2481-2489.
Künstner A, Nabholz B, Ellegren H. 2011b. Significant selective constraint at 4-fold degenerate sites in the avian genome and its consequence for detection of positive selection. *Genome Biol Evol* 3: 1381-1389.
Kutter C, Watt S, Stefflova K, Wilson MD, Goncalves A, Ponting CP, Odom DT, Marques AC. 2012. Rapid turnover of long noncoding RNAs and the evolution of gene expression. *PLoS Genetics* 8(7): e1002841.
Kvikstad EM, Duret L. 2014. Strong heterogeneity in mutation rate causes misleading hallmarks of natural selection on indel mutations in the human genome. *Mol Biol Evol* 31(1): 23-36.
Lamarck J. 1809. *Zoological philosophy*. Dentu, Paris

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

Li WH, Ellsworth DL, Krushkal J, Chang BH, Hewett-Emmett D. 1996. Rates of nucleotide substitution in primates and rodents and the generation-time effect hypothesis. *Mol Phylogenet Evol* 5(1): 182-187.

Lin MF, Kheradpour P, Washietl S, Parker BJ, Pedersen JS, Kellis M. 2011. Locating protein-coding sequences under selection for additional, overlapping functions in 29 mammalian genomes. *Genome Res* 21(11): 1916-1928.

Lindblad-Toh K, Garber M, Zuk O, Lin MF, Parker BJ, Washietl S, Kheradpour P, Ernst J, Jordan G, Mauceli E et al. 2011. A high-resolution map of human evolutionary constraint using 29 mammals. *Nature* 478(7370): 476-482.

Lindblad-Toh K, Wade CM, Mikkelsen TS, Karlsson EK, Jaffe DB, Kamal M, Clamp M, Chang JL, Kulbokas EJ, 3rd Zody MC et al. 2005. Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature* 438(7069): 803-819.

Loges NT, Olbrich H, Fenske L, Mussaffi H, Horvath J, Fliegauf M, Kuhl H, Baktai G, Peterffy E, Chodhari R et al. 2008. DNAI2 mutations cause primary ciliary dyskinesia with defects in the outer dynein arm. *Am J Hum Genet* 83(5): 547-558.

Lowe CB, Kellis M, Siepel A, Raney BJ, Clamp M, Salama SR, Kingsley DM, Lindblad-Toh K, Haussler D. 2011. Three periods of regulatory innovation during vertebrate evolution. *Science* 333(6045): 1019-1024.

Lu J, Zhang F, Xu S, Fire AZ, Kay MA. 2012. The extragenic spacer length between the 5' and 3' ends of the transgene expression cassette affects transgene silencing from plasmid-based vectors. *Molecular therapy : the Journal of the American Society of Gene Therapy* 20(11): 2111-2119.

Ludwig MZ, Bergman C, Patel NH, Kreitman M. 2000. Evidence for stabilizing selection in a eukaryotic enhancer element. *Nature* 403(6769): 564-567.

Lunter G. 2007. Probabilistic whole-genome alignments reveal high indel rates in the human and mouse genomes. *Bioinformatics* 23(13): i289-296.

Lunter G, Ponting CP, Hein J. 2006. Genome-wide identification of human functional DNA using a neutral indel model. *PLoS Comput Biol* 2(1): e5.

Lunter G, Rocco A, Mimouni N, Heger A, Caldeira A, Hein J. 2008. Uncertainty in homology inferences: assessing and improving genomic sequence alignment. *Genome Res* 18(2): 298-309.

Ma W, Dong FF, Stavrinides J, Guttman DS. 2006. Type III effector diversification via both pathoadaptation and horizontal transfer in response to a coevolutionary arms race. *PLoS Genetics* 2(12): e209.

Makova KD, Yang S, Chiaromonte F. 2004. Insertions and deletions are male biased too: a whole-genome analysis in rodents. *Genome Res* 14(4): 567-573.

Mallarino R, Grant PR, Grant BR, Herrel A, Kuo WP, Abzhanov A. 2011. Two developmental modules establish 3D beak-shape variation in Darwin's finches. *Proc Natl Acad Sci U S A* 108(10): 4057-4062.

Mallick S, Gnerre S, Muller P, Reich D. 2009. The difficulty of avoiding false positives in genome scans for natural selection. *Genome Res* 19(5): 922-933.

Margulies EH, Birney E. 2008. Approaches to comparative sequence analysis: towards a functional view of vertebrate genomes. *Nat Rev Genet* 9(4): 303-313.

Margulies EH, Blanchette M, Haussler D, Green ED. 2003. Identification and characterization of multi-species conserved sequences. *Genome Res* 13(12): 2507-2518.

Margulies EH, Cooper GM, Asimenos G, Thomas DJ, Dewey CN, Siepel A, Birney E, Keefe D, Schwartz AS, Hou M et al. 2007. Analyses of deep mammalian sequence alignments and constraint predictions for 1% of the human genome. *Genome Res* 17(6): 760-774.

May D, Blow MJ, Kaplan T, McCulley DJ, Jensen BC, Akiyama JA, Holt A, Plajzer-Frick I, Shoukry M, Wright C et al. 2012. Large-scale discovery of enhancers from human heart tissue. *Nat Genet* 44(1): 89-93.

Mayr E. 1946. The Number of Species of Birds. *The Auk* 63(1): 67.

McLean CY, Reno PL, Pollen AA, Bassan AI, Capellini TD, Guenther C, Indjeian VB, Lim X, Menke DB, Schaar BT et al. 2011. Human-specific loss of regulatory DNA and the evolution of human-specific traits. *Nature* 471(7337): 216-219.

Meader S. 2010. Application of the Neutral Indel Model to Genome Sequences of Diverse Metazoans. In *Physiology, Anatomy, and Genetics*, Vol Dphil. University of Oxford, Oxford.

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

Mendel G. 1865. Experiments in Plant Hybridisation, trans. *Translated in Castle's Genetics and Eugenics, 1916* 6: 353.

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

Montgomery SB, Goode D, Kvikstad E, Albers CA, Zhang Z, Mu XJ, Ananda G, Howie B, Karczewski KJ, Smith KS et al. 2013. The origin, evolution and functional impact of short insertion-deletion variants identified in 179 human genomes. *Genome Res*.

Moses AM, Pollard DA, Nix DA, Iyer VN, Li XY, Biggin MD, Eisen MB. 2006. Large-scale turnover of functional transcription factor binding sites in Drosophila. *PLoS Comput Biol* 2(10): e130.

Mouchiroud D, D'Onofrio G, Aissani B, Macaya G, Gautier C, Bernardi G. 1991. The distribution of genes in the human genome. *Gene* 100: 181-187.

Mummery-Widmer JL, Yamazaki M, Stoeger T, Novatchkova M, Bhalerao S, Chen D, Dietzl G, Dickson BJ, Knoblich JA. 2009. Genome-wide analysis of Notch signalling in Drosophila by transgenic RNAi. *Nature* 458(7241): 987-992.

Mustonen V, Lassig M. 2005. Evolutionary population genetics of promoters: predicting binding sites and functional phylogenies. *Proc Natl Acad Sci U S A* 102(44): 15936-15941.

Nabholz B, Kunstner A, Wang R, Jarvis ED, Ellegren H. 2011. Dynamic evolution of base composition: causes and consequences in avian phylogenomics. *Mol Biol Evol* 28(8): 2197-2210.

Nachman MW, Crowell SL. 2000. Estimate of the mutation rate per nucleotide in humans. *Genetics* 156(1): 297-304.

Nakabachi A, Yamashita A, Toh H, Ishikawa H, Dunbar HE, Moran NA, Hattori M. 2006. The 160-kilobase genome of the bacterial endosymbiont Carsonella. *Science* 314(5797): 267.

Nam K, Mugal C, Nabholz B, Schielzeth H, Wolf JB, Backstrom N, Kunstner A, Balakrishnan CN, Heger A, Ponting CP et al. 2010. Molecular evolution of genes in avian genomes. *Genome Biol* 11(6): R68.

Necsulea A, Soumillon M, Warnefors M, Liechti A, Daish T, Zeller U, Baker JC, Grutzner F, Kaessmann H. 2014. The evolution of lncRNA repertoires and expression patterns in tetrapods. *Nature* 505(7485): 635-640.

Nesse RM, Williams GC. 1994. *Why we get sick : the new science of Darwinian medicine*. Times Books, New York.

Nielsen R, Bustamante C, Clark AG, Glanowski S, Sackton TB, Hubisz MJ, Fledel-Alon A, Tanenbaum DM, Civello D, White TJ et al. 2005. A scan for positively selected genes in the genomes of humans and chimpanzees. *PLoS Biol* 3(6): e170.

Nielsen R, Hellmann I, Hubisz M, Bustamante C, Clark AG. 2007. Recent and ongoing selection in the human genome. *Nat Rev Genet* 8(11): 857-868.

Notredame C, Higgins DG, Heringa J. 2000. T-Coffee: A novel method for fast and accurate multiple sequence alignment. *J Mol Biol* 302(1): 205-217.

Nozawa M, Suzuki Y, Nei M. 2009. Reliabilities of identifying positive selection by the branch-site and the site-prediction methods. *Proc Natl Acad Sci U S A* 106(16): 6700-6705.

Odom DT, Dowell RD, Jacobsen ES, Gordon W, Danford TW, MacIsaac KD, Rolfe PA, Conboy CM, Gifford DK, Fraenkel E. 2007. Tissue-specific transcriptional regulation has diverged significantly between human and mouse. *Nat Genet* 39(6): 730-732.

Ohno S. 1972. So much "junk" DNA in our genome. *Brookhaven Symposia in Biology* 23: 366-370.

Organ CL, Shedlock AM, Meade A, Pagel M, Edwards SV. 2007. Origin of avian genome size and structure in non-avian dinosaurs. *Nature* 446(7132): 180-184.

Pappas CT, Bliss KT, Zieseniss A, Gregorio CC. 2011. The Nebulin family: an actin support group. *Trends Cell Biol* 21(1): 29-37.

Parker SC, Hansen L, Abaan HO, Tullius TD, Margulies EH. 2009. Local DNA topography correlates with functional noncoding regions of the human genome. *Science* 324(5925): 389-392.

Parsch J. 2003. Selective constraints on intron evolution in Drosophila. *Genetics* 165(4): 1843-1851.

Paten B, Herrero J, Beal K, Fitzgerald S, Birney E. 2008. Enredo and Pecan: genome-wide mammalian consistency-based multiple alignment with paralogs. *Genome Res* 18(11): 1814-1828.

Pellicer J, Fay MF, Leitch IJ. 2010. The largest eukaryotic genome of them all? *Botanical Journal of the Linnean Society* 164(1): 10-15.

Penn O, Privman E, Landan G, Graur D, Pupko T. 2010. An alignment confidence score capturing robustness to guide tree uncertainty. *Mol Biol Evol* 27(8): 1759-1767.

Pennisi E. 2012. Genomics. ENCODE project writes eulogy for junk DNA. *Science* 337(6099): 1159, 1161.

Pereira SL, Baker AJ. 2006. A molecular timescale for galliform birds accounting for uncertainty in time estimates and heterogeneity of rates of DNA substitutions across lineages and sites. *Mol Phylogenet Evol* 38(2): 499-509.

Perrone CA, Tritschler D, Taulman P, Bower R, Yoder BK, Porter ME. 2003. A novel dynein light intermediate chain colocalizes with the retrograde motor for intraflagellar transport at sites of axoneme assembly in chlamydomonas and Mammalian cells. *Mol Biol Cell* 14(5): 2041-2056.

Petren K, Grant PR, Grant BR, Keller LF. 2005. Comparative landscape genetics and the adaptive radiation of Darwin's finches: the role of peripheral isolation. *Mol Ecol* 14(10): 2943-2957.

Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A. 2010. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res* 20(1): 110-121.

Pollard KS, Salama SR, King B, Kern AD, Dreszer T, Katzman S, Siepel A, Pedersen JS, Bejerano G, Baertsch R et al. 2006. Forces shaping the fastest evolving regions in the human genome. *PLoS Genetics* 2(10): e168.

Ponting CP, Hardison RC. 2011. What fraction of the human genome is functional? *Genome Res* 21(11): 1769-1776.

Ponting CP, Nellaker C, Meader S. 2011. Rapid turnover of functional sequence in human and other genomes. *Annu Rev Genomics Hum Genet* 12: 275-299.
Price AL, Jones NC, Pevzner PA. 2005. De novo identification of repeat families in large genomes. *Bioinformatics* 21 Suppl 1: i351-358.
Price TD, Grant PR, Gibbs HL, Boag PT. 1984. Recurrent patterns of natural selection in a population of Darwin's finches. *Nature* 309(5971): 787-789.
Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26(6): 841-842.
R Core Team. 2012. R: A Language and Environment for Statistical Computing.
Rambaut A. 2008. FigTree. http://tree.bio.ed.ac.uk/software/figtree/
Rambaut A, Grassly NC. 1997. Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Comput Appl Biosci* 13(3): 235-238.
Rands CM, Darling A, Fujita M, Kong L, Webster MT, Clabaut C, Emes RD, Heger A, Meader S, Hawkins MB et al. 2013. Insights into the evolution of Darwin's finches from comparative analysis of the Geospiza magnirostris genome sequence. *BMC Genomics* 14: 95.
Romiguier J, Ranwez V, Douzery EJ, Galtier N. 2010. Contrasting GC-content dynamics across 33 mammalian genomes: relationship with life-history traits and chromosome sizes. *Genome Res* 20(8): 1001-1009.
Sabeti PC Varilly P Fry B Lohmueller J Hostetter E Cotsapas C Xie X Byrne EH McCarroll SA Gaudet R et al. 2007. Genome-wide detection and characterization of positive selection in human populations. *Nature* 449(7164): 913-918.
Sanchez-Pulido L, Devos D, Genevrois S, Vicente M, Valencia A. 2003. POTRA: a conserved domain in the FtsQ family and a class of beta-barrel outer membrane proteins. *Trends in Biochemical Sciences* 28(10): 523-526.
Sanchez-Pulido L, Devos D, Valencia A. 2002. BRICHOS: a conserved domain in proteins associated with dementia, respiratory distress and cancer. *Trends in Biochemical Sciences* 27(7): 329-332.
Sanchez-Pulido L, Ponting CP. 2011. Structure and evolutionary history of DISC1. *Hum Mol Genet* 20(R2): R175-181.
Satija R, Hein J, Lunter GA. 2010. Genome-wide functional element detection using pairwise statistical alignment outperforms multiple genome footprinting techniques. *Bioinformatics* 26(17): 2116-2120.
Sato A, Tichy H, O'HUigin C, Grant PR, Grant BR, Klein J. 2001. On the origin of Darwin's finches. *Mol Biol Evol* 18(3): 299-311.
Schmidt D, Wilson MD, Ballester B, Schwalie PC, Brown GD, Marshall A, Kutter C, Watt S, Martinez-Jimenez CP, Mackay S et al. 2010. Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding. *Science* 328(5981): 1036-1040.
Schwartz S, Kent WJ, Smit A, Zhang Z, Baertsch R, Hardison RC, Haussler D, Miller W. 2003. Human-mouse alignments with BLASTZ. *Genome Res* 13(1): 103-107.
Sharp PM, Li WH. 1987. The rate of synonymous substitution in enterobacterial genes is inversely related to codon usage bias. *Mol Biol Evol* 4(3): 222-230.
Shendure J, Ji H. 2008. Next-generation DNA sequencing. *Nature biotechnology* 26(10): 1135-1145.
Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S et al. 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* 15(8): 1034-1050.

Sims D, Ilott NE, Sansom SN, Sudbery IM, Johnson JS, Fawcett KA, Berlanga-Taylor AJ, Luna-Valero S, Ponting CP, Heger A. 2014. CGAT: computational genomics analysis toolkit. *Bioinformatics*. doi: 10.1093/bioinformatics/btt756

Sinervo B, Bleay C, Adamopoulou C. 2001. Social causes of correlational selection and the resolution of a heritable throat color polymorphism in a lizard. *Evolution* 55(10): 2040-2052.

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

Smit AFA, Hubley R, Green P. 1996-2010. RepeatMasker Open-3.0.

Smith JM, Haigh J. 1974. The hitch-hiking effect of a favourable gene. *Genetical Research* 23(1): 23-35.

Smith NG, Brandström M, Ellegren H. 2004. Evidence for turnover of functional noncoding DNA in mammalian genome evolution. *Genomics* 84(5): 806-813.

Snell GD. 1929. Dwarf, a New Mendelian recessive character of the house mouse. *Proc Natl Acad Sci U S A* 15(9): 733-734.

Sonnhammer EL, Hollich V. 2005. Scoredist: a simple and robust protein sequence distance estimator. *BMC Bioinformatics* 6: 108.

Stark A, Lin MF, Kheradpour P, Pedersen JS, Parts L, Carlson JW, Crosby MA, Rasmussen MD, Roy S, Deoras AN et al. 2007. Discovery of functional elements in 12 Drosophila genomes using evolutionary signatures. *Nature* 450(7167): 219-232.

Strope CL, Scott SD, Moriyama EN. 2007. indel-Seq-Gen: a new protein family simulator incorporating domains, motifs, and indels. *Mol Biol Evol* 24(3): 640-649.

Suzuki Y. 2008. False-positive results obtained from the branch-site test of positive selection. *Genes Genet Syst* 83(4): 331-338.

Swanson WJ, Vacquier VD. 2002. The rapid evolution of reproductive proteins. *Nat Rev Genet* 3(2): 137-144.

Takahata N, Satta Y, Klein J. 1992. Polymorphism and balancing selection at major histocompatibility complex loci. *Genetics* 130(4): 925-938.

Talavera G, Castresana J. 2007. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Syst Biol* 56(4): 564-577.

Tam LW, Lefebvre PA. 2002. The Chlamydomonas MBO2 locus encodes a conserved coiled-coil protein important for flagellar waveform conversion. *Cell Motil Cytoskeleton* 51(4): 197-212.

Thomas CA, Jr. 1971. The genetic organization of chromosomes. *Annual Review of Genetics* 5: 237-256.

Thomas JW, Touchman JW, Blakesley RW, Bouffard GG, Beckstrom-Sternberg SM, Margulies EH, Blanchette M, Siepel AC, Thomas PJ, McDowell JC et al. 2003. Comparative analyses of multi-species sequences from targeted genomic regions. *Nature* 424(6950): 788-793.

Ulitsky I, Shkumatava A, Jan CH, Sive H, Bartel DP. 2011. Conserved function of lincRNAs in vertebrate embryonic development despite rapid sequence evolution. *Cell* 147(7): 1537-1550.

Ureta-Vidal A, Ettwiller L, Birney E. 2003. Comparative genomics: genome-wide analysis in metazoan eukaryotes. *Nat Rev Genet* 4(4): 251-262.

Venter JC Adams MD Myers EW Li PW Mural RJ Sutton GG Smith HO Yandell M Evans CA Holt RA et al. 2001. The sequence of the human genome. *Science* 291(5507): 1304-1351.

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

Vitti JJ, Grossman SR, Sabeti PC. 2013. Detecting natural selection in genomic data. *Annual review of genetics* 47: 97-120.

Voight BF, Kudaravalli S, Wen X, Pritchard JK. 2006. A map of recent positive selection in the human genome. *PLoS Biol* 4(3): e72.

Vorbach C, Harrison R, Capecchi MR. 2003. Xanthine oxidoreductase is central to the evolution and function of the innate immune system. *Trends Immunol* 24(9): 512-517.

Wang ZQ, Fung MR, Barlow DP, Wagner EF. 1994. Regulation of embryonic growth and lysosomal targeting by the imprinted Igf2/Mpr gene. *Nature* 372(6505): 464-467.

Ward LD, Kellis M. 2012. Evidence of abundant purifying selection in humans for recently acquired regulatory functions. *Science* 337(6102): 1675-1678.

Ward LD, Kellis M. 2013. Response to comment on "Evidence of abundant purifying selection in humans for recently acquired regulatory functions". *Science* 340(6133): 682.

Warren WC, Clayton DF, Ellegren H, Arnold AP, Hillier LW, Künstner A, Searle S, White S, Vilella AJ, Fairley S et al. 2010. The genome of a songbird. *Nature* 464(7289): 757-762.

Waterston RH Lindblad-Toh K Birney E Rogers J Abril JF Agarwal P Agarwala R Ainscough R Alexandersson M An P et al. 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature* 420(6915): 520-562.

Weatherly KL, Ramesh R, Strange H, Waite KL, Storrie B, Proudman JA, Wong EA. 2001. The turkey transcription factor Pit-1/GHF-1 can activate the turkey prolactin and growth hormone gene promoters in vitro but is not detectable in lactotrophs in vivo. *Gen Comp Endocrinol* 123(3): 244-253.

Wheeler TJ, Clements J, Eddy SR, Hubley R, Jones TA, Jurka J, Smit AF, Finn RD. 2013. Dfam: a database of repetitive DNA based on profile hidden Markov models. *Nucleic Acids Res* 41: D70-82.

White NE, Phillips MJ, Gilbert MT, Alfaro-Nunez A, Willerslev E, Mawson PR, Spencer PB, Bunce M. 2011. The evolutionary history of cockatoos (Aves: Psittaciformes: Cacatuidae). *Mol Phylogenet Evol* 59(3): 615-622.

Williamson SH, Hubisz MJ, Clark AG, Payseur BA, Bustamante CD, Nielsen R. 2007. Localizing recent adaptive evolution in the human genome. *PLoS Genetics* 3(6): e90.

Wilson DE, Reeder DM. 2005. *Mammal Species of the World. A Taxonomic and Geographic Reference*. Johns Hopkins University Press.

Wright S. 1931. Evolution in Mendelian Populations. *Genetics* 16(2): 97-159.

Wu CH, Apweiler R, Bairoch A, Natale DA, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R et al. 2006. The Universal Protein Resource (UniProt): an expanding universe of protein information. *Nucleic Acids Res* 34(Database issue): D187-191.

Yang Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput Appl Biosci* 13(5): 555-556.

Yang Z. 2006. *Computational Molecular Evolution*. Oxford University Press Oxford.

Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol* 24(8): 1586-1591.

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

Yang Z, Nielsen R, Goldman N. 2009. In defense of statistical methods for detecting positive selection. *Proc Natl Acad Sci U S A* 106(36): E95.

Yang Z, Wong WS, Nielsen R. 2005. Bayes empirical bayes inference of amino acid sites under positive selection. *Mol Biol Evol* 22(4): 1107-1118.

Zhang G, Parker P, Li B, Li H, Wang J. 2012. The genome of Darwin's Finch (Geospiza fortis). In *GigaScience*. doi:10.5524/100040

Zhang J, Nielsen R, Yang Z. 2005. Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. *Mol Biol Evol* 22(12): 2472-2479.

Zhang J, Webb DM, Podlaha O. 2002. Accelerated protein evolution and origins of human-specific features: Foxp2 as an example. *Genetics* 162(4): 1825-1835.

Zhang Z, Raghavachari B, Hardison RC, Miller W. 1994. Chaining multiple-alignment blocks. *Journal of Computational Biology* 1(3): 217-226.

Appendix

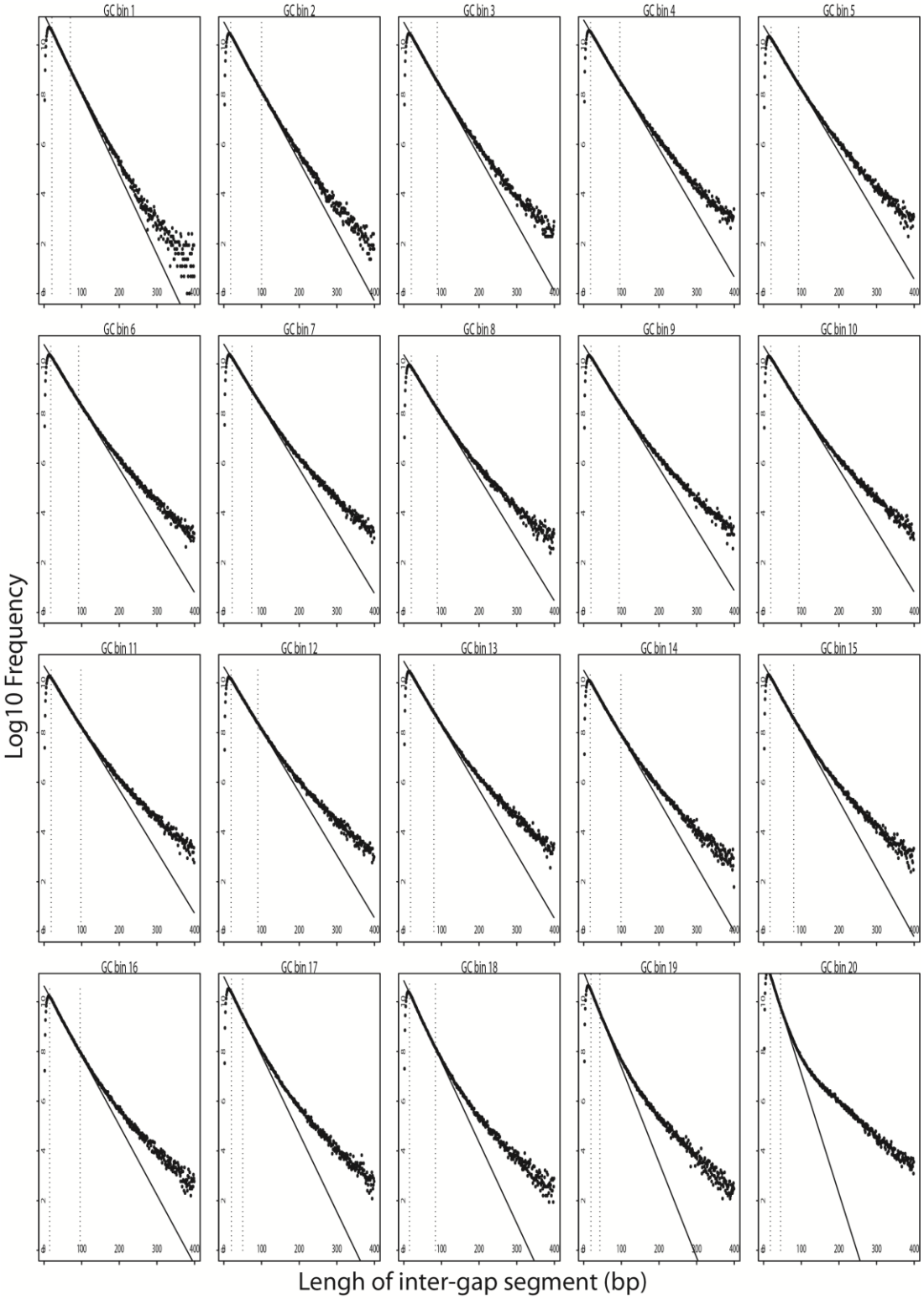

Figure A.1: IGS histograms across 20 equally populated GC-bins from trimmed human (hg19) – horse (equCab2) alignments.

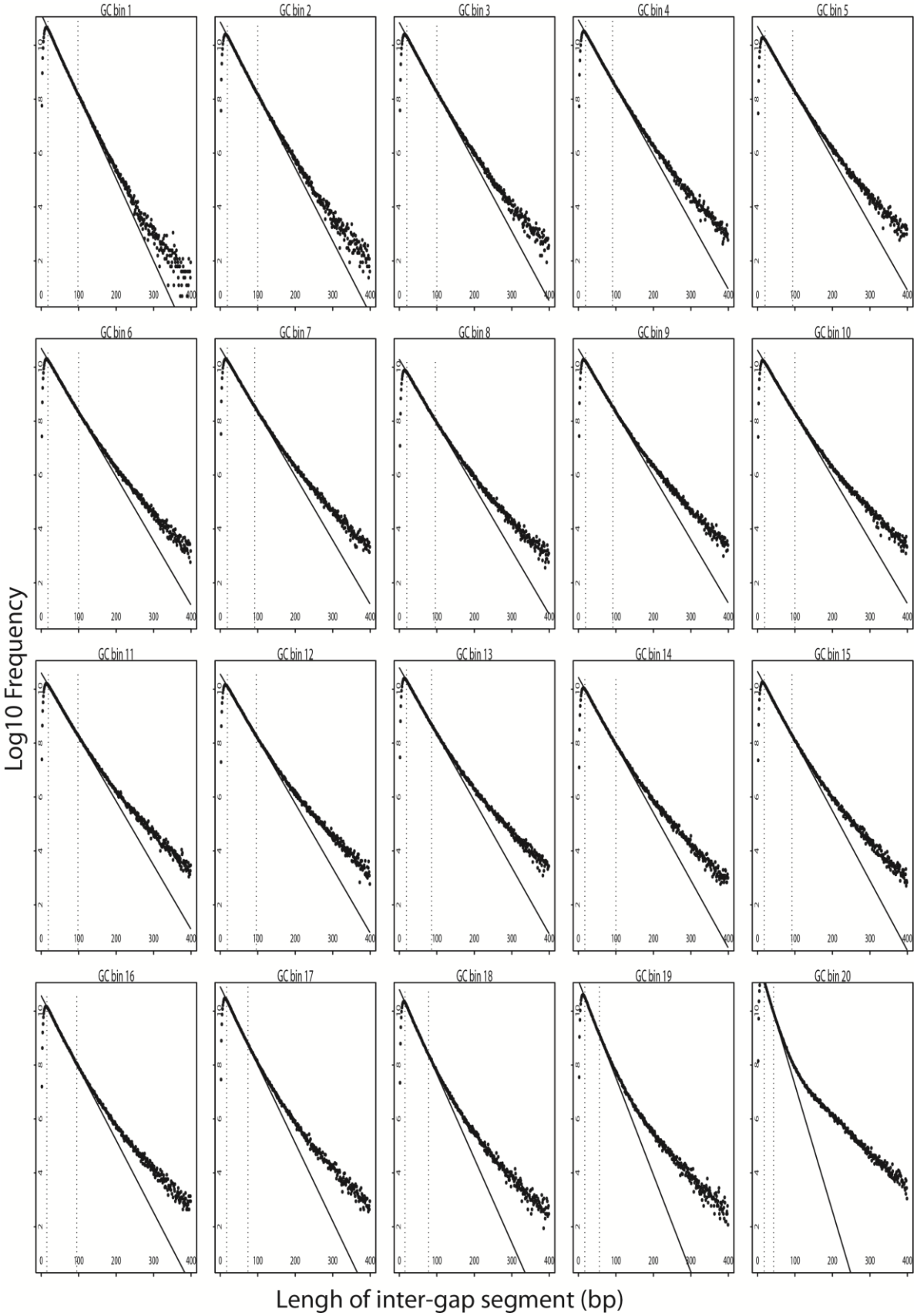

Figure A.2: IGS histograms across 20 equally populated GC-bins from trimmed human (hg19) – rhino (cerSim1) alignments.

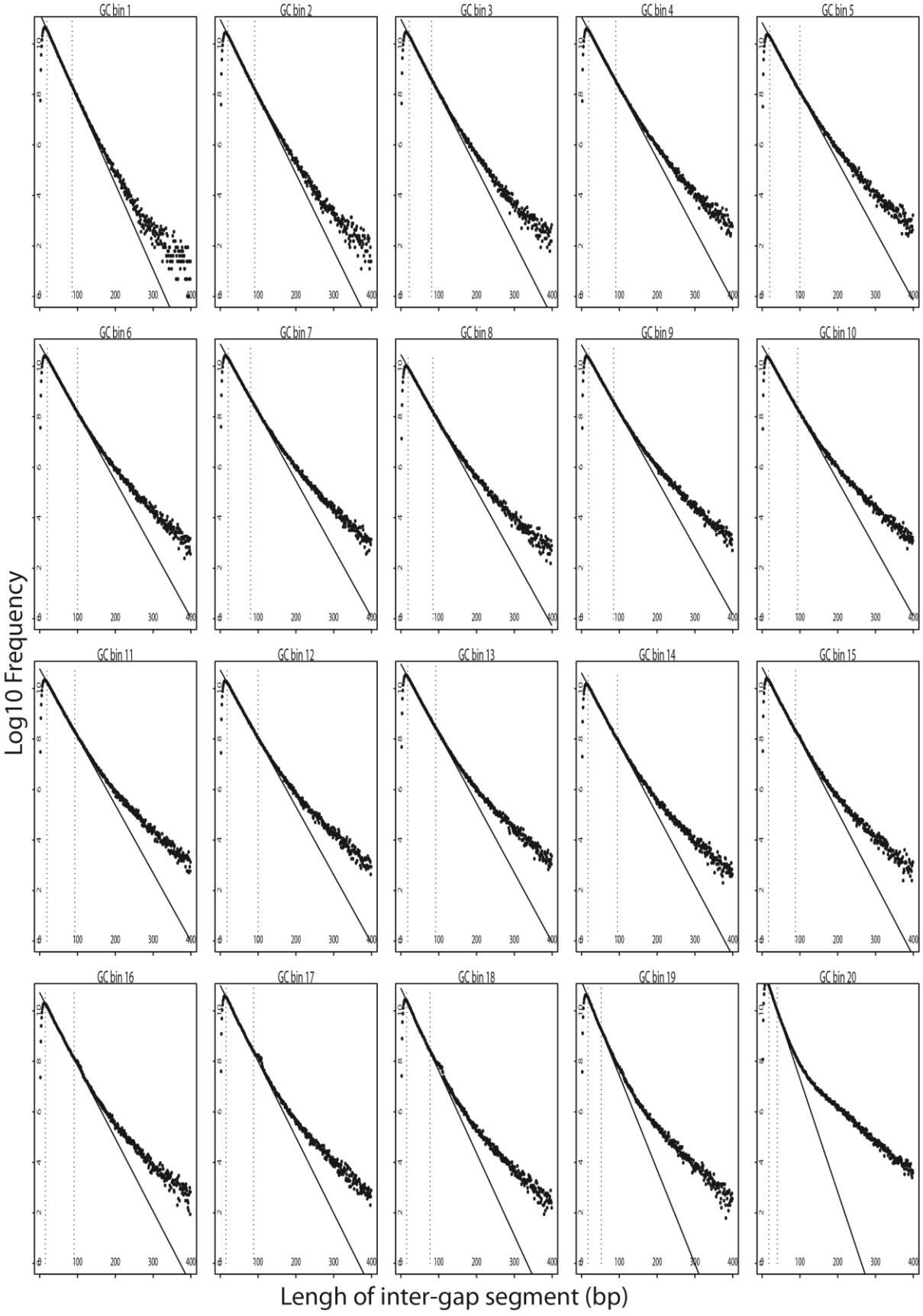

Figure A.3: IGS histograms across 20 equally populated GC-bins from trimmed human (hg19) – bushbaby (otoGar3) alignments.

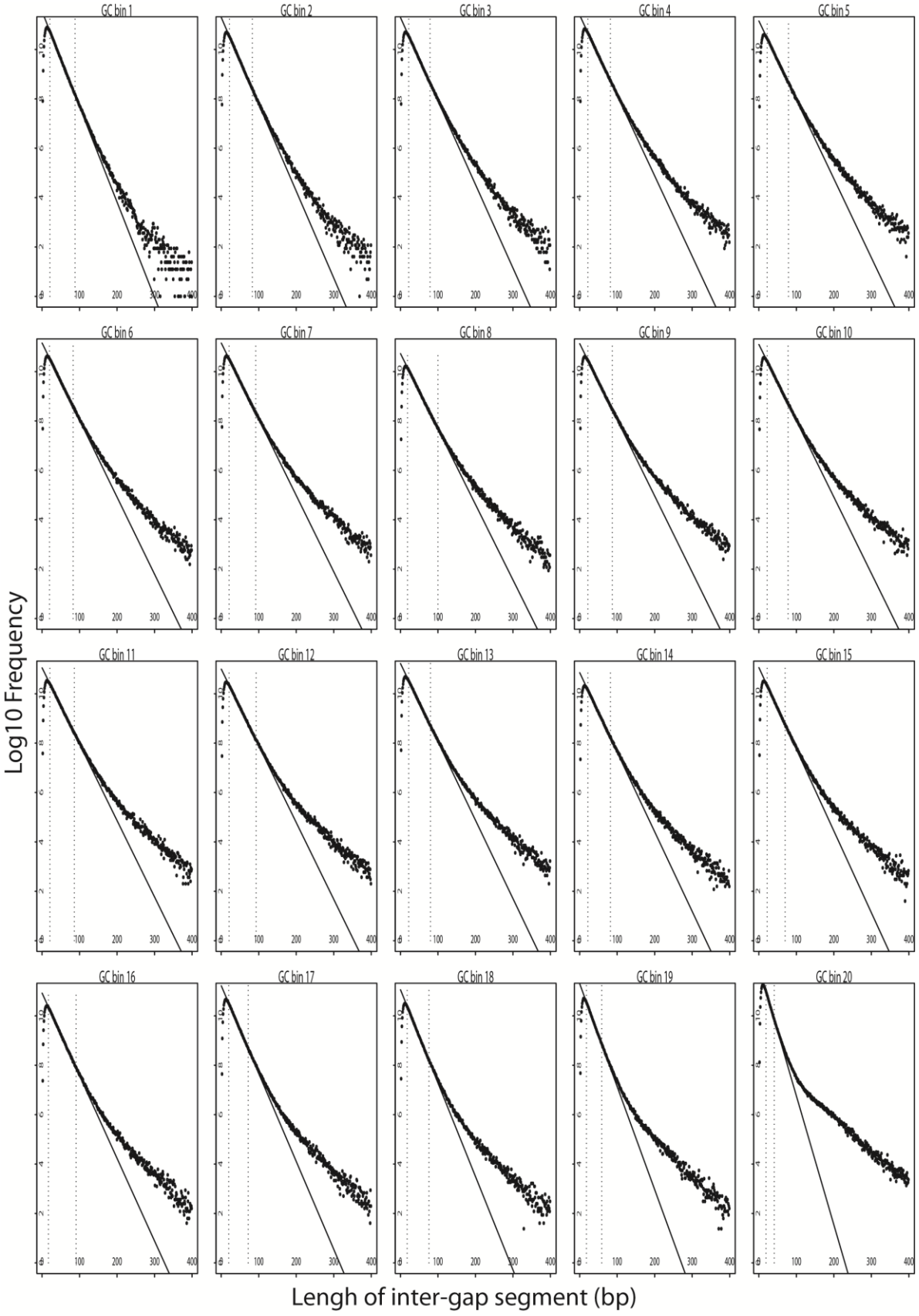

Figure A.4: IGS histograms across 20 equally populated GC-bins from trimmed human (hg19) – dog (canFam2) alignments.

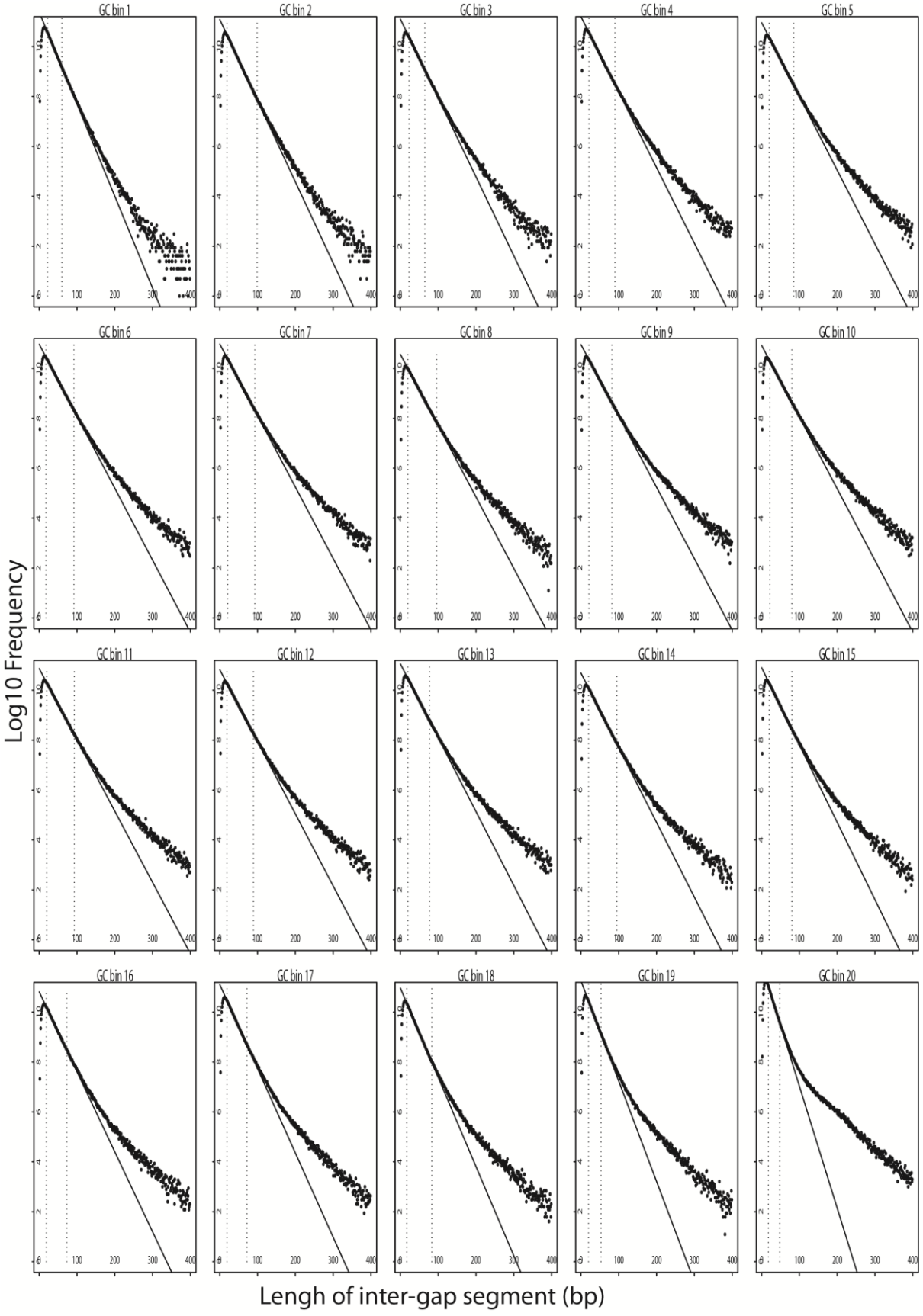

Figure A.5: IGS histograms across 20 equally populated GC-bins from trimmed human (hg19) – panda (ailMel1) alignments.

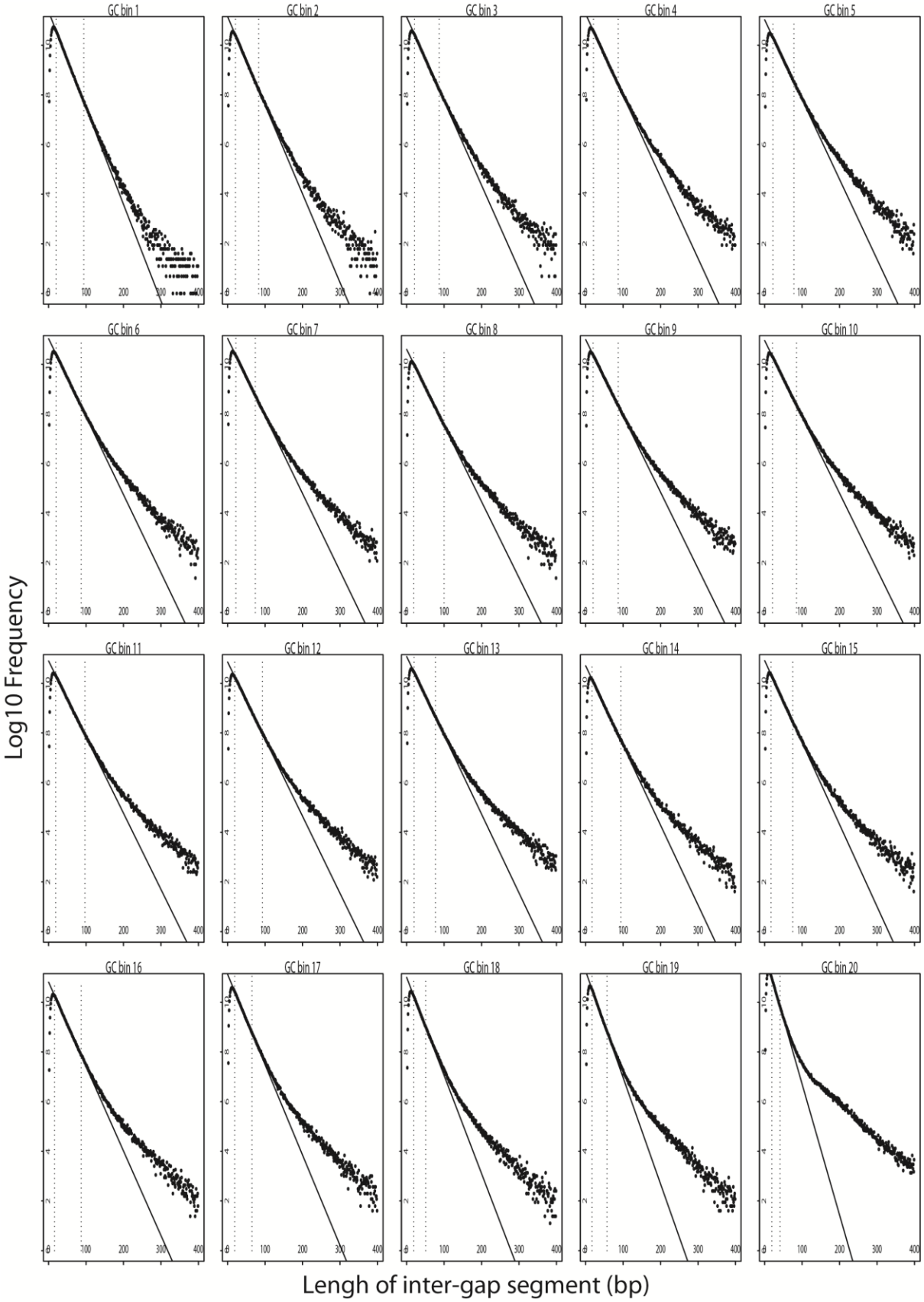

Figure A.6: IGS histograms across 20 equally populated GC-bins from trimmed human (hg19) – cow (bosTau7) alignments.

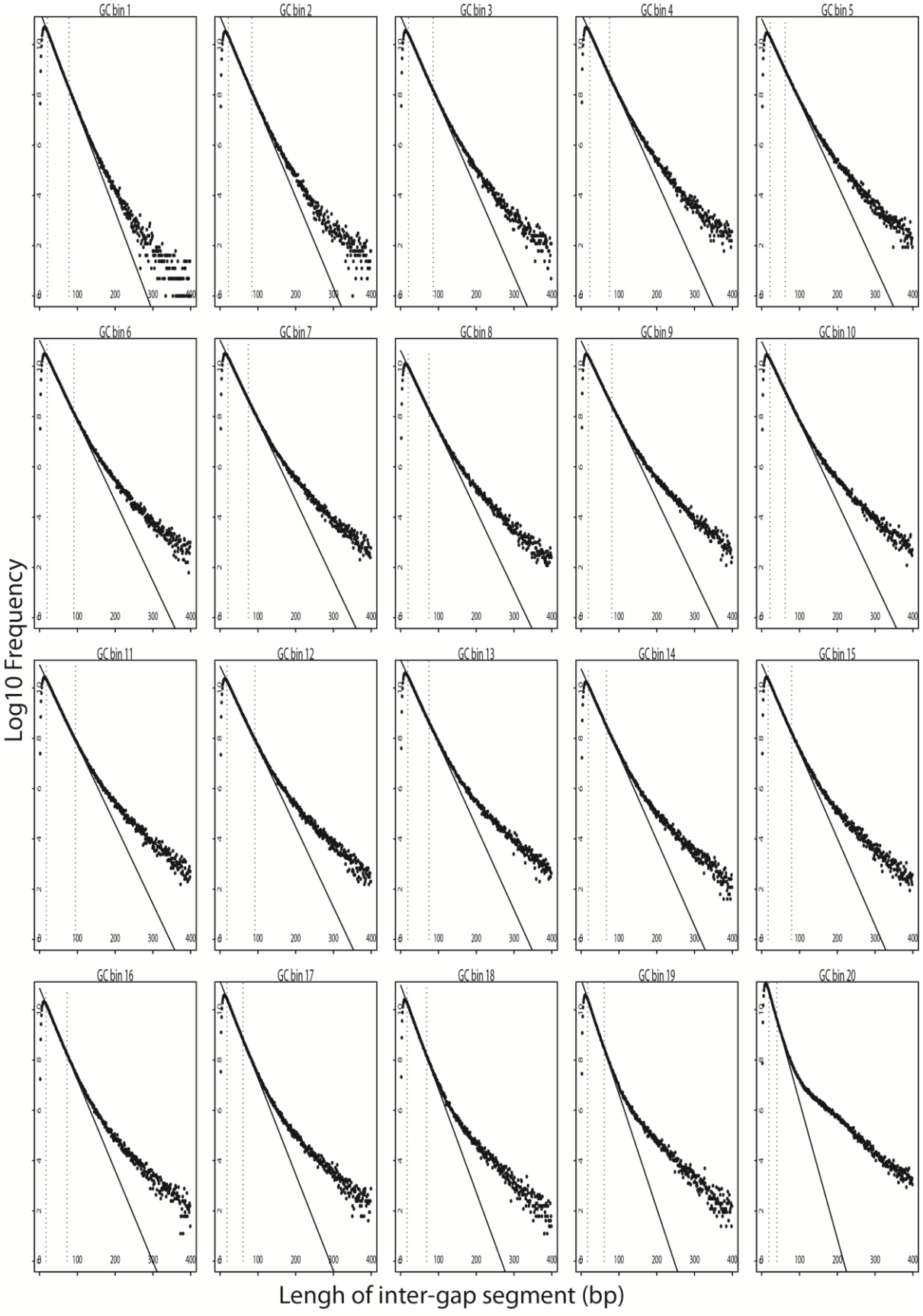

Figure A.7: IGS histograms across 20 equally populated GC-bins from trimmed human (hg19) – rabbit (oryCun2) alignments.

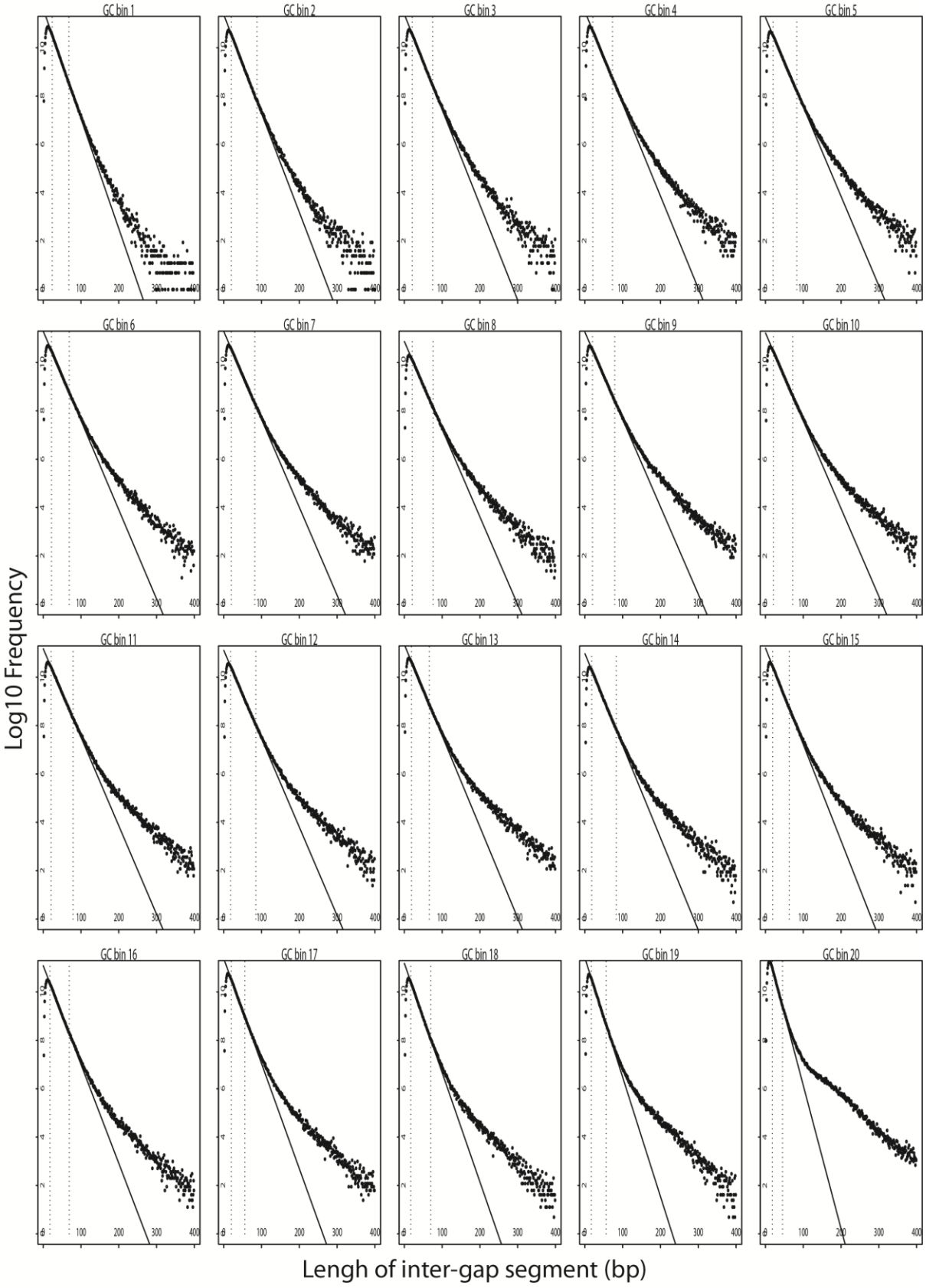

Figure A.8: IGS histograms across 20 equally populated GC-bins from trimmed human (hg19) – guinea pig (cavPor3) alignments.

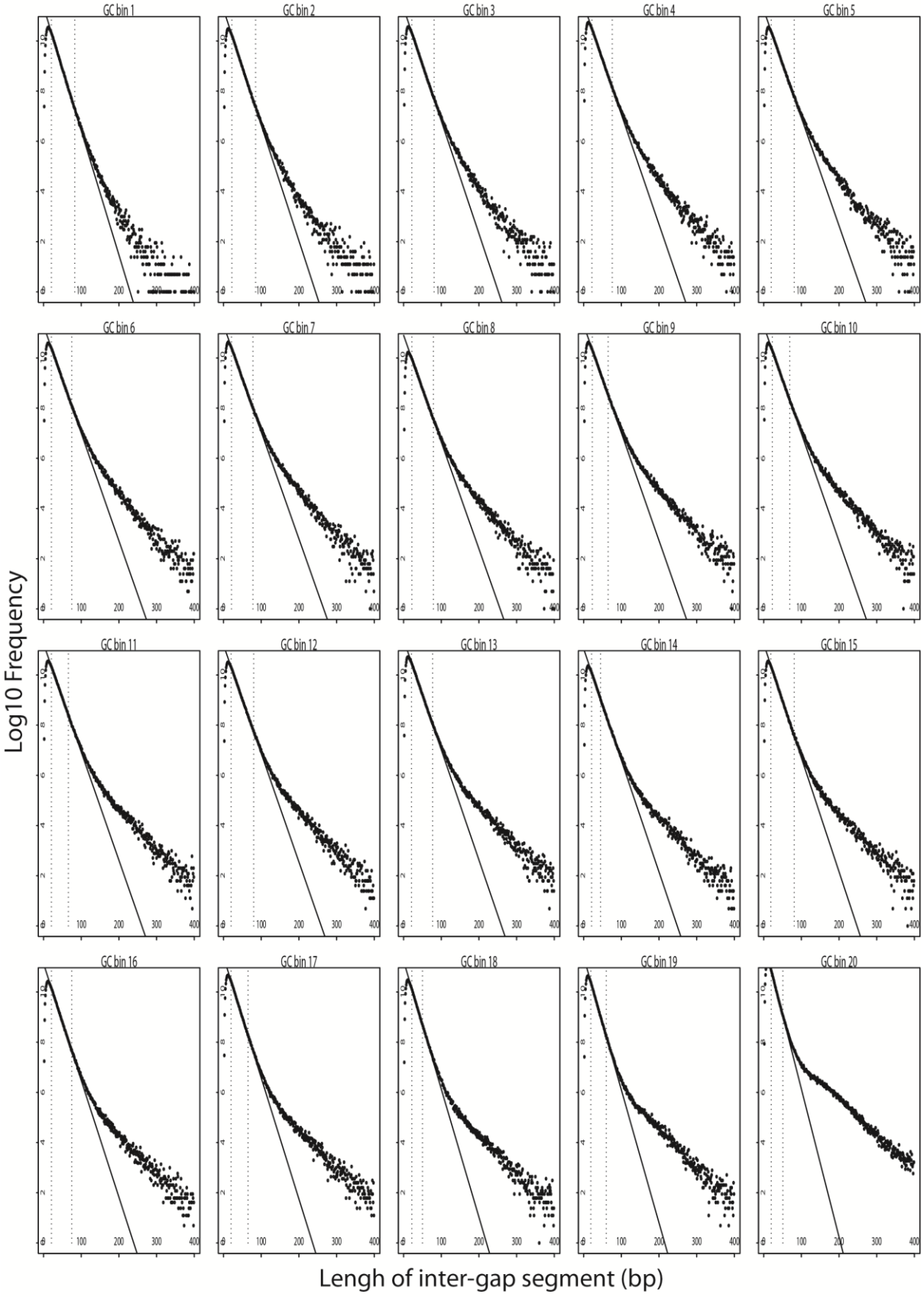

Figure A.9: IGS histograms across 20 equally populated GC-bins from trimmed human (hg19) – mouse (mm10) alignments.

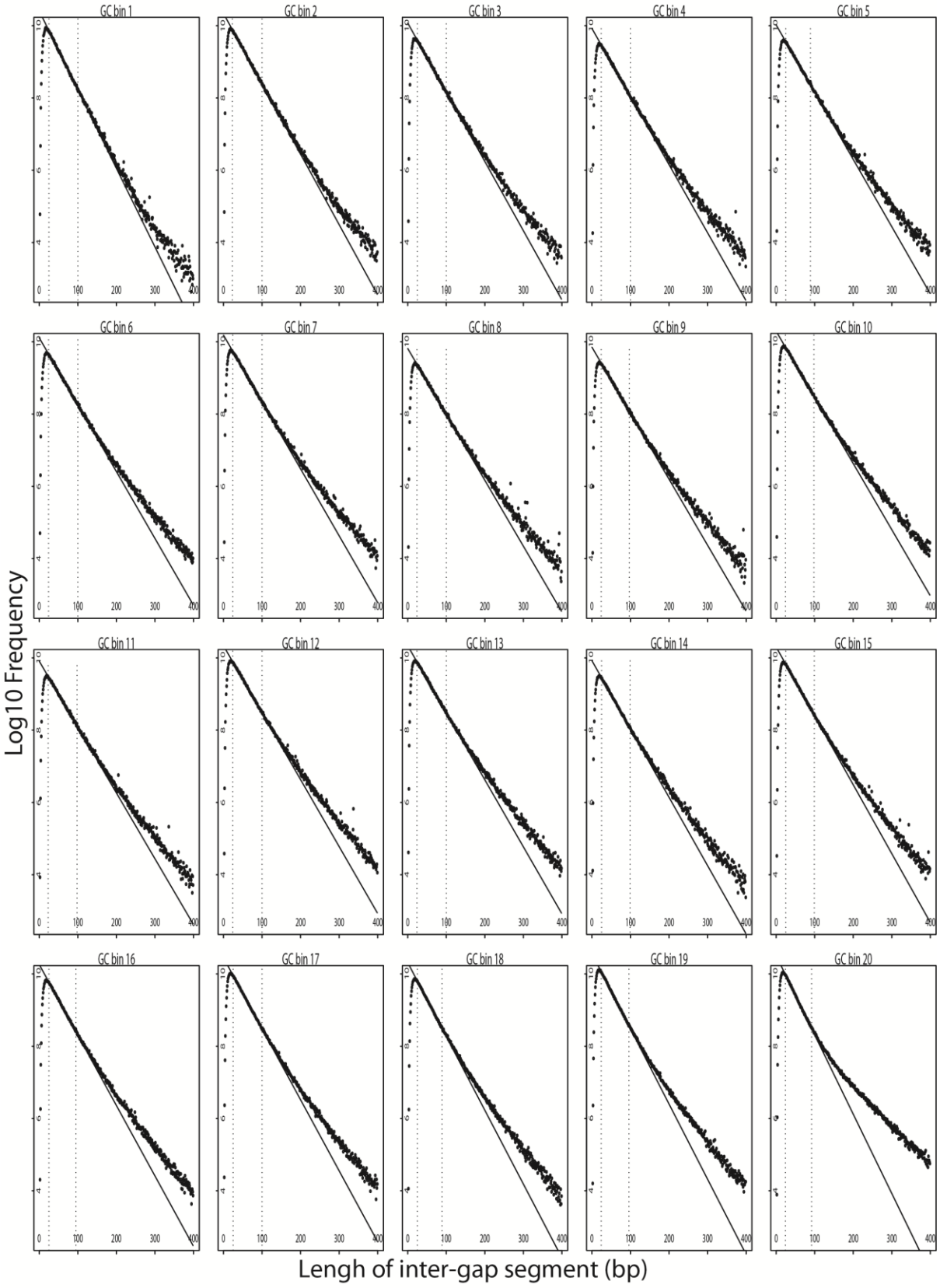

Figure A.10: IGS histograms across 20 equally populated GC-bins from trimmed mouse (mm10) – rat (rn5) alignments.

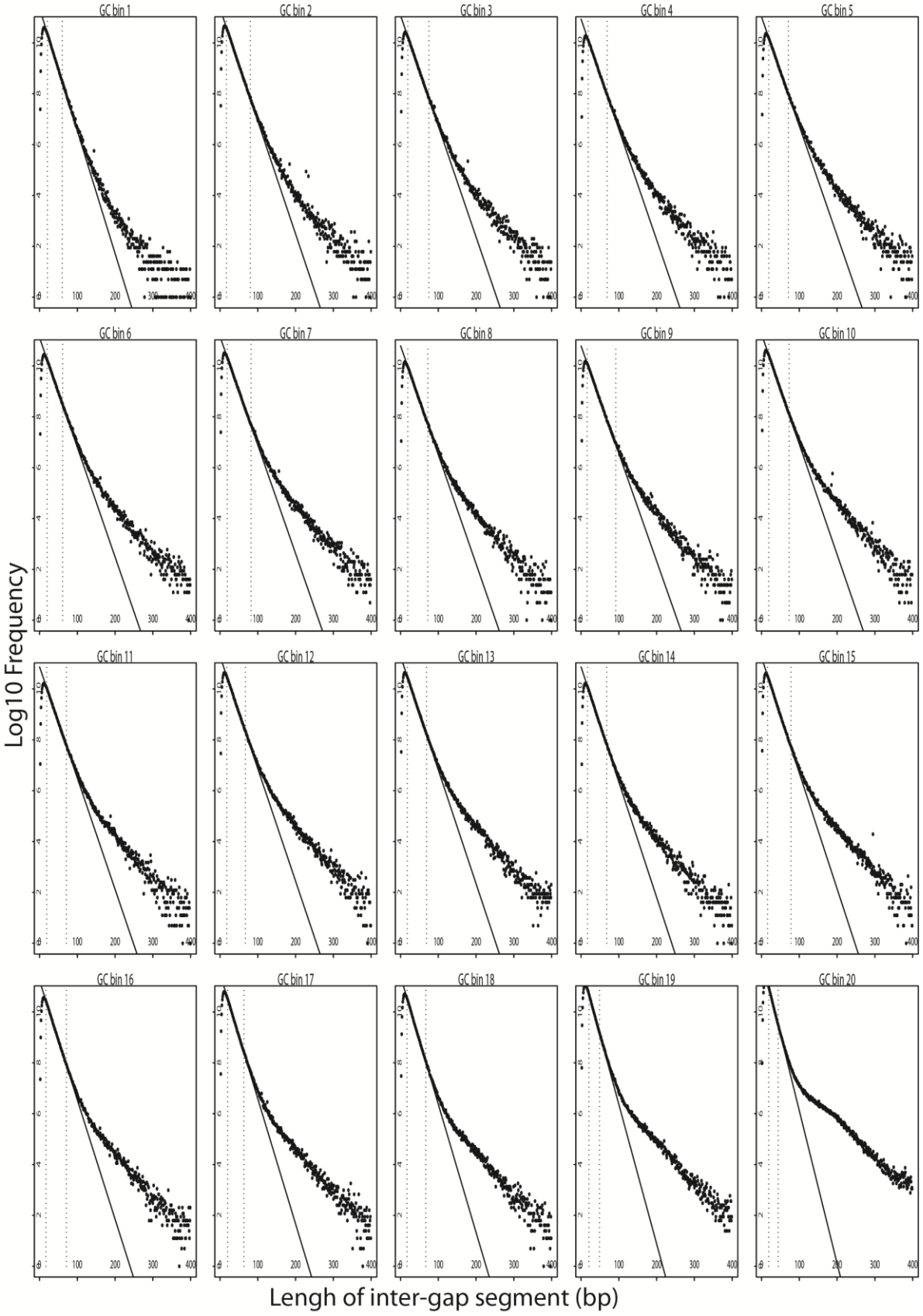

Figure A.11: IGS histograms across 20 equally populated GC-bins from trimmed mouse (mm10) – horse (equCab2) alignments.

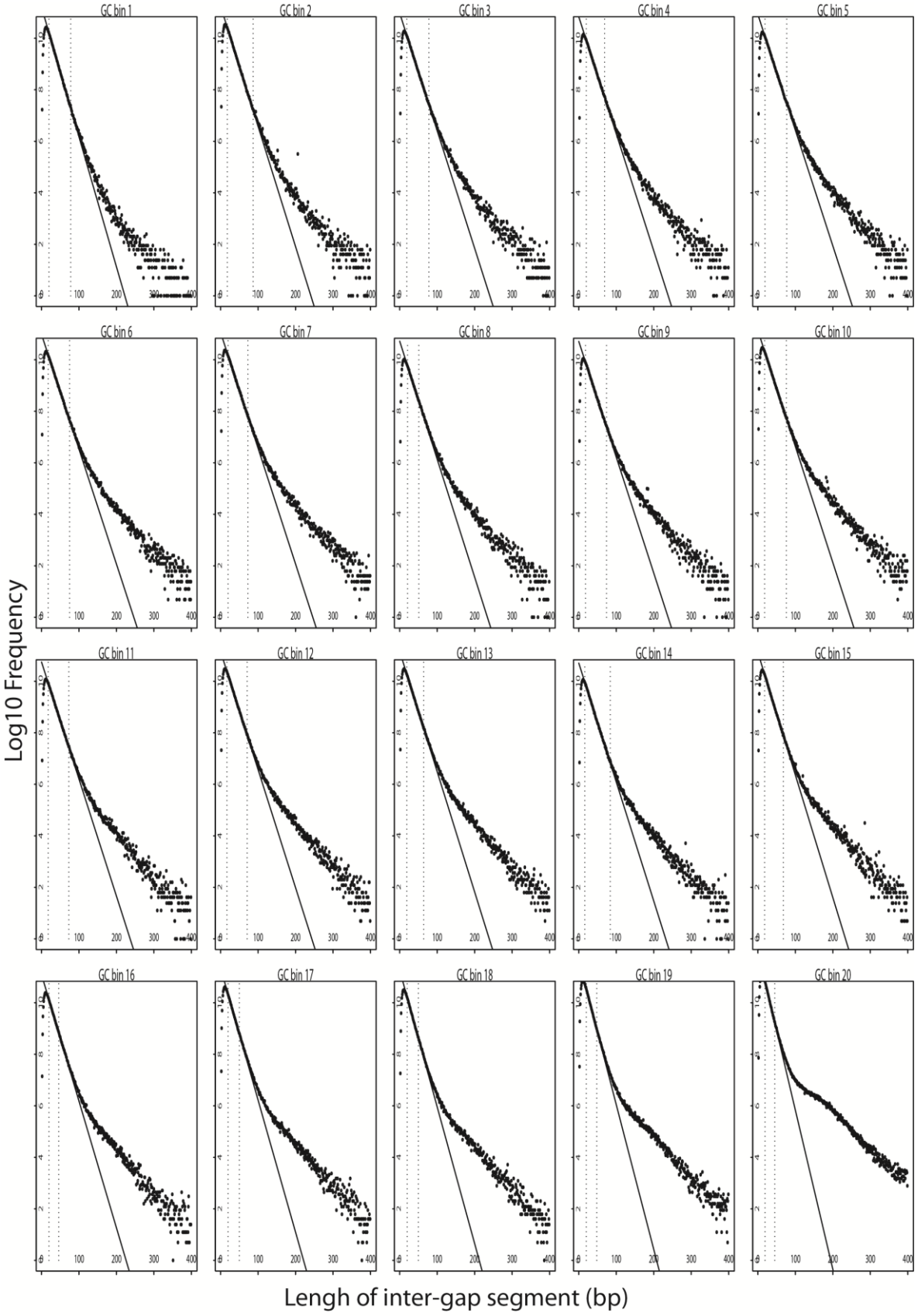

Figure A.12: IGS histograms across 20 equally populated GC-bins from trimmed mouse (mm10) – dog (canFam2) alignments.

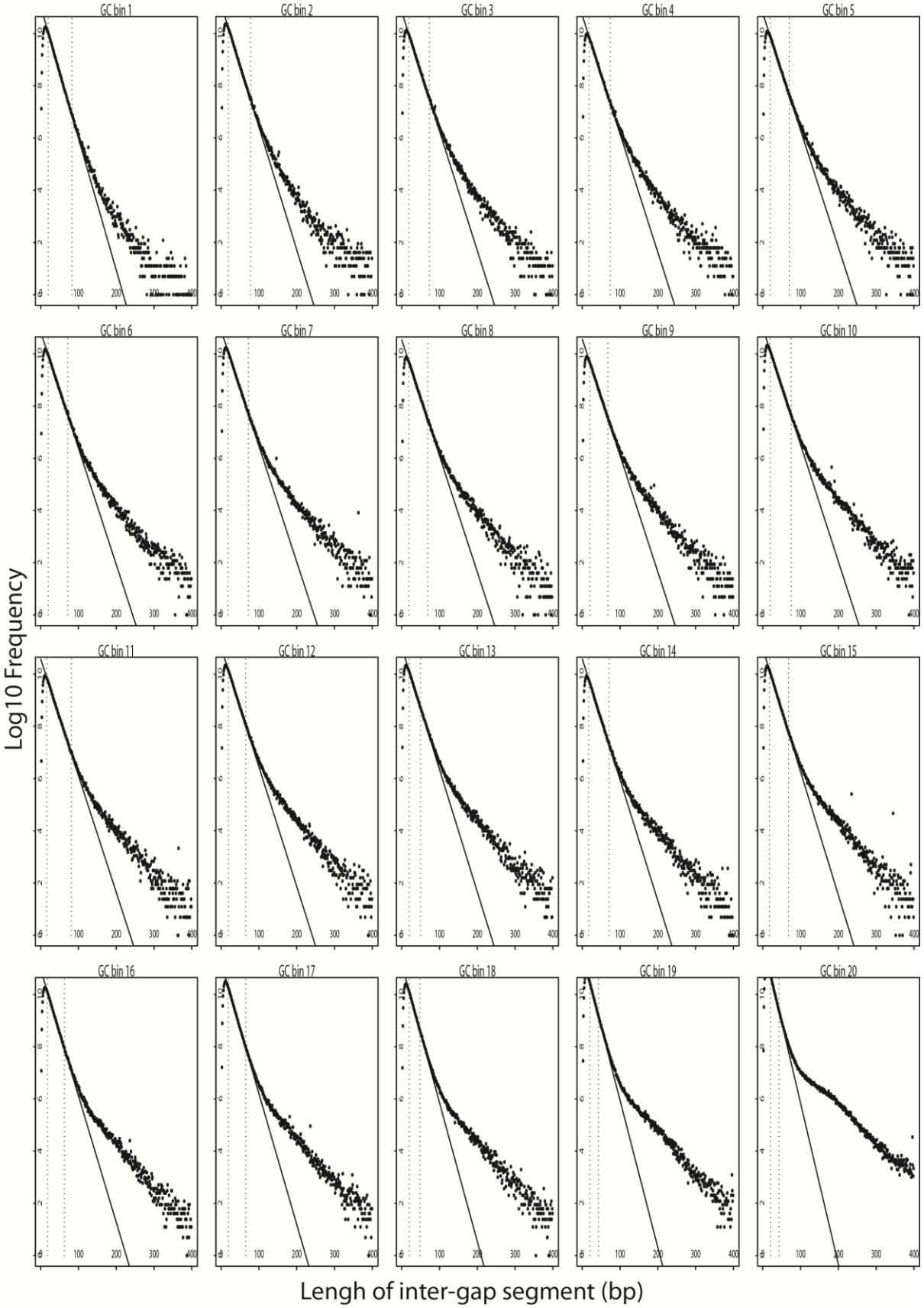

Figure A.13: IGS histograms across 20 equally populated GC-bins from trimmed mouse (mm10) – cow (bosTau7) alignments.

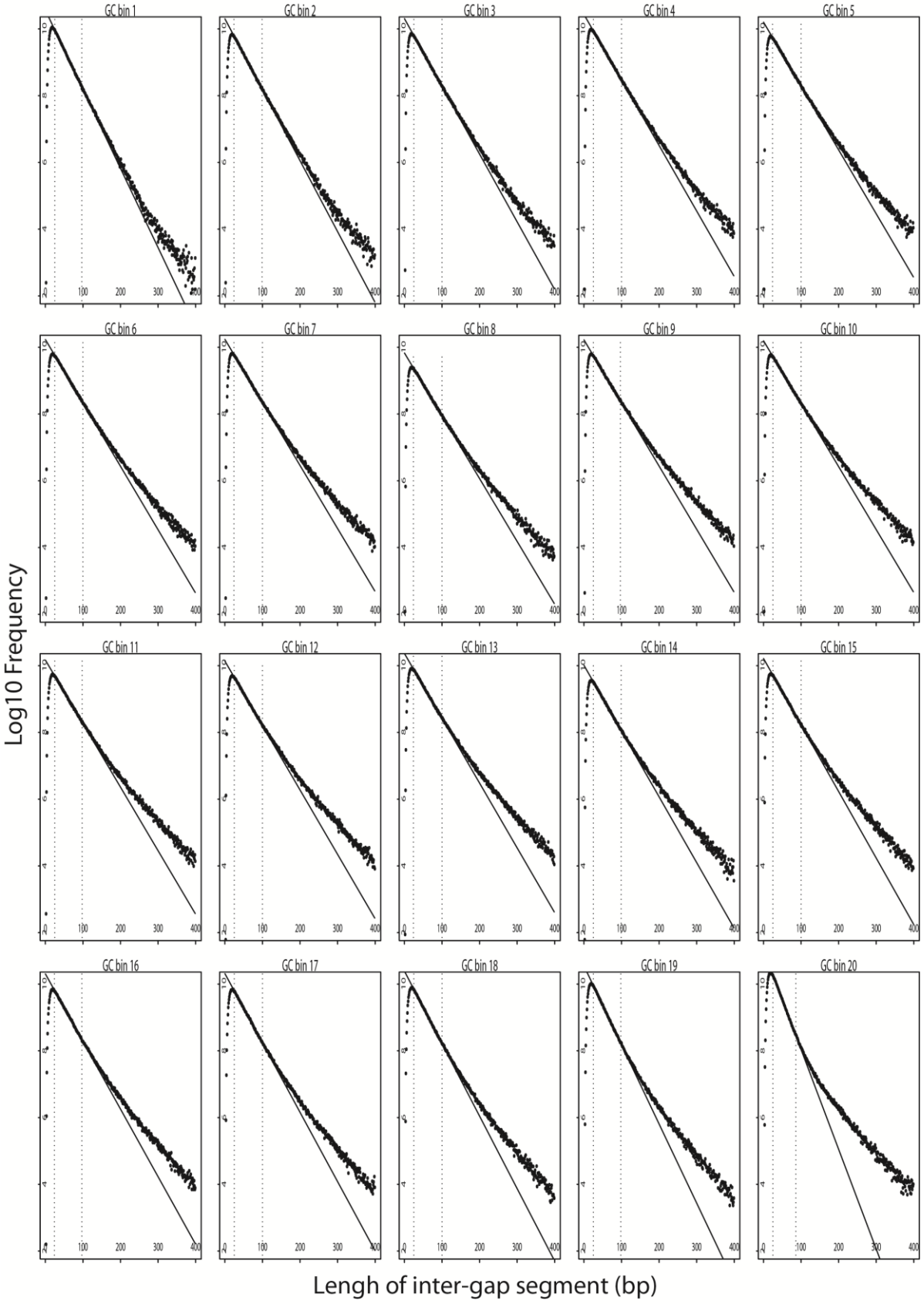

Figure A.14: IGS histograms across 20 equally populated GC-bins from trimmed dog (canFam2) – ferret (mpf_v1) alignments.

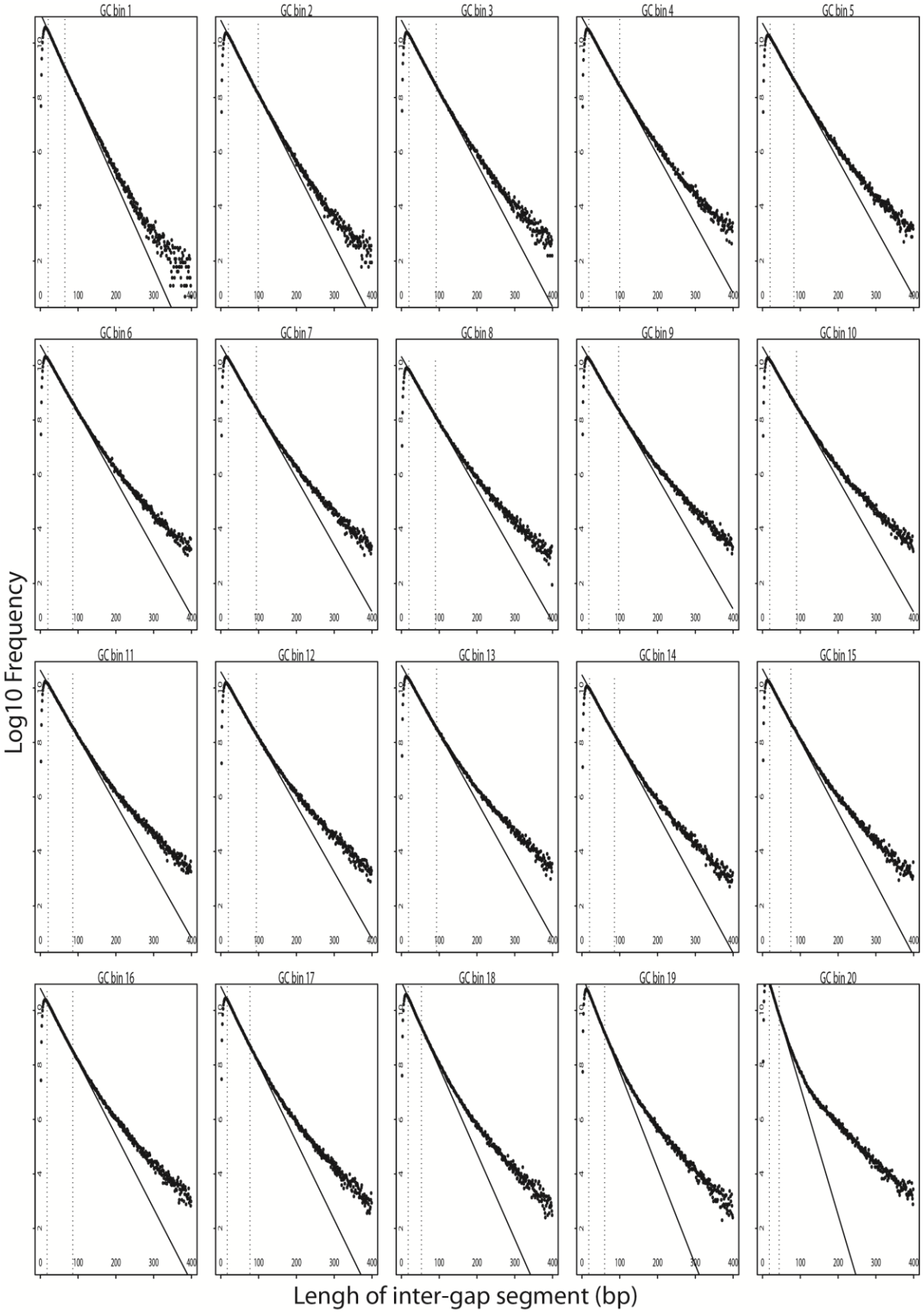

Figure A.15: IGS histograms across 20 equally populated GC-bins from trimmed dog (canFam2) – horse (equCab2) alignments.

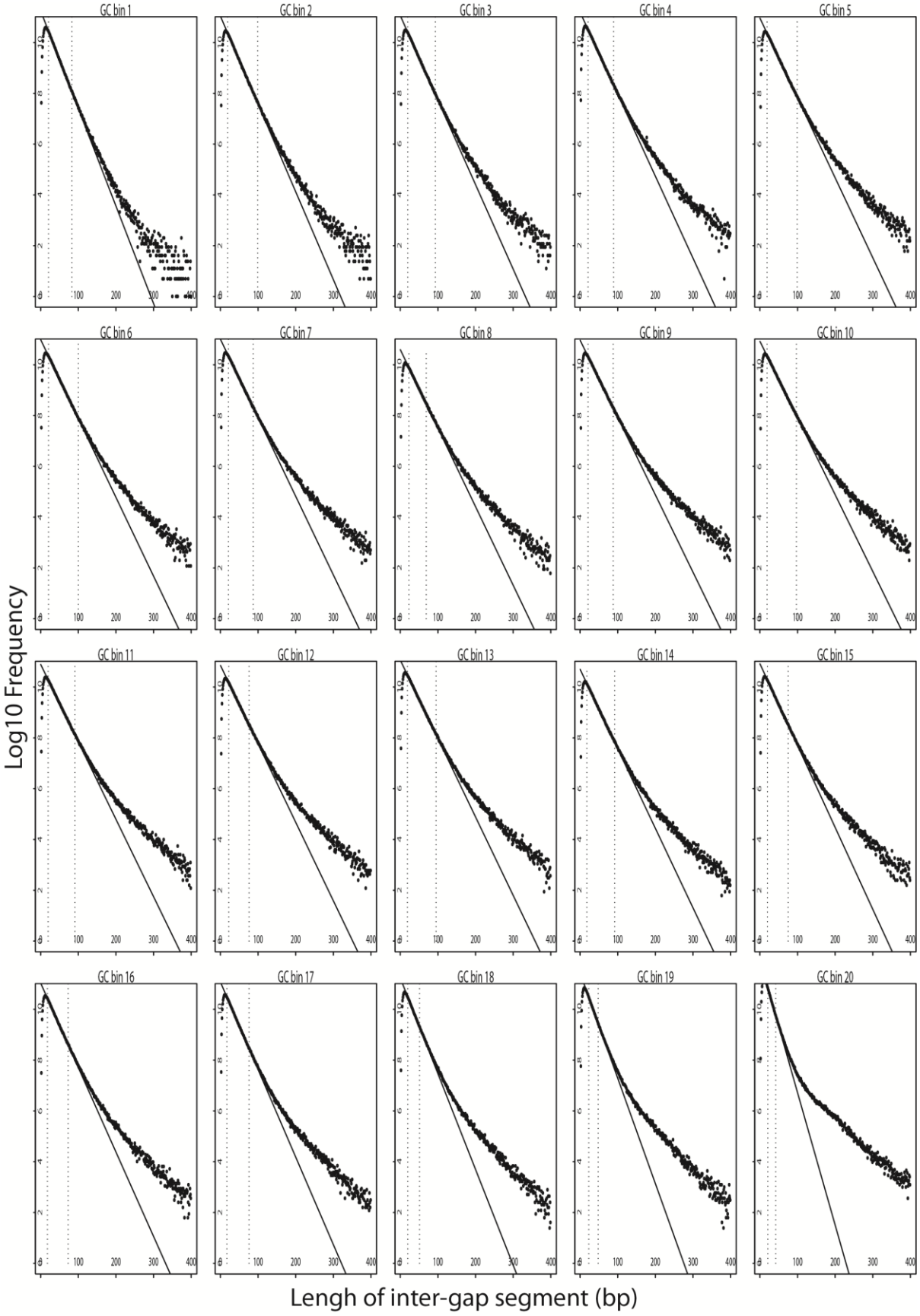

Figure A.16: IGS histograms across 20 equally populated GC-bins from trimmed dog (canFam2) – cow (bosTau7) alignments.

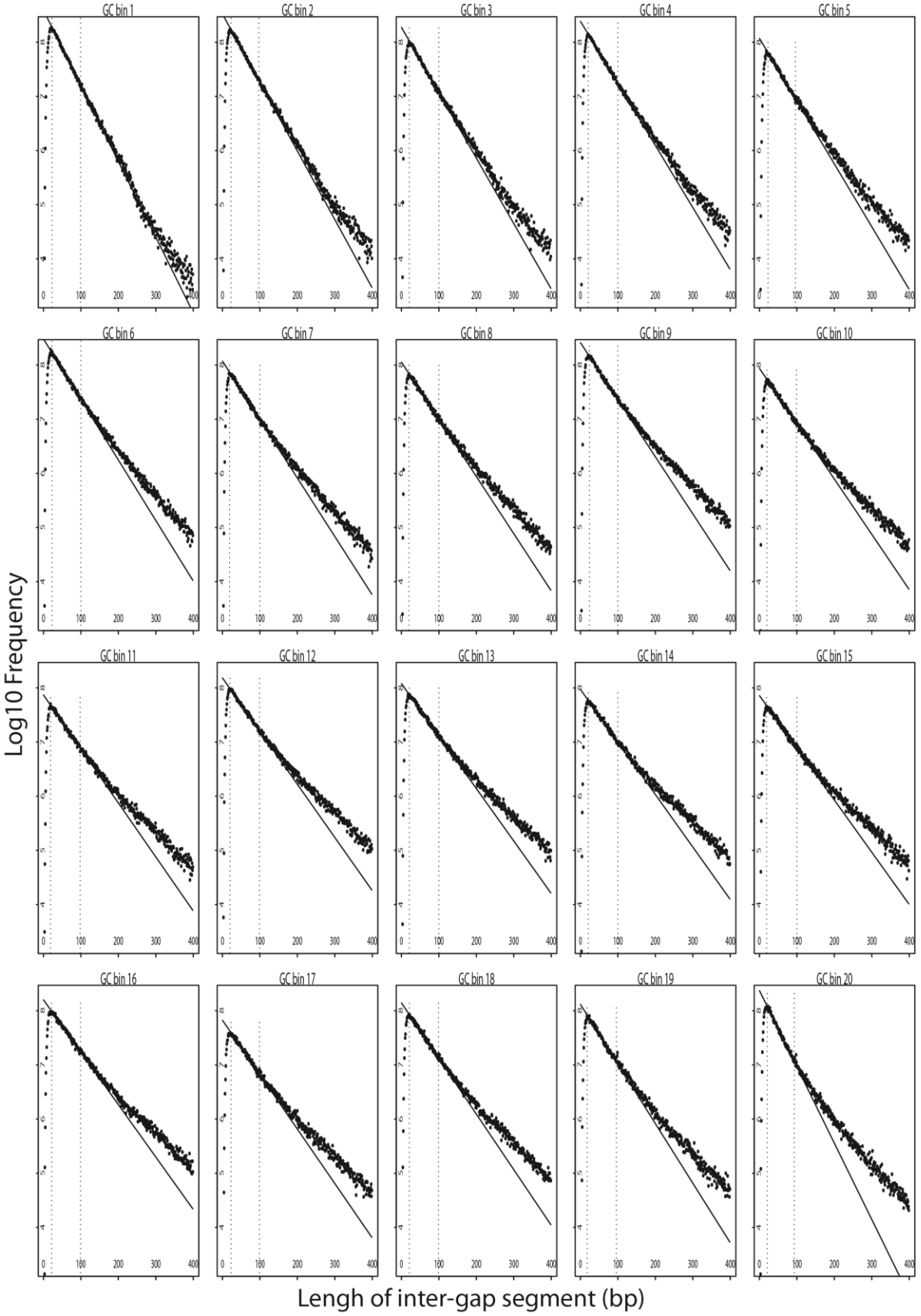

Figure A.17: IGS histograms across 20 equally populated GC-bins from trimmed chicken (galGal3) – turkey (melGal1) alignments.

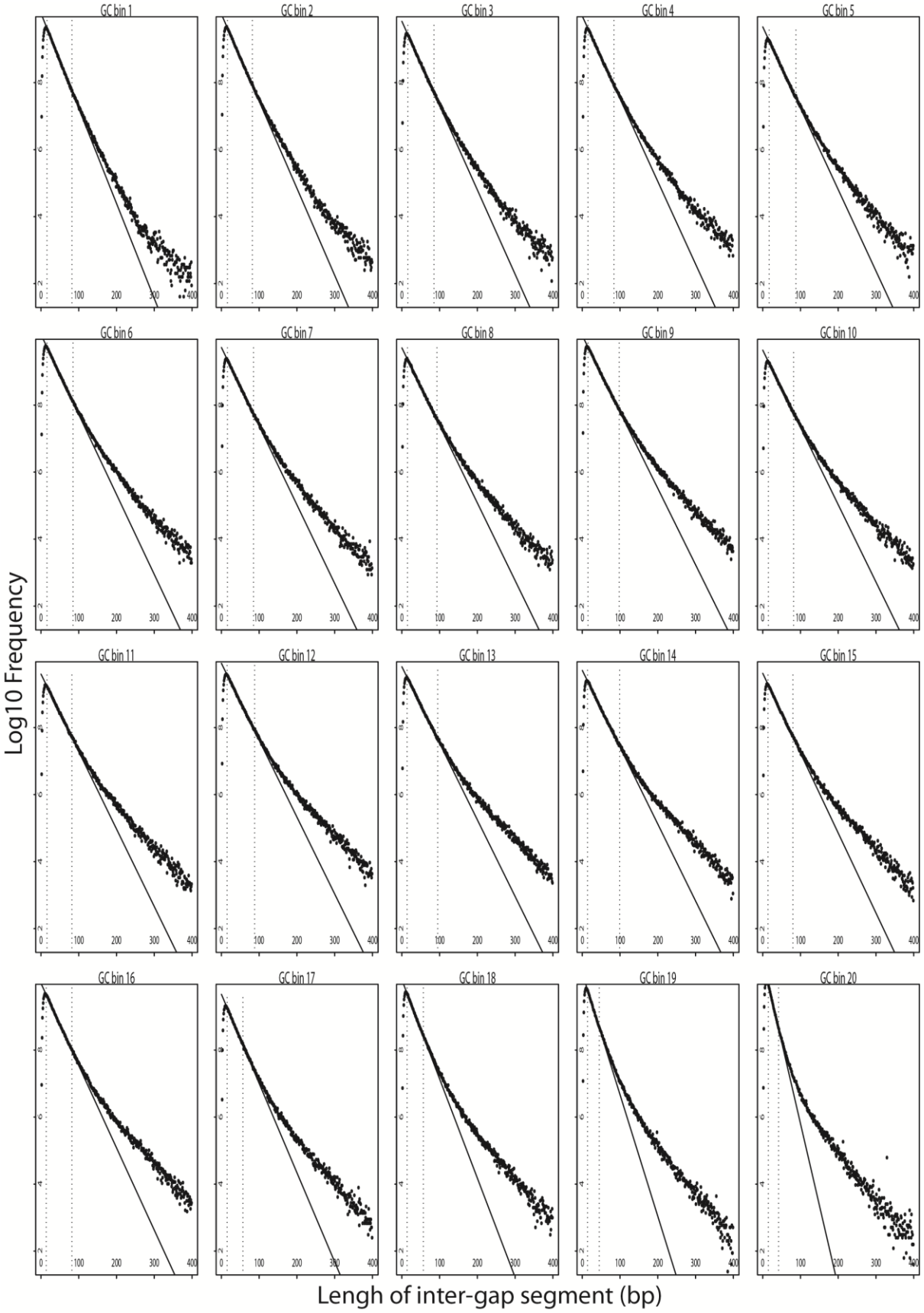

Figure A.18: IGS histograms across 20 equally populated GC-bins from trimmed chicken (galGal3) – adelie (pa_v1) alignments.

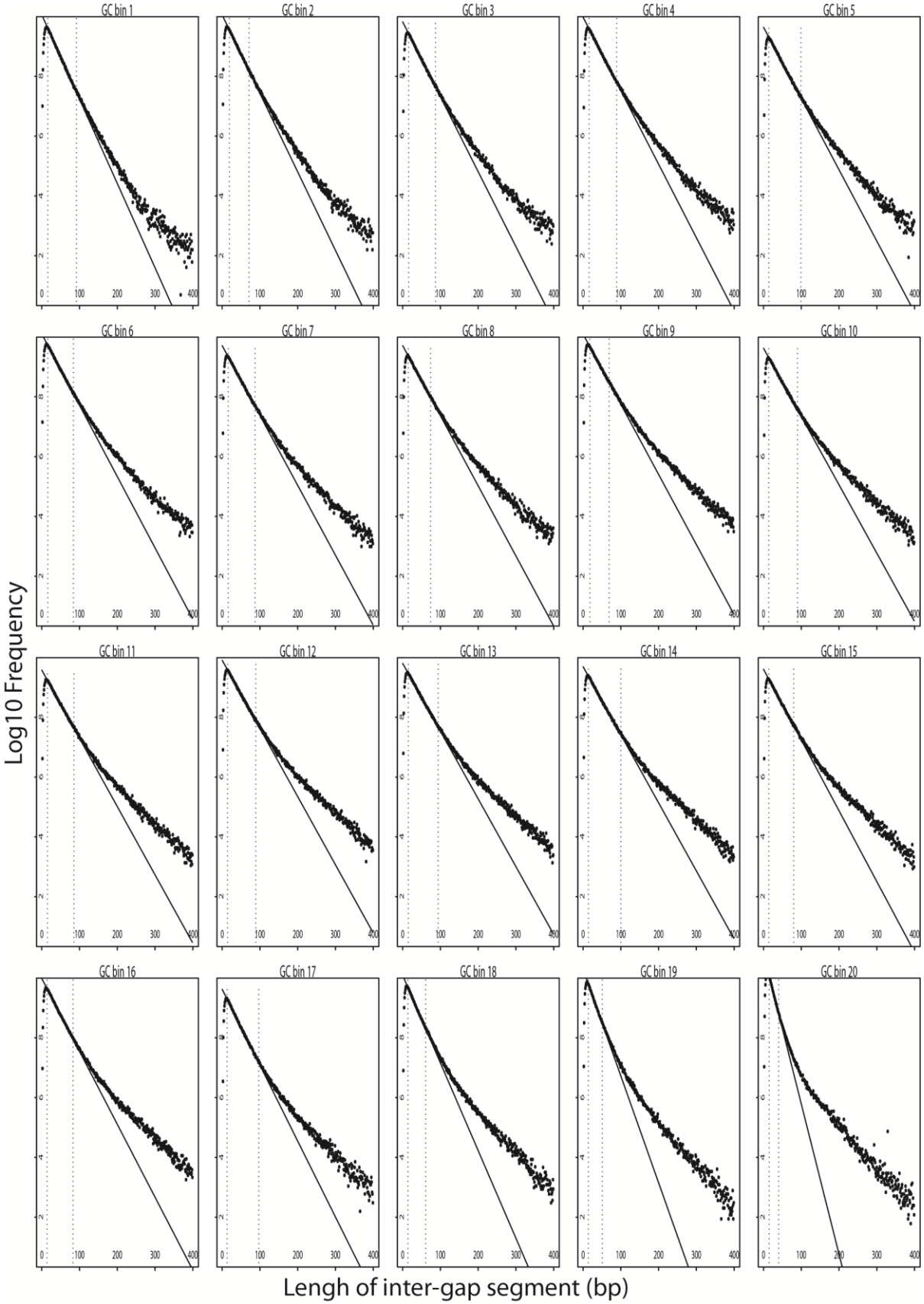

Figure A.19: IGS histograms across 20 equally populated GC-bins from trimmed chicken (galGal3) – emperor (af_v1) alignments.

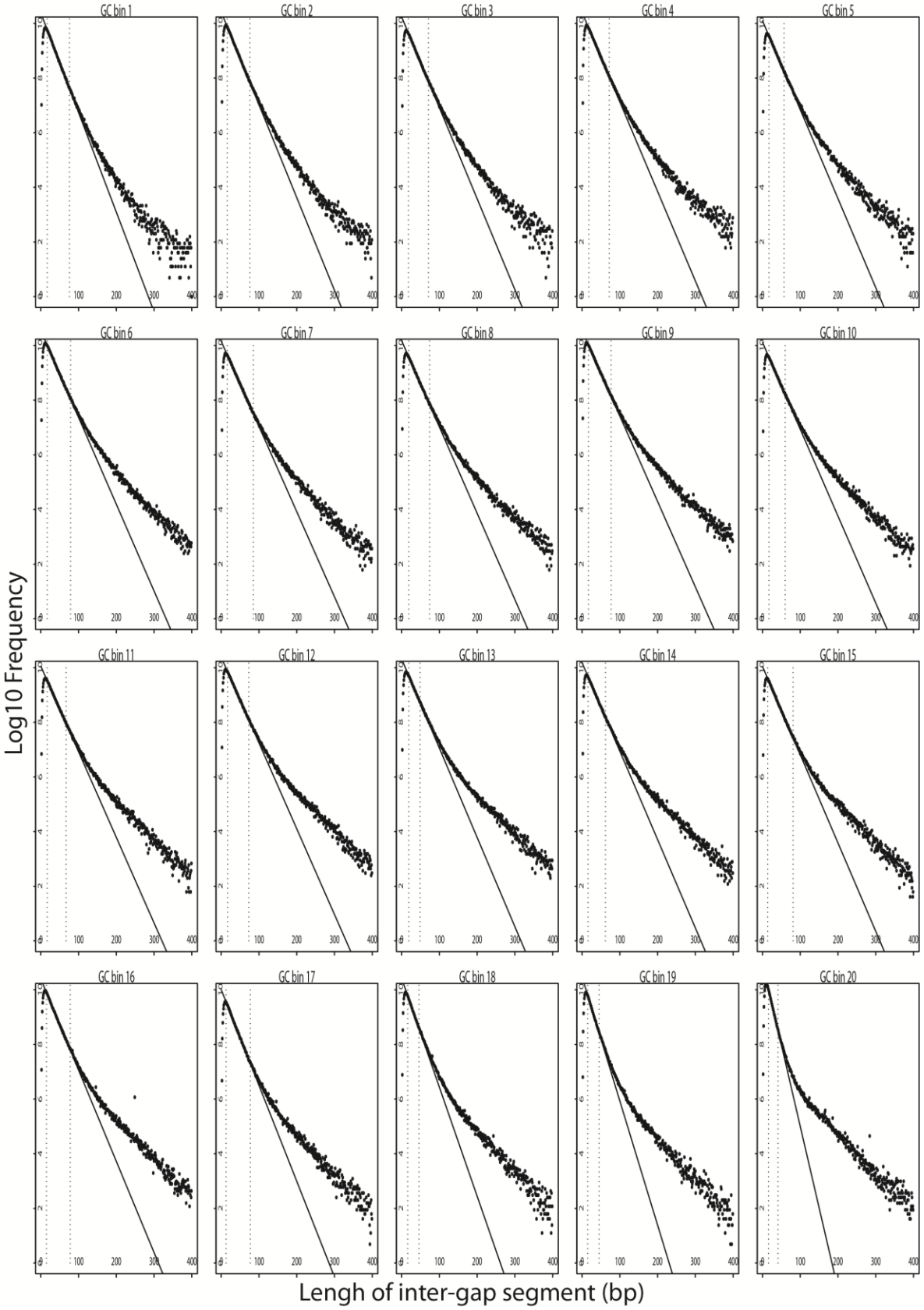

Figure A.20: IGS histograms across 20 equally populated GC-bins from trimmed chicken (galGal3) – zebra finch (taeGut1) alignments.

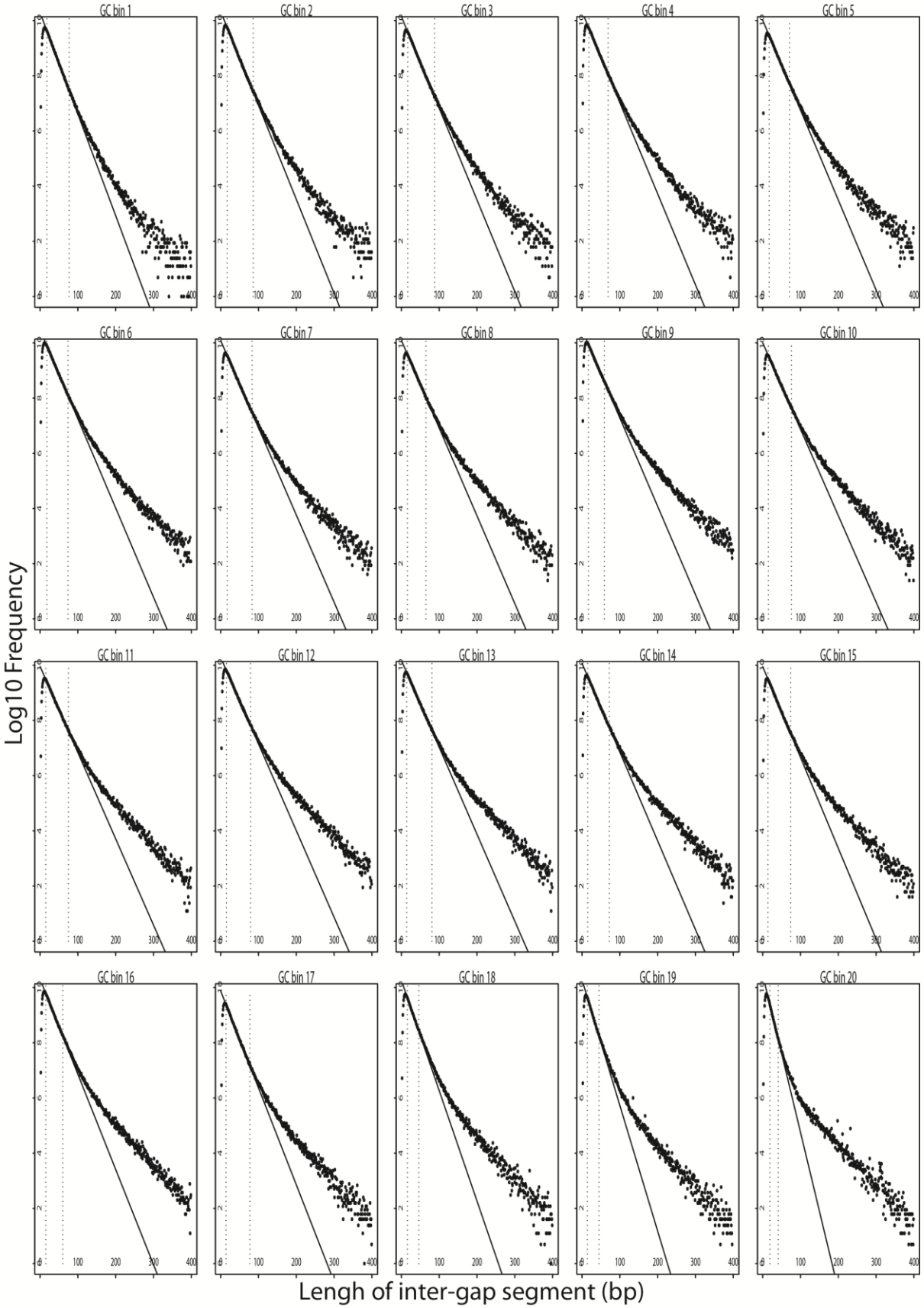

Figure A.21: IGS histograms across 20 equally populated GC-bins from trimmed chicken (galGal3) – Darwin's finch (gm_v1) alignments.

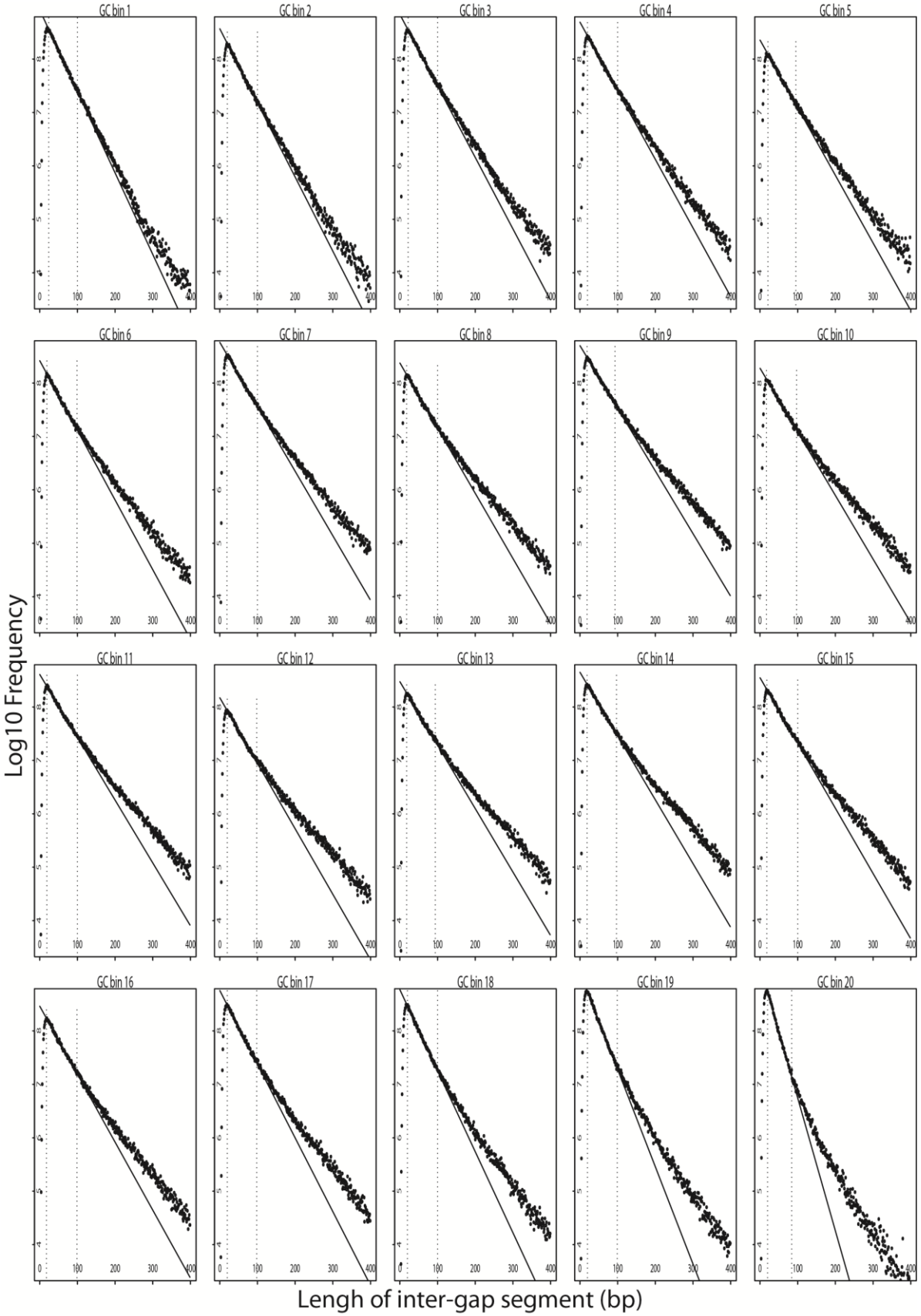

Figure A.22: IGS histograms across 20 equally populated GC-bins from trimmed zebra finch (taeGut1) – Darwin's finch (gm_v1) alignments.

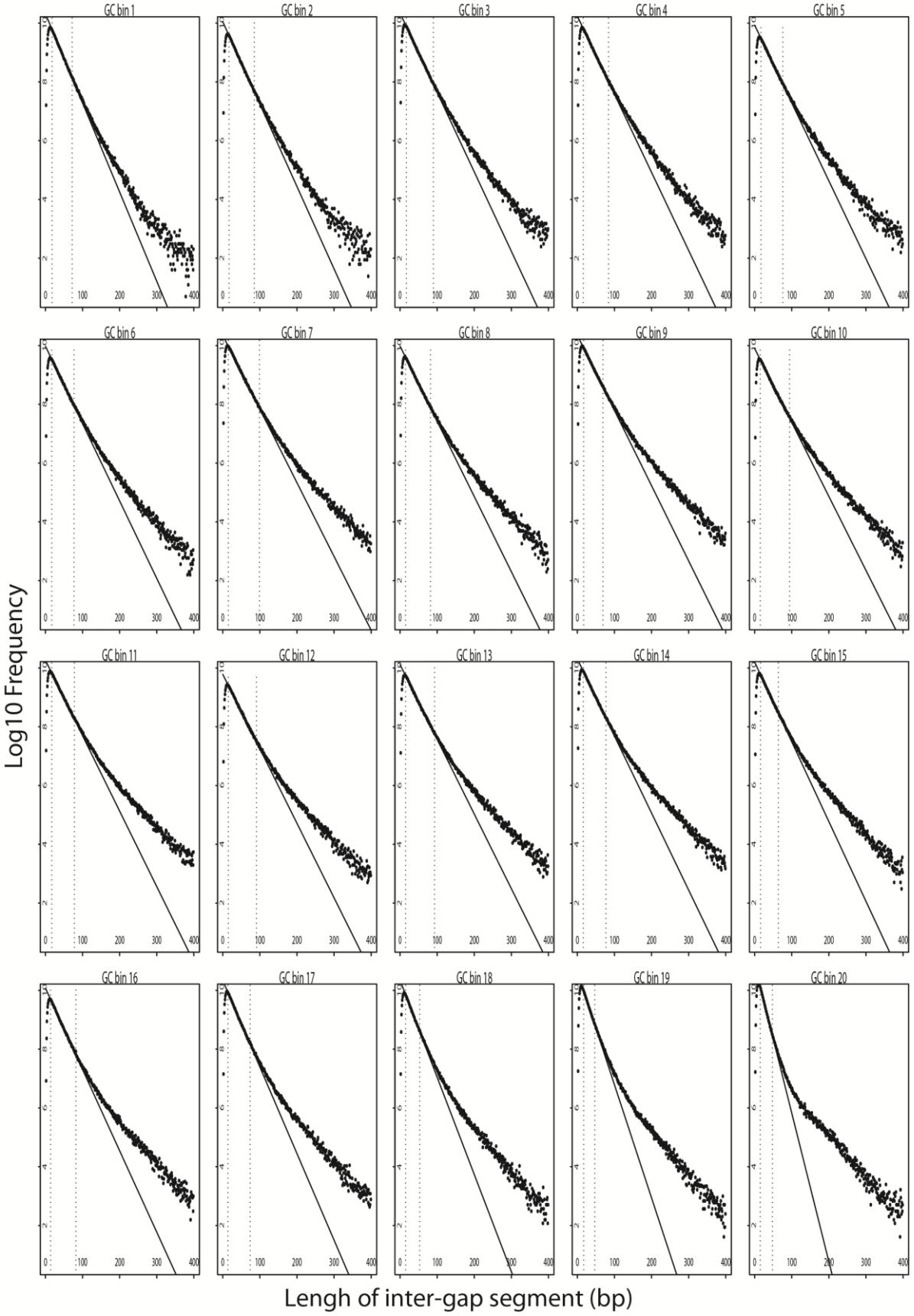

Figure A.23: IGS histograms across 20 equally populated GC-bins from trimmed zebra finch (taeGut1) – budgerigar (melUnd1) alignments.

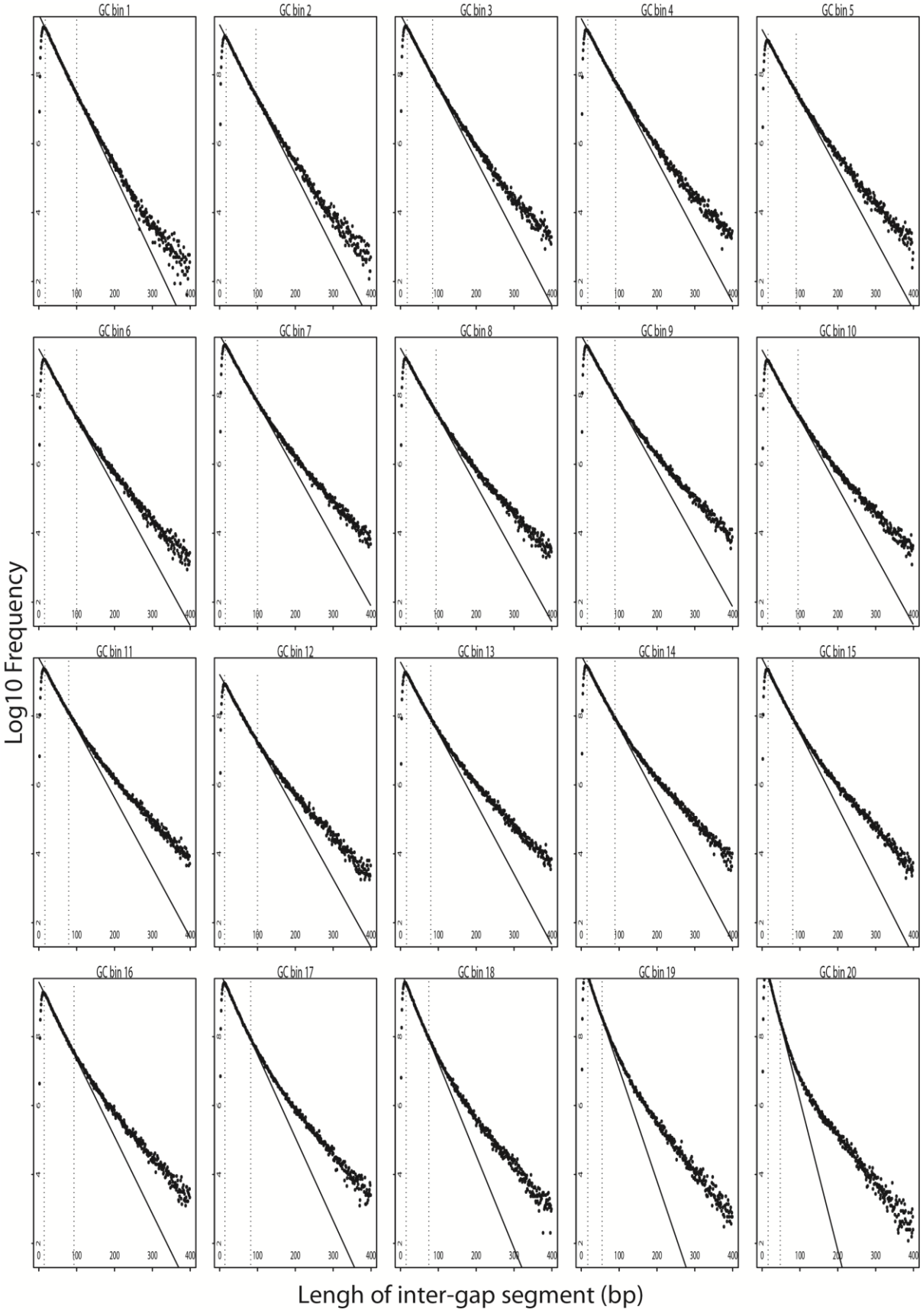

Figure A.24: IGS histograms across 20 equally populated GC-bins from trimmed zebra finch (taeGut1) – adelie (pa_v1) alignments.

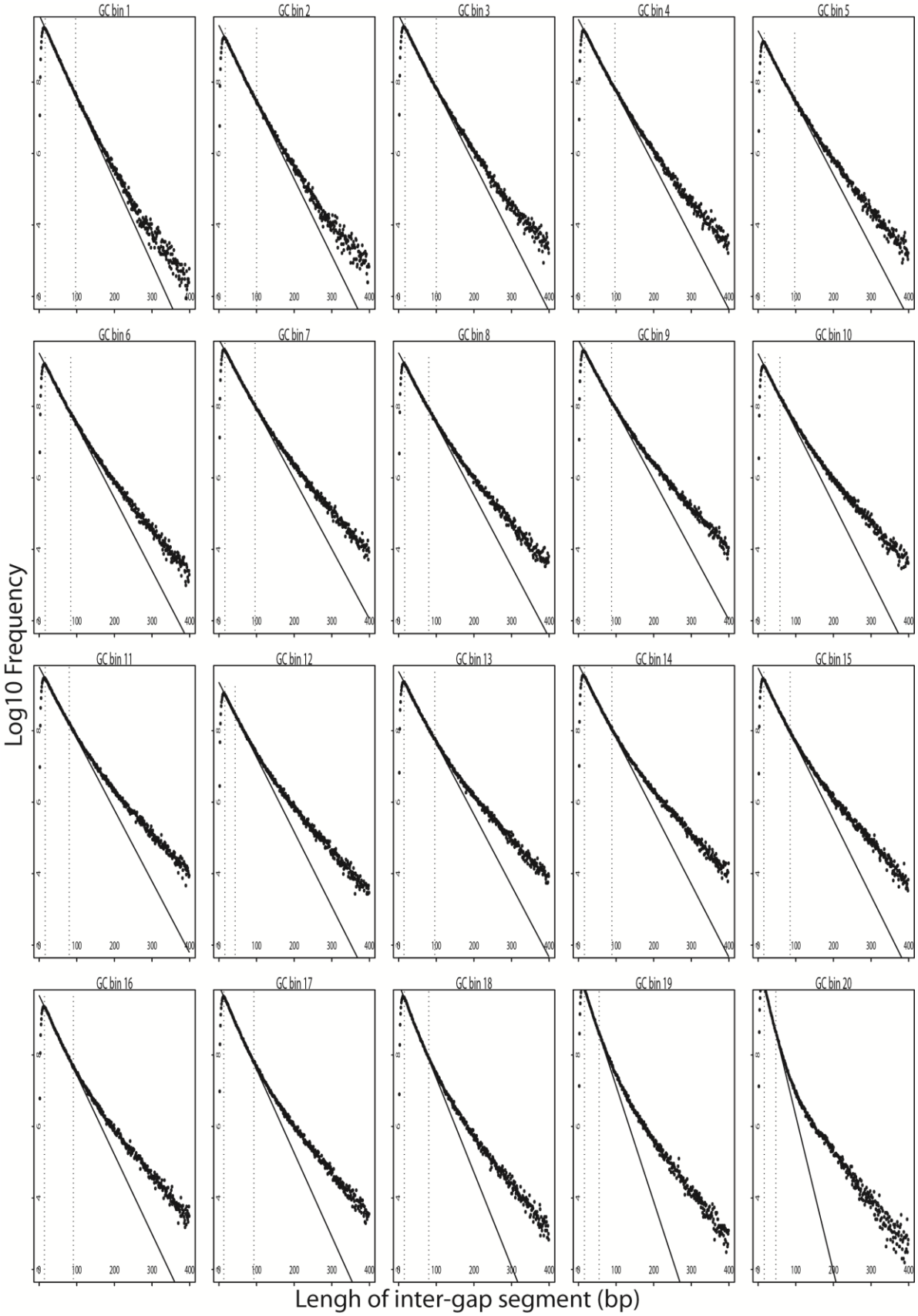

Figure A.25: IGS histograms across 20 equally populated GC-bins from trimmed zebra finch (taeGut1) – emperor (af_v1) alignments.