Increased genome sampling reveals novel insights into vertebrate molecular evolution

A thesis submitted to the National University of Ireland for the Degree of
Doctor of Philosophy

Presented by:
Aoife Doherty
Department of Biology,
NUI Maynooth,
Maynooth,
Co. Kildare, Ireland.

NUI MAYNOOTH

Ollscoil na hÉireann Má Nuad

October 2012

Supervisor
Professor James McInerney B. Sc., Ph.D. (Galway)

Head of Department
Professor Paul Moynagh, B.A (mod), Ph.D. (Dublin)

Table of Contents

For my parents and sister

Acknowledgements

Firstly, I want to sincerely thank Prof. James McInerney for the guidance over the three years. I am extremely lucky to have had such a supportive and patient supervisor.

To everyone in the Bioinformatics unit, thank you so much for all the help, particularly when I first started out. Thanks to Dr. Davide Pisani for all the insightful comments and advice. To Dr. David Alvarez-Ponce with whom I collaborated on a section of this thesis, thank for all the patience. A particular thank you to Leanne and Sinead, conferences and thesis writing and all that fun stuff just wouldn't have been the same without you guys. I want to thank Mr. Brian Daly for all his patience with computer resources when I first started. Similarly, I want to thank Dr. Vanush Paturyan for the technical support and for general brain-picking every once in a while.

This research would not have been possible without funding from IRCSET and the computational resources of ICHEC and HPC at NUI Maynooth. I genuinely appreciate both of these resources.

Cat, don't cry when you read this ☺ Thanks so much for everything. The trouble I've gotten into with you could fill a book far longer than this thesis! I couldn't have asked for a nicer person to grow up with, I am SO lucky to still have the same best friend as I did well over a decade ago. To the girls, thanks for all the encouragement and for organizing plenty of post-thesis banter to make up for what I've missed the last while, love you guys.

Peter, I'm so happy you're back. It just wasn't the same without your awful puns. For a while, I had no one to "share" food with, beat at pool or accidently ruin the ending of films for ☺.

And finally to my family. To mam and dad, thank you SO MUCH for always giving up everything for me ☺. I appreciate everything you've done for Niamh and I. If science doesn't work out, I definitely have a career playing patience in Vegas after the amount I've won the last few months ☺ Also thank you for everything you've done especially in the last few months that I've been home, it's going to be hard to get me to leave when I know how spoilt I am here ☺ Nurse Niamhy, thanks for all the lolling. Also, thanks for being able to sleep 23 hours a day, leaving me a lovely quiet house to write in. To my extended family and friends and everyone around me, love you now and always ☺ xxx

Declaration

This thesis has not been submitted in whole, or in part, to this, or any other University for any other degree and is, except where otherwise stated, the original work of the author.

Signed: _____

Aoife Marie Doherty

Abbreviations

AIC	Akaike Information Criterion
AU	Approximately Unbiased
BF	Bayes Factors
BIC	Bayesian Information Criterion
BLAST	Basic Local Alignment Search Tool
BLOSUM	Blocks of Amino Acid Substitution Matrix
BMGE	Block Mapping and Gathering with Entropy
BNC	Base Nucleotide Composition
CA	Correspondence Analysis
CDC	Codon Deviation Co-efficient
CIR	Cox-Ingersoll-Ross
CUB	Codon Usage Bias
DM	Distance Matrix
DNA	Deoxyribonucleic Acid
E-value	Expect-value
F81	Felsenstein 1981
FSA	Fast Statistical Alignment
GO	Gene Ontology
GTR	General Time Reversible
HIV	Human Immunodeficiency Virus
HKY85	Hasegawa et al., 1985
JC	Jukes-Cantor
JTT	Jones-Taylor-Thornton

K2P	Kimura-2-Parameter
LBA	Long Branch Attraction
LogN	Lognormal
MCMC	Markov Chain Monte Carlo
ML	Maximum Likelihood
MP	Maximum Parsimony
MSA	Multiple Sequence Alignment
MVA	Multivariate Analysis
MYA	Million Years Ago
N_c	Effective Number of Codons
NJ	Neighbor Joining
NNI	Nearest Neighbor Interchange
OTU	Operational Taxonomic Unit
PAM	Point Accepted Mutation
PIN	Protein Interaction Network
PP	Posterior Probability
PRANK	Probabilistic Alignment Kit
RNA	Ribonucleic Acid
RSCU	Relative Synonymous Codon Usage
SCU	Synonymous Codon Usage
SPR	Subtree Pruning and Regrafting
tRNA	Transfer RNA
UGam	Uncorrelated Gamma
UPGMA	Unweighted Pair Group Method with Arithmetic Mean

WAG	Whelan and Goldman
WGD	Whole Genome Duplication

Index of Tables

Chapter 1

Chapter 2

Chapter 3

Chapter 4

Index of Figures

Chapter 1

Chapter 2

Chapter 3

Chapter 4

Index of Equations

Index of Scripts

Index of Appendices

Index of Electronic Appendices

Abstract

In this thesis, increased vertebrate genome sampling and recent methodological advancements were combined to address three distinct questions pertaining to vertebrate molecular evolution.

Gene duplicability is the tendency to retain multiple gene copies after a duplication event. Various factors correlate with gene duplicability, such as protein function and timing of expression during development. The position of a gene's encoded product in the protein-protein interaction network recently emerged as an additional factor determining gene duplicability. The first investigation described in this thesis coupled comparative genomics with protein-protein interaction data to assess the dynamic relationship between gene duplicability and network structure in primates.

Deciphering the timing of the Ursidae (bear) phylogeny speciation events has proven to be a challenging task. A valuable node to calibrate in such studies is that separating giant panda and polar bear. The exact timing of this important calibration node is currently disputed. The second investigation described in this thesis applied the largest amount of nuclear data currently available in a Bayesian framework to attempt to accurately estimate the timing of divergence between the giant panda and polar bear.

It is known that synonymous codon usage is governed by a combination of selective and neutral processes. Currently, it is thought that primarily neutral processes govern synonymous codon usage in vertebrates, possibly due to their lower long-term effective population sizes. The third investigation described in this thesis combined increased genomic sampling and a novel codon usage bias index to conduct the first systematic investigation into the forces that govern synonymous codon usage in vertebrates.

Chapter 1: Introduction

Since the latter half of the 20th century, molecular data has become a standard means of exploring evolutionary hypotheses. This has culminated in the present era, in which a meteoric rise in whole genome sequence availability offers unprecedented opportunities to explore the central tenets of evolution from a molecular perspective (Eisen and Fraser, 2003, Wolfe and Li, 2003, Mardis, 2008).

This thesis encompasses three separate investigations that explore vertebrate molecular evolution from several angles. Each investigation is distinct regarding the hypothesis under examination and the tools used to consider each hypothesis. The purpose of this introductory chapter is to present fundamental concepts and methods that are commonly employed in molecular evolutionary studies. Each subsequent chapter contains a separate introductory section that is relevant to that particular investigation.

1.1 Molecules as documents of evolutionary history

In the last century, several significant discoveries allowed the concepts and tools of molecular biology to permeate evolutionary investigations. For example, deoxyribonucleic acid (DNA) was determined as the genetic material of a cell (Avery et al., 1944, Hershey and Chase, 1952), the molecular structure of DNA was unraveled (Watson and Crick, 1953) and the genetic code was deciphered (Crick et al., 1961, Nirenberg and Matthaei, 1961, Nirenberg, 2004).

However, two seminal publications ultimately united the field of evolution and molecular biology. First, it was proposed that the number of amino acid residue differences between a pair of molecules could be indicative of the time elapsed since their evolutionary divergence (Zuckerkandl and Pauling, 1962). Second, it was suggested that semantides (a class of molecules encompassing DNA, ribonucleic acid (RNA) and polypeptides) were the most informative molecules for the investigation of evolutionary events (Zuckerkandl and Pauling, 1965). These two influential studies, accompanied by complementary methodological and conceptual advances (for example; Fitch and Margoliash (1967), Dayhoff and Eck (1968), Kimura (1968), Ohno (1970)) provided a foundation upon which the field of modern molecular evolution has been built. Since this time, the discipline has progressed at an exceptional pace, greatly broadening the spectrum of questions that can be addressed from a molecular perspective (Koonin, 2009, Lander, 2011). Over the course of the following sections, some of the key concepts relating to molecular evolution will be discussed.

1.2 Homology and alignment

As all living organisms descended from a common ancestor, homology is a fundamental concept of molecular evolution. Sir Richard Owen introduced the term homolog as "*the same organ in different animals under every variety of form and function*" (Owen, 1843, quoted in Koonin, 2005). It is unsurprising that common ancestry is not included in this definition, as it was a pre-Darwinian and pre-Mendelian era. Currently, homology may be defined in an evolutionary sense as "*the relationship of any two characters that have descended, usually with divergence, from a common ancestral character*" (Fitch, 2000). Prominent forms of homology include orthology and paralogy, describing specific cases that arise from speciation and gene duplication events, respectively.

Homology between genetic sequences is commonly assessed using an expect value ("E-value") statistic that is derived from the Basic Local Alignment Search Tool (BLAST) (Altschul et al., 1990). An E-value describes the number of hits that would be expected to be returned for a particular sequence by chance, given a database of a particular size. In chapter 4 of this thesis, a reciprocal blast strategy was employed. This is a commonly implemented technique to identify potential orthologs, in which two genes are deemed to be orthologs if they identify each other as best hits in the opposite genome.

Once homology between a set of genetic sequences is established, these genes comprise a homologous gene family. The alignment of the members of such a family (i.e. a multiple sequence alignment (MSA)) serves two important purposes. First, it reveals common features that are important for the structure and function of the set of homologs. Second, it uncovers poorly conserved regions that are less important for the underlying common function or structure, but may define specificity in each homolog. Numerous MSA algorithms exist, including those that are implemented in Clustal, Muscle and Probabilistic Alignment Kit (PRANK) (Thompson et al. (1994), Edgar (2004) and

Löytynoja and Goldman (2005), respectively). Curation of the resultant MSA is commonly carried out to resolve ambiguously aligned regions. Although manual curation is routinely performed, an automated approach such as that implemented in Gblocks (Castresana, 2000), trimAl (Capella-Gutiérrez et al., 2009) or Block Mapping and Gathering with Entropy (BMGE) (Criscuolo and Gribaldo, 2010) may be more suitable in studies that involve large amounts of sequence data.

1.3 Supermatrices

The rise in fully sequenced genomes has led to an increased number of molecular evolutionary studies that incorporate multiple homologous gene families (for example, Wildman et al. (2007), Dunn et al. (2008)). Methodologies such as the supermatrix and supertree approach combine such large amounts of data into a single framework, upon which phylogenetic hypotheses are generated. In a supertree approach, a phylogenetic tree is reconstructed for each gene family alignment and the derived source trees are merged to generate a phylogenetic hypothesis. Alternatively, a supermatrix approach may be employed, in which all of the gene family alignments (i.e. input matrices) are amalgamated into a single phylogenetic matrix and subsequently analysed simultaneously (Kluge 1989, De Queiroz and Gatesy, 2007).

There has been considerable debate over which approach is a more suitable mode of data concatenation and phylogenetic inference (for example, see Beninda-Emonds, 2004, De Queiroz and Gatesy, 2007, Von Haeseler, 2012). However, supertrees were not used throughout this research, and as such, shall not be discussed in further detail. Supermatrices use the phylogenetic information encoded in the characters more fully than supertree methods (de Queiroz and Gatesy, 2007). Furthermore, the lack of requirement for fully overlapping data sets is commonly cited as an additional advantage of supermatrices. In the event that a taxon is not present in a given source matrix, this taxon is represented in

the supermatrix by a series of question marks that extend the length of the other sequences in the source matrix. The effect of missing data on the biological hypotheses that are inferred from a supermatrix approach has been disputed (Wiens, 2003, Wiens, 2006, Sanderson et al., 2010). Regardless of these potential concerns, supermatrices have been successfully employed in answering an array of evolutionary questions, such as investigating the diversification of rodents (Fabre et al., 2012) and primates (Chatterjee et al., 2009). The supermatrix approach to phylogenetic tree reconstruction is described in Figure 1.1.

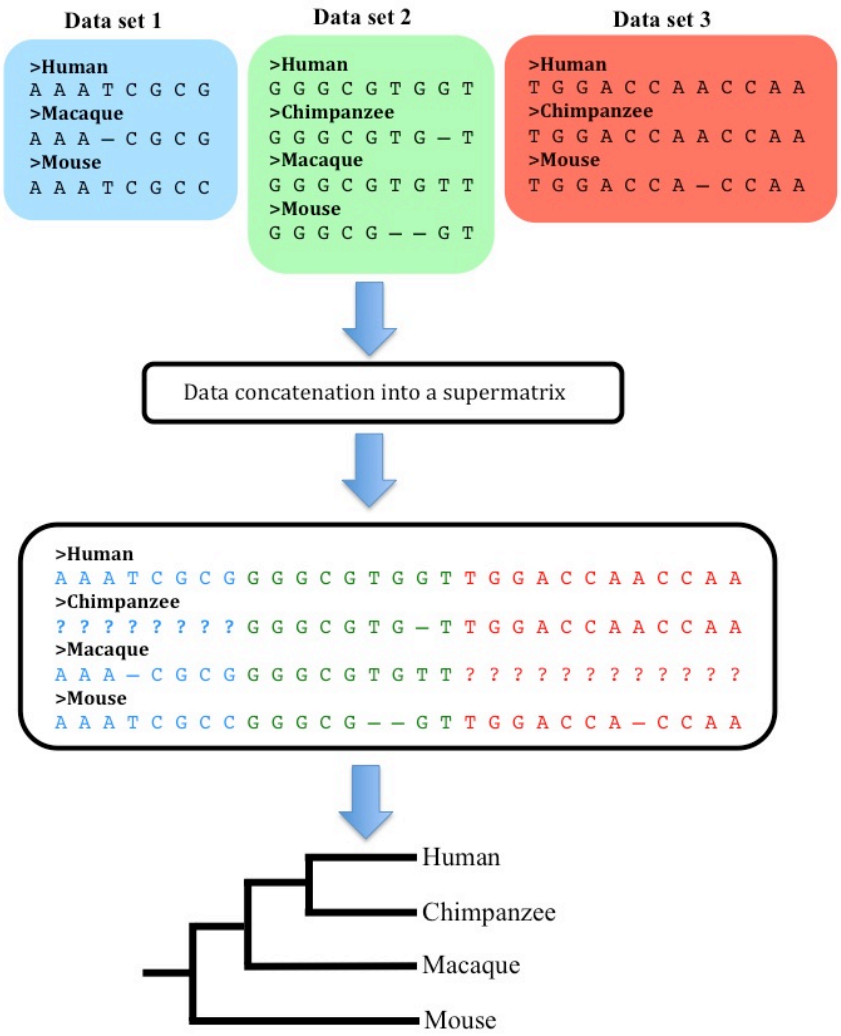


Figure 1.1 A supermatrix approach to homologous gene family concatenation.

In this example, there are three input homologous gene family alignments (labelled data set 1, 2 and 3). These families are concatenated to generate a supermatrix. The supermatrix may be used in conjunction with a phylogenetic tree reconstruction method to infer a phylogenetic hypothesis.

1.4 Phylogenetic tree reconstruction

Without a phylogenetic framework, each species would remain an independent mystery upon which a limited number of evolutionary investigations could be conducted. Although Lamarck is accredited with the publication of the first evolutionary tree (Lamarck, 1809) (Figure 1.2), it was essentially the sole figure in Darwin's renowned publication "*On the Origin of Species*" (Figure 1.3) (Darwin, 1859) that popularized the phylogenetic tree concept.

Historically, a phylogeny reconstruction was used simply to understand the pattern of common ancestry between taxa. Recently, phylogenies reconstructed using genetic sequences have addressed more diverse biological questions. For example, phylogenetic trees have been applied to studies involving human health (Worobey et al., 2004), forensics (Metzker et al., 2002) and conservation (Erwin, 1991, Vézquez and Gittleman, 1998). Specifically in an evolutionary context, phylogenetic trees have previously aided the identification of entities that are under the influence of natural selection (Suzuki et al., 2001, Kosiol et al., 2008) and the detection of orthology and paralogy events (Goodman et al., 1979, Goodstadt and Ponting, 2006, Gabaldón, 2008).

At present, there are several methods routinely used to reconstruct phylogenetic trees, each of which may be categorized as a distance matrix or character-based method. The primary difference between the two categories is centred on the treatment of the data. Distance matrix methods transform aligned sequences into a pairwise distance matrix and identify a phylogenetic hypothesis that predicts the observed set of distances a closely as possible (Felsenstein, 2004). In contrast, character-based methods consider each aligned site individually to calculate a probability for each possible hypothesis. Subsequently, the most probable phylogenetic hypothesis is identified (Felsenstein, 2004). In the following sections, both sets of methods will be discussed.

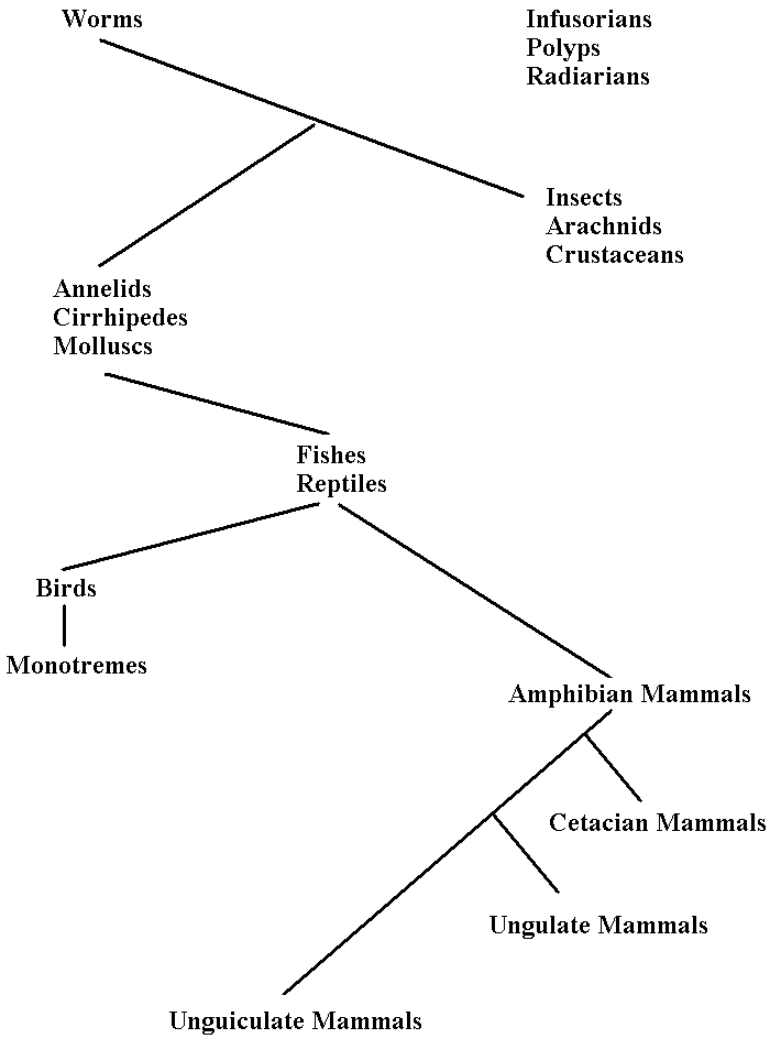


Figure 1.2 Lamarck's evolutionary tree.

This tree is considered to be the first example of an evolutionary tree. It is (an english translation of) Lamarck's interpretation of the evolution of animals (Lamarck, 1809).

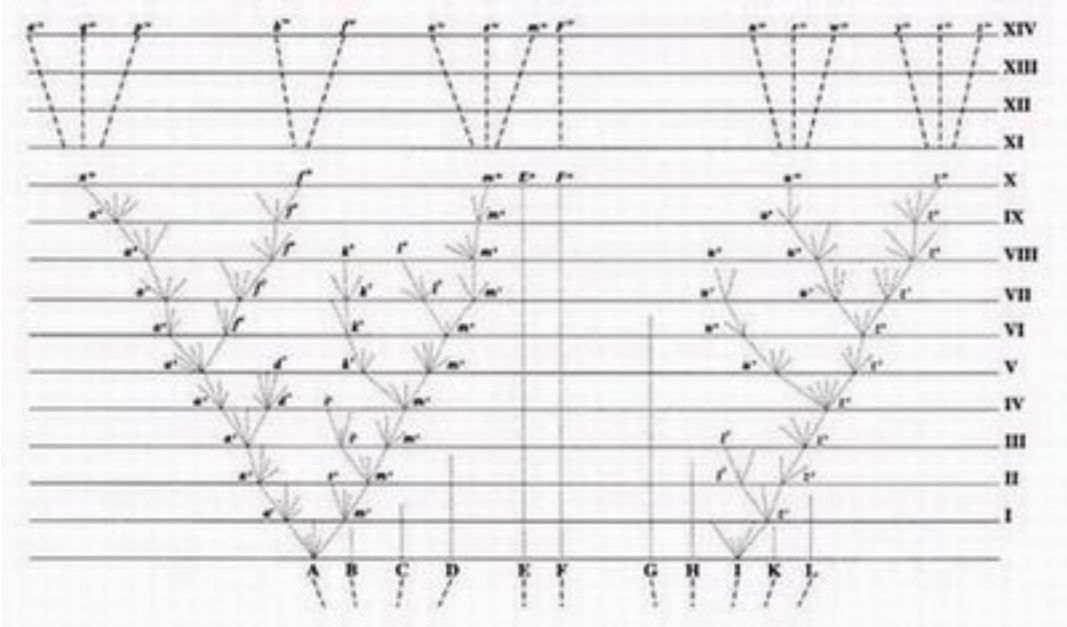


Figure 1.3 Darwin's tree of life.

This depiction of a phylogenetic tree is the sole illustration in Darwin's celebrated publication "*On the Origin of Species*" (Darwin 1859).

1.4.1 Distance matrix methods

Distance matrix (DM) methods were introduced into the field of phylogenetics independently by Cavalli-Sforza & Edwards (1967) and Fitch & Margoliash (1967). All DM methods calculate an evolutionary distance between all pairs of taxa, and identify a phylogenetic tree that predicts the observed set of distances as accurately as possible (Felsenstein, 2004). Several DM methods exist, such as the Nei's minimum evolution method (Kidd and Sgaramella-Zonta, 1971, Rzhetsky and Nei, 1992, 1993) and the commonly used neighbor joining (NJ) protocol (Saitou and Nei, 1987). NJ was the sole DM method implemented in this thesis and so shall be discussed further.

In NJ, a star-like tree topology is initially assumed for a set of sequences, in which no taxa cluster together. The sequences are converted into a distance matrix that estimates the evolutionary distance between them. Based on the computed distance matrix, a summed branch length is calculated between each pair of taxa, and the two taxa with the smallest summed branch length are selected and connected together. This pair of taxa is now considered as an operational taxonomic unit (OTU). The OTU is treated as another taxon, with the branch length of the OTU calculated as an average of the branch lengths of the initial two taxa. This calculation is repeated for each pair of taxa in the data set, until all of the internal branches have been identified (Saitou and Nei, 1987).

NJ is regularly used as a preliminary rapid tree building method to serve as a basis for other methods and has proven to be a valuable tool in phylogenetic tree reconstruction exercises that incorporate a large number of sequences (Tamura et al., 2004). However, if the error in the distance estimation is large, it can be difficult to obtain a reliable distance matrix that is the input for NJ. Obviously, if the input to the algorithm is poor, the algorithm has little chance of success (Holder and Lewis, 2003).

1.4.2 Character-based methods

1.4.3 Maximum parsimony

In a character-based phylogeny concept, maximum parsimony (MP) was proposed by Camin and Sokal (1965), who defined the term and devised the algorithm to estimate evolutionary change and construct a phylogenetic tree. The principle of this method is that precedence is given to simplicity. For each phylogenetic tree proposed, the minimum number of substitutions required to reconstruct the evolutionary history of each character site is determined. Subsequently, the hypothesis that requires the fewest changes to explain the observed topology is considered the best estimate of the phylogeny, and is described as the most parsimonious tree (Yang, 1996b).

Although the concept of maximum parsimony is straightforward, the method is criticized because of its implicit assumption that multiple substitutions at the same site are rare. Felsenstein (1978) crucially demonstrated that in some cases, parsimony is statistically inconsistent. This means that there are situations in which an incorrect conclusion will be obtained with a greater amount of confidence as the amount of data increases (discussed in section 1.6.1). Consequently, probabilistic methods (that attempt to account for multiple substitutions at the same site) have progressively supplanted MP as a method of phylogenetic tree inference in molecular sequence studies (Gadagkar and Kumar, 2005, Spencer et al., 2005).

1.4.4 Models of substitution

"*All models are wrong, but some are useful*"

– Box (1976)

Sequence divergence is linear for a limited time after a divergence event as multiple substitutions that occur at the same site eventually obscure the true evolutionary distance between two sequences. This is a particularly worrying scenario in the investigation of anciently diverged or fast-evolving sequences. Thus, models of substitution are commonly employed in phylogenetic tree reconstruction methods. These models combine the observed evolutionary distance between sequences with an estimate of the unobservable change that cannot be deterministically calculated to approximate a number of substitutions that more accurately reflects reality. Two categories of probabilistic models of substitution were used in this thesis: those that model DNA and protein sequence evolution, each of which shall be described.

1.4.5 Models of DNA sequence evolution

The most commonly implemented models of DNA sequence evolution generally differ in three parameters: (i) substitution parameters, (ii) base frequency parameters and (iii) rate heterogeneity parameters. Depending on their structure, nucleotides may be categorized as either purines (adenine and guanine) or pyrimidines (cytosine and thymine). Substitutions that exchange a purine for another purine, or a pyrimidine for another pyrimidine, are referred to as transitions. Conversely, substitutions that convert a purine to a pyrimidine (or vice versa) are described as transversions (Figure 1.4). Models of DNA sequence evolution differ by the relative rate at which transitions (α) and transversions (β) are proposed to occur. Some models, such as the Jukes-Cantor (JC) model (Jukes and Cantor, 1969) and Felsenstein (F81) model (Felsenstein, 1981) expect that transitions and transversions occur at equal rates. Other models, such as the Kimura-2-Parameter (K2P)

model (Kimura, 1980), Hasegawa et al. (HKY85) model (Hasegawa et al., 1985) and General Time Reversible (GTR) model (Tavaré, 1986) suggest that such substitutions occur at different rates, in line with previous observations that transitions occur more frequently than transversions (Brown et al., 1982, Gojobori et al., 1982) (Figure 1.5).

The second parameter that distinguishes the various models of DNA sequence evolution is the average frequency at which each nucleotide base is estimated to occur (π_x, where x is either A, C, G or T). Some models (for example, JC and K2P) assume that base composition is at equilibrium, while others (such as F81, HKY85 and GTR) posit that this may not be the case (Figure 1.5).

The last issue to consider when analyzing the history of molecular sequences is that the evolutionary rate between different sites in a sequence may vary considerably due to, for example, constraints of the genetic code or selection for gene function. Such variation is often modelled using a gamma distribution (Yang, 1996a). The shape of this distribution is controlled by a parameter, α , that specifies the range of mutation rate variation that is observed between sites. A small α value represents extreme rate variation, while a large α value indicates minor rate variation. However, it has been suggested that assuming that all sites in a sequence are free to vary may lead to an incorrect estimation of subsitutions when there are sites that do not change. This suggestion may be incorporated into a sequence evolutionary model through the designation of a proportion of the sites as invariant (or invariable), signifying that they do not vary in their rate of substitution (Hasegawa, Kishino et al. 1985).

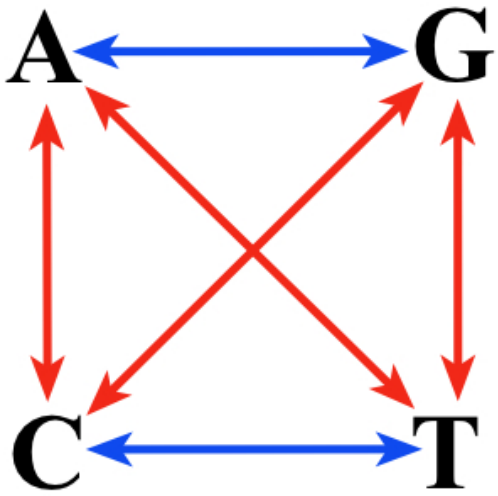


Figure 1.4 Nucleotide substitution matrix.

A and G are purine bases. C and T are pyrimidine bases. Blue arrows indicate transitions. Red arrows indicate transversions. This figure was reproduced from Page and Holmes (1998) (Chapter 5: Measuring Genetic Change, page 146).

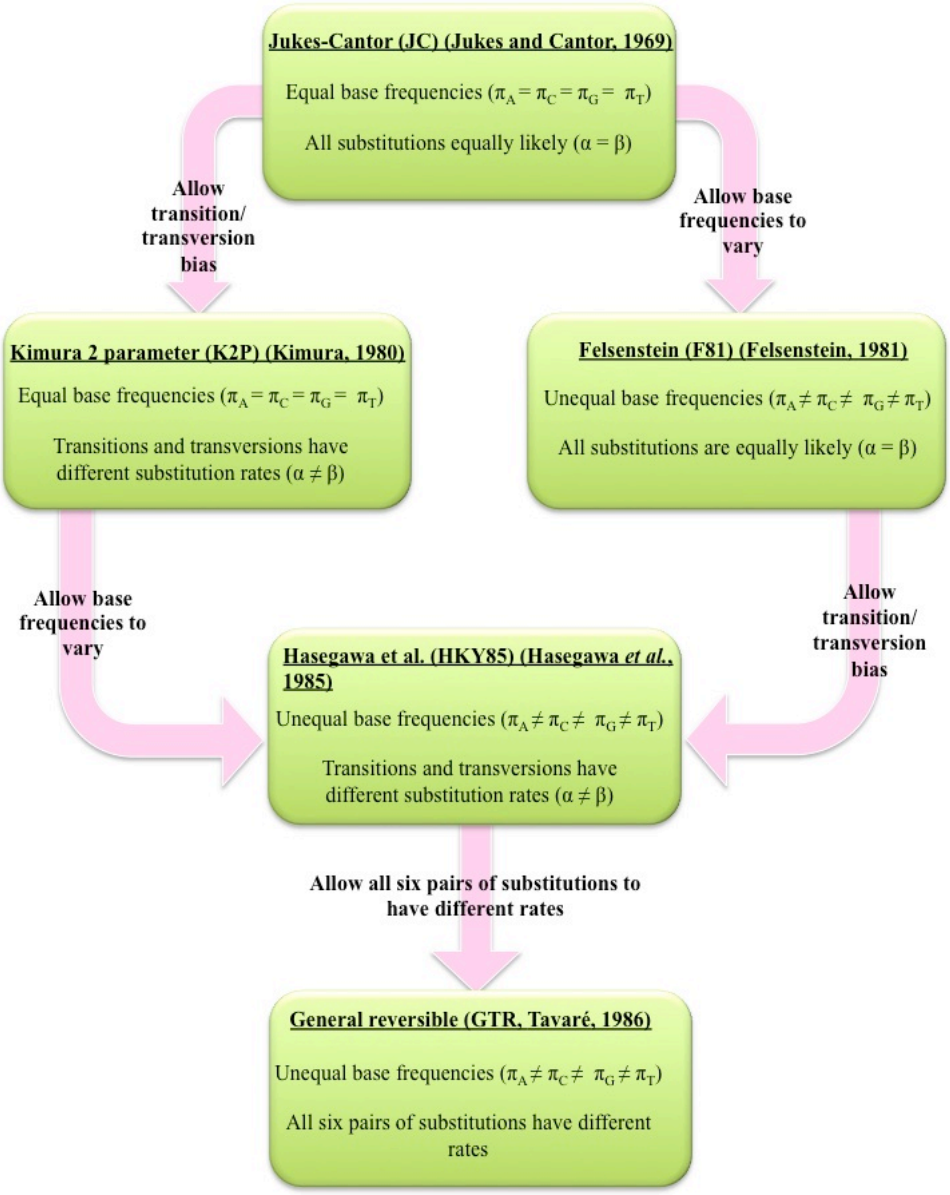


Figure 1.5 A summary of five commonly implemented DNA sequence evolution models.

The various models differ in their approach to calculating how often various substitutions are expected to occur, and to the relative frequency of each of the nucleotide bases. This figure was reproduced from Page and Holmes (1998) (Chapter 5: Measuring Genetic Change, page 153).

1.4.6 Models of protein sequence evolution

Models of protein evolution are pre-computed matrices that describe the probability that one amino acid changes to any other amino acid. Dayhoff and colleagues pioneered a counting method to generate point accepted mutation (PAM) matrices from a limited amount of protein sequence data that was available at the time (Dayhoff and Eck, 1968, Dayhoff et al., 1972). Subsequently, Jones et al. (1992) applied Dayhoff's methodology to produce a replacement matrix from a much larger database (the JTT matrix). However, these two counting methods effectively employ parsimony to estimate amino acid replacement matrices. In order to remove the potentially problematic parsimonous aspect of calculating substitution probabilities, Whelan and Goldman (2001) applied a likelihood framework to generate the WAG matrix that more effectively handles multiple substitutions at the same site.

1.4.7 Model selection

Model misspecification under- or over-estimates the magnitude of substitution that has occurred in a set of aligned sequences, in turn affecting the outcome of a phylogenetic analysis. For example, an inappropriate model may influence branch length estimation and bias statistical support values (Buckley et al., 2001, Buckley and Cunningham, 2002). As such, it is important to carefully select a model of substitution that adequately captures the underlying evolutionary process in a statistically rigorous manner. There are different criteria available to aid model selection, three of which were used throughout this thesis: Akaike Information Criterion (AIC) (Akaike, 1974), Bayesian Information Criterion (BIC) (Schwarz, 1978) and Bayes Factors (BF) (Kass and Raftery, 1995).

AIC is a popular model selection criterion measured using the following formula:

$$AIC = 2k - 2\ln(L) \qquad [1.1]$$

where k is the number of model parameters, and L is the maximum likelihood value. As AIC penalizes excessive parameter use, while rewarding a well fitting model, the optimal model is that which obtains the lowest AIC score. Alternatively, assessment of model suitability may be facilitated in a Bayesian framework through the implementation of the Bayesian Information Criterion. BIC is similar to AIC, with the exception that the penalty term that is included for each additional parameter introduced by BIC is larger than in AIC. Although it is debated which of the two aforementioned criteria is more suited as a selection criterion (Posada and Buckley, 2004, Alfaro and Huelsenbeck, 2006, Ripplinger and Sullivan, 2008), both measures are implemented in commonly used model selection software programs, such as ModelGenerator (Keane et al., 2006) and ProtTest (Abascal et al., 2005). Finally, BF represent the probability of the data given a null hypothesis, over the probability of the data given an alternative hypothesis (Goodman, 1999). The BF returned when two hypotheses are compared is generally interpreted according to the table of Kass and Raftery (1995) (Table 1.1).

Table 1.1 Interpretation of Bayes Factors (redrawn from Kass and Raftery, 1995).

Bayes Factors Interpretation		
Log_{10}(Bayes Factor)	**Bayes Factor**	**Evidence against null hypothesis**
0 – 0.5	1 – 3.2	Barely worth a mention
0.5 – 1	3.2 – 10	Substantial
1 – 2	10 – 100	Strong
> 2	> 100	Decisive

1.4.8 Maximum likelihood

The concept of maximum likelihood (ML) is accredited to the statistician R. A. Fisher (Fisher, 1912, Fisher, 1922). However, his research generally focussed on quantitative genetics. It was Edwards and Cavalli-Sforza (1964) who subsequently suggested that maximum likelihood methods could have a phylogenetic application. After a number of progressions, Felsenstein (1981) adapted the method for DNA sequence data and developed the "pruning algorithm" that allowed ML principles to be carried out on a realistic number of sequences. ML may be understood as a parametric approach to tree building that estimates the probability of observing data (*d*) (i.e an aligned set of genetic sequences), given a hypothesis (a tree (τ) and a substitution model (θ)). It may be defined mathematically as:

$$L(\tau, \theta) = Pr(d \mid \tau, \theta) \qquad [1.2]$$

For each site in an alignment (the data), given a substitution model and a phylogenetic tree (the hypothesis), a likelihood is calculated. The product of these site likelihoods provides the total likelihood for the considered hypothesis. The optimal tree is that which returns the highest likelihood (that is, the lowest negative log-transformed likelihood) (Felsenstein, 1981). ML estimates are generally deemed to be more robust to systematic error and model violation than other methods, particularly parsimony (Huelsenbeck 1995, Whelan et al., 2001, although see Warnow (2012)). However, multiple equally good answers render it difficult to guarantee that the ML tree is actually maximal. In addition, for ML, each possible hypothesis be assessed individually, which is technically prohibitive to calculate for any reasonable number of taxa (Huelsenbeck et al., 2001, Delsuc et al., 2005). For example, with a 5-taxa data set, there are 15 possible unrooted

trees, However, a 10-taxa data set results in an exponential rise to 2,027,025 possible unrooted trees. This issue may be circumvented through the use of heuristics that reduce tree search space. In these heuristics, a starting tree is progressively altered to maximise the likelihood function. There are a number of heuristic tree rearrangement algorithms. For example, nearest neigbour interchange (NNI) invokes the exchange of two neighbouring branches, while subtree pruning and regrafting (SPR) involves the detachment of a subtree from an existing tree, followed by a re-attachment of the subtree to a different part of the existing tree (for a review, see Felsenstein, 2004). ML remains an incredibly popular method of phylogenetic tree inference at present. For example, Whelan and Goldman (2001) used ML to estimate the WAG model of amino acid substitution (see section 1.4.6), and Perelman et al. (2011) implemented an ML approach to understand the divergence of primates.

1.4.9 Bayesian inference

Bayesian inference is closely allied to ML as both are character-based methods of phylogenetic tree reconstruction that incorporate probabilistic models of sequence evolution. A fundamental conceptual difference between the two is that Bayesian inference incorporates prior knowledge into its calculation. Bayesian phylogenetics is centred on the posterior probability (PP) of a hypothesis. This may be defined using Bayes' theorem:

$$\Pr(H|D) = \frac{\Pr(H) \times \Pr(D|H)}{\Pr(D)} \qquad [1.3]$$

This equation calculates the PP of hypothesis *H* (i.e. a phylogenetic tree) given data *D* (i.e. an aligned set of genetic sequences), a substitution model and a prior probability distribution for the set of all available alternative hypotheses (in this case, all the possible

phylogenetic trees). The denominator is a normalising constant that is computationally intensive to calculate as it requires a summation of the likelihoods of all the possible hypotheses (i.e. trees) (Yang and Rannala, 1997). However, the PP of a tree may be approximated using Markov Chain Monte Carlo (MCMC) methods that consider a random sample from the posterior distribution, providing an approximation for the true posterior probability (Yang and Rannala, 1997, Huelsenbeck et al., 2001, Felsenstein, 2004).

In phylogenetics, the most commonly used MCMC method is the Metropolis-Hastings algorithm (Metropolis et al., 1953, Hastings, 1970). In this approach, the PP of an arbitrary starting tree is calculated, which shall be defined as the "current tree". This tree is randomly perturbed, producing a "new tree", whose PP is also calculated. The Metropolis-Hastings algorithm decides whether to accept or reject the newly proposed tree as the next state in the Markov chain by calculating the height ratio of the PP of the two trees. Through repetition of tree proposal, acceptance and rejection, a Markov chain of trees is created. For a properly constructed and adequately run Markov chain, the proportion of time that any tree is visited in the chain is a valid approximation of its PP (Tierney, 1994). Thus, the chain tends to remain in regions of high PP. The algorithm may be terminated when multiple independent Markov chains converge (i.e. exhibit similar posterior distributions).

Bayesian inference may be superior to ML as it seeks to glean information about the shape of the PP landscape rather than locating the global maximum. This means that the proposed tree is less likely to be a local maximum in tree space (Huelsenbeck et al., 2001). Recently, the Bayesian approach has become a popular method of phylogenetic inference that has been implemented in a diverse array of studies, such as the inference and evaluation of uncertainty in phylogenies (Huelsenbeck et al., 2000; Murphy et al., 2001), and in the estimation of divergence times (Rota-Stabelli et al., 2011).

1.5 Assessment of confidence in phylogenetic inference

It is important to assess the level of confidence in an inferred phylogenetic hypothesis, to allow for comparison between different topologies and methodologies. There are a number of ways to determine the robustness of an inferred tree, some of which shall be described.

Bootstrapping is a non-parametric statistical technique that was originally devised by Efron (1979) but formally proposed as a method of obtaining confidence limits of phylogenies by Felsenstein (1985). It currently represents one of the most common forms of confidence interval inference on a phylogenetic hypothesis. For example, of all the publications in the journal "Systematic Biology" in 2001, at least 50% of those presented phylogenetic analyses, and all of the studies that reconstructed phylogenies used the bootstrap method to measure nodal support (Soltis and Soltis, 2003). In this technique, a data matrix (i.e. an alignment) is randomly re-sampled with replacement multiple times to produce pseudo-replicate data matrices of equal length. Phylogenetic trees are generated for each pseudo-replicate data matrix, and the resultant trees are summarized into a single tree using a majority-rule component consensus approach (Margush and McMorris, 1981). The support value on each node of a bootstrapped tree is indicitive of the proportion of times that a given clade is found in the replicate data sets. It is important to clarify that bootstrapping measures the repeatability of a data set, rather than accuracy of the inferred phylogeny.

One of the most popular methods for assessing tree incongruence is the Approximately Unbiased (AU) topology test (Shimodaira, 2002), which was implemented in chapter 2 of this thesis. This multiscale bootstrap procedure is designed to assess whether a given tree is a significantly more likely hypothesis than another tree. Several sets of bootstrap replicates are generated by changing the sequence length, which may differ

from that of the original data. The number of times that a hypothesis is supported by each of the replicates is counted for each set to obtain bootstrap probability values for different sequence lengths. The AU-test calculates the approximately unbiased P-value from the change in bootstrap probability values along the changing sequence length.

1.6 Sources of phylogenetic error

There are two principal categories of phylogenetic error: stochastic (sampling) error, and systematic error. Stochastic error constitutes one of the major limitations of phylogenetic analyses that are based on single or few genes. As the number of positions in a single gene is usually quite low, random noise may lead to poor resolution of the phylogenetic hypothesis under consideration (Philippe et al., 2005). For example, Rokas et al. (2003) investigated phylogenies that were obtained from 106 orthologous genes belonging to eight yeast species. Once a phylogenetic tree was reconstructed for each homologous gene family individually, more than twenty alternative topologies were generated. Subsequent concatenation of all genes in the data set produced a robust phylogeny that significantly rejected all other potential phylogenies obtained in the gene-scale analyses. The increased resolution provided by augmented gene sampling has led to the optimistic view that using genomic data will diminish the stochastic error that is often observed in analyses that focus on a single or few genes (Gee, 2003, Jeffroy et al., 2006).

Systematic errors occur when a reconstruction method arrives upon an incorrect solution with stronger support as the amount of data considered increases (Philippe et al., 2005). Such inconsistencies may result in alternative conflicting phylogenies or absolute clade support for incorrect topologies. For instance, let us refer to the Rokas et al. (2003) study described in the previous paragraph. The concatenation of data was an integral component in the reconstruction of a robust phylogeny. However, Phillips et al. (2004) demonstrated that the data set contained non-phylogenetic signal. Dependent upon the

optimality criterion or model assumptions, mutually incongruent, yet 100% supported phylogenetic trees, could be obtained. Three common causes of systematic inconsistencies are long branch attraction, compositional attraction and heterotachy, each of which shall be described.

1.6.1 Long branch attraction

Felsenstein (1978) crucially observed that parsimony might be statistically inconsistent, as rapidly evolving lineages would be inferred as closely related, regardless of their true relationships. It is possible that two long branch sister taxa separated by a short internode will acquire more identical bases by chance than the few number of inherited changes on the short internode that groups the long branch with its true relative. In this case, the most parsimonious solution would be to erroneously group the two longer branches as sister taxa, resulting in long branch attraction (LBA).

Although originally observed in parsimony analyses, LBA has subsequently been demonstrated to affect other tree reconstruction methods, if the proposed model of substitution is strongly violated (Swofford, 2001, Lemonn and Moriarty, 2004 Bergsten, 2005). However, probabilistic methods that incorporate an estimate of the expected amount of change along each branch of the tree are generally thought to be more robust to the effects of LBA (Kuhner and Felsenstein, 1994, Swofford et al., 2001; Bergsten, 2005).

There has been an array of methods proposed to diminish the impact that LBA may have on a phylogenetic inference. In some situations, LBA may be ameliorated through increased taxonomic sampling (Pollock et al., 2002, Poe, 2003, Heath et al., 2008). Alternatively, the effect of LBA may be reduced through the elimination of fast-evolving sites that can be saturated by multiple substitutions (Brinkmann and Philippe, 1999, Cummins and McInerney, 2011). Finally, selection of an appropriate outgroup may be critical in avoiding LBA because in situations where the selected outgroup is extremely

divergent, a fast evolving ingroup taxon may be artifactually attracted to the long branch of the outgroup (Bergsten, 2005). However, often the simple lack of suitable outgroup availability is a limiting factor.

1.6.2 Compositional attraction

Genetic sequences tend to exhibit compositional bias that causes sequences to be erroneously grouped together based upon the similarity of their nucleotide or amino acid composition. Although originally suggested to be an issue confined to nucleotide-based phylogenies (Loomis and Smith, 1990), compositional bias was later demonstrated to also occur in phylogenies that were derived with amino acid sequences (Foster et al., 1997, Chang and Campbell, 2000).

Common approaches to alleviate the impact of compositional bias include recoding, RY coding in the case of nucleotides (in which purines are recoded as "R" and pyrimidines as "Y") (Woese, 1991) or Dayhoff recoding in the case of amino acids (in which amino acids are recoded into various functional categories) (Hrdy et al., 2004). Alternatively, the use of heterogeneous models that account for compositional variation throughout the tree are becomingly increasingly popular (for example, Foster et al. (2009)).

1.6.3 Heterotachy

Heterotachy describes the situation in which the evolutionary rate of a site varies with time due to changing functional or structural constraints (Philippe and Lopez, 2001). As a result, fast-evolving sites may become slowly evolving (or vice versa) in different lineages. Heterotachy leaves no observable signature on genetic sequences, making it extremely difficult to identify (Kolaczkowski and Thornton, 2004). Recently, some approaches have been developed to detect heterotachy. For example, a non-homogeneous gamma model was suggested to account for variability in site rates over time (Fitch and

Markowitz, 1970, Kolaczkowsi and Thornton, 2004). More recently, Whelan et al. (2011) developed a likelihood ratio test based approach that allows substitution rates to vary independently among the branches of the tree.

1.7 The subphylum Vertebrata

Vertebrata is a diverse subphylum of Phylum Chordata, comprising of approximately 63,000 named species (Hoffmann et al., 2010) that is further divided into seven distinct phylogenetic classes. These are Agnatha, Chondrichthyes, Osteichthyes, Amphibia, Aves, Reptilia and Mammalia (Figure 1.6). For reasons discussed in the following sections, most of the vertebrate species that have been sequenced to date are belong to the Class Mammalia. Thus, Mammalia is the only vertebrate class that will be discussed in further detail.

1.7.1 The Class Mammalia

At present, it is estimated that there are 5,676 extant mammalian species and approximately 4,000 fossil genera (Wilson, 2005) that are categorized as Protheria (monotremes) and Theria. The latter is composed of the infraclasses Metatheria (marsupials) and Eutheria (placental mammals).

The radiation of mammals is a richly documented transition that is continuously re-written by key fossil discoveries and increasingly comprehensive phylogenies. In 2002, *Eomaia scansoria* was discovered and described as an early Eutherian that lived approximately 125 million years ago (Ji et al., 2002). More recently, the oldest known placental mammal fossil, *Juramaia sinensis*, was estimated to have lived approximately 160 million years ago, extending the first appearence of the Eutherian clade by 35 MY (Luo et al., 2011). In addition, this new fossil corroborates with marsupial-eutherian

divergences recently estimated to have occured 143 – 193 MYA using molecular data (van Rheede et al., 2006, Bininda-Emonds et al., 2008, Phillips et al., 2009).

Placental mammals represent the most diverse of the three extant mammalian lineages (Novacek, 1992, McKenna et al., 1997). A number of studies suggest that morphological data alone are unreliable for resolving the affinities of this particular clade (Springer and Murphy, 2007, Lee and Camens, 2009). It has been the advent of molecular phylogenetics, sequencing of the human genome (Lander et al., 2001, Venter et al., 2001), and subsequent sequencing of other mammalian genomes that has greatly improved our understanding of Eutherian phylogeny.

Currently, the Ensembl gene sequence database (Hubbard et al., 2002) includes more mammalian genomes than other vertebrate classes, in spite of the fact that there are approximately 5,676 named mammalian species, almost double the number of bird species (10,027) and six times as many fish species (31,327) (Hoffman et al., 2010). Why are vertebrate genome sequencing projects biased towards mammals? Incentives for sequencing mammalian genomes mirror historical motives for model organism selection. For example, the understanding and treatment of complex disorders such as diabetes, hypertension and obesity has been advanced considerably through the use of various mammalian models (Truett et al. (1991), Jacob and Kwitek (2002), Farrer (2006)). The sequencing of such organisms may provide further insight into genetic underlying causes of common medical conditions (for example, Stoll, 2000).

A still unfulfilled promise of comparative biology is a unified view of the origin and evolutionary divergence of mammalian species. Until recently, paleontological, morphological, and small amounts of molecular sequence data struggled to identify and date ancient mammalian speciation events. Thus, mammalian taxa genome comparison provides incredibly powerful tools for a deeper understanding of phylogenetic divergences, general trends in genomic complexity evolution and lineage specific adaptations.

The realization of the benefits of sequencing non-human mammalian genomes such as those described above motivated the initiation of the "Mammalian Genome Project" (http://www.broadinstitute.org/scientific-community/science/projects/mammals-models/mammalian-genome-project), aimed at sequencing multiple placental mammalian genomes at "low coverage". To understand the phrase "low coverage" in a genome sequencing context, and its implications for molecular evolutionary analyses, a short summary of genome sequencing shall be provided.

1.7.2 Vertebrate genome sequencing

In 1977, Sanger et al. published two methodological papers regarding the rapid determination of DNA sequences that provided opportunites to examine molecular evolutionary hypotheses from a whole new viewpoint (Sanger et al., 1977a, Sanger et al., 1977b). In the last decade, a new wave of cost-effective sequencing technologies have been introduced (for a review of the various sequencing strategies, see Mardis (2008), Schuster (2008), Shendure and Ji (2008), Ansorge (2009), Metzker, (2009)). Essentially, to sequence a genome, DNA is fragmented into short sections. The base pairs composing these fragments are obtained. Through repeated rounds of the base decipherication stage, multiple overlapping reads of the bases composing each DNA fragment are retrieved. The coverage of a genome, such as "2×" refers to the number of overlapping sequences that built a region of gene assembly.

There are inherent limitations associated with genomes that are sequenced at low coverage. For example, all of the usual genome sequencing issues (such as missing sequence, sequence fragmentation and inaccurate insertions/deletions/substitutions) will be exasperated in a low-coverage genome (Green, 2007, Milinkovitch, 2010). Thus, some argue that in spite of the additional data that low-coverage genome sequencing provides, it

will remain difficult to truly differentiate artifacts from evolutionary events until improved homogeneity in both taxon sampling and sequencing coverage (Milinkovitch et al., 2010).

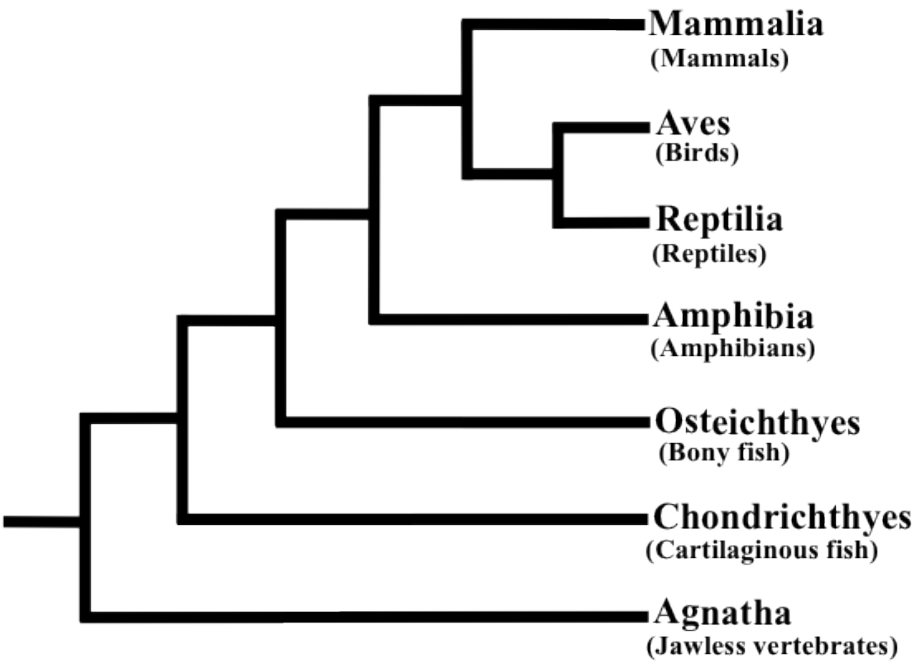


Figure 1.6 Phylogeny of Phylum Vertebrata.

Figure re-drawn and adapted from Kasahara (2007).

1.8 Aims of thesis

Monumentous technological and conceptual advances coupled with a remarkable rise in genomic-scale data availability allow a diverse array of evolutionary hypotheses to be examined from a molecular perspective. The ultimate aim of this thesis is to exploit such recent progress to address various questions pertaining to vertebrate evolution.

In chapter 2, I will describe networks (specifically protein-protein interaction networks), gene duplication and the mammalian order Primata. Then, I will present an investigation that was conducted in collaboration with Dr. David Alvarez-Ponce to assess the relationship between primate protein-protein interaction network structure and gene duplicability.

Chapter 3 commences with a discussion of molecular dating and the mammalian taxonomic family Ursidae. Until recently, the date of divergence between two members of this clade (polar bear and giant panda) has been examined using either mitochondrial data or a small amount of nuclear data. BGI Shenzhen, China provided an unpublished polar bear genome that allowed the speciation event between giant panda and polar bear to be addressed using the largest amount of nuclear data to date in a Bayesian framework.

In chapter 4, I will describe the forces that govern synonymous codon usage bias. Then, I will examine the evidence that vertebrate synonymous codon usage is maintained by a balance between neutral and selective processes, using novel gene expression data and codon usage bias index. The thesis closes with some concluding remarks in chapter 5.

Chapter 2: The dynamic relationship between gene duplicability and network structure in primates.

2.1 Introduction

One of the key insights provided by fully sequenced genomes is the pervasiveness of gene duplication and loss in organisms (Ohno, 1970, Zhang, 2003). However, not all genes are equally likely to duplicate. While some gene families are represented by dozens of members in a given genome, others remain as singletons over time. This observation naturally leads to the question as to what constrains gene duplication.

There are a number of biological factors known to correlate with gene duplicability. One such factor is the position of a gene's encoded product in a protein-protein interaction network (PIN). This chapter commences with an introduction to gene duplication, networks and primates. Subsequently, a study is described that explored if the structure of the primate PIN constrains the duplicability of its encoding genes from two distinct perspectives. First, it is known that the relationship between gene duplicability and network centrality is not universal. In some species, such as *Saccharomyces cerevisiae,* non-duplicated genes tend to occupy more central PIN positions, whereas in humans, the opposite trend is true (Prachumwat and Li, 2006, Liang and Li, 2007). This led us to question the relationship between gene duplicability and protein centrality over the course of primate evolution. Second, it has recently been demonstrated that physically interacting proteins exhibit similar evolutionary histories, such as similar rates of evolution (for a review, see Lovell and Robertson (2010)). However, it is less clear whether the phylogenetic tree topologies of interacting proteins are more similar than expected at random. Fryxell (1996) argued that this might be the case, due to co-duplication. Although a number of examples of correlated tree topologies have been reported (for example, Fryxell, 1996, Koretke, 2000), an analysis at the level of the entire primate interactome is

yet to be conducted. Thus, our second null hypothesis stated that the phylogenetic trees of physically interacting proteins are no more similar than expected at random, and that there is no evidence for widespread co-duplication in primates.

This study was conducted in collaboration with Dr. David Alvarez-Ponce. I assembled a primate data set, reconstructed homologous gene family phylogenetic trees and inferred the duplication events that occurred throughout primate evolution. In addition, I mapped the inferred duplications to the human interactome and conducted the biological enrichment, age and suspected pseudogene content analyses. Finally, I compared the phylogenetic tree topologies of physically interacting proteins (in the actual network, and in the randomized networks). Dr. Alvarez-Ponce assembled the PIN, sub-networks and randomized networks. He calculated the network centralities and determined the proportion of duplicated genes that interact in the network. Results presented in this chapter have been published in the journal *Molecular Biology and Evolution* (see Doherty et al., 2012; see also the Publication section of this thesis).

2.1.1 The evolutionary significance of gene duplication in vertebrates

A number of early geneticists recognized the potential evolutionary significance of gene duplication (reviewed in Taylor and Raes, 2004). However, it was ultimately one highly prescient publication that explicitly proposed gene duplication as a significant factor in organismal diversification (Ohno, 1970). Since this time, whole genome sequence analyses have confirmed the existence of paralogous genes in species belonging to all three domains of life (Zhang, 2003).

There are a number of molecular mechanisms that generate gene duplications (reviewed in Hahn, 2009). For example, unequal crossing over is an outcome of homologous recombination between sequences. Retroposition describes the integration of reverse transcribed mature RNA at random sites in the genome. Finally, a whole genome duplication (WGD) generates a duplicate for every gene in the genome, that are differentially preserved and lost based on various biological factors (Maere et al., 2005, Hakes et al., 2007, Amoutzias and Van De Peer, 2010). It is generally hypothesized that vertebrate genomes were shaped by two rounds of WGD (the "2R hypothesis") (Panopoulou and Poustka (2005), Kasahara (2007)).

Exploring patterns of gene duplication is often key to understanding the origin and evolution of important vertebrate traits. For example, the acquisition of vertebrate colour vision is the result of visual pigment gene duplication (Yokoyama, 1994). Furthermore, differential duplication of salivary amylase genes among different human populations and primate species is correlated with the level of starch consumed in the diet (Perry et al., 2007). Lastly, gene duplication has been implicated in the pathogenesis of various common diseases (Zhang et al., 2009).

2.1.2 Gene duplicability

Not every gene that duplicates in a genome becomes fixed in a population, as a duplicate is likely to be lost unless it offers a selective advantage to the organism in which it is found (Ohno, 1970, Lipinski et al., 2011). Gene duplicability defines the propensity to retain multiple gene copies in a genome. Over the last decade, some of the factors that correlate with gene duplicability have been identified. For example, in yeast, a higher duplicability is observed in proteins that are exposed to the extracellular environment than for those that are localized to intracellular components (Prachumwat and Li, 2006). This has been attributed to the fact that yeast inhabit a wide range of biological niches, so genetic diversity for proteins that interact with the external environment may confer benefits to the organism. Other biological factors that correlate with gene duplicability include gene function (Marland et al., 2004, Conant and Wagner, 2002), complexity of the encoded proteins (Papp et al., 2003) and timing of expression during development (Castillo-Davis and Hartl, 2002, Yang and Li, 2004). In the investigation described in this chapter, we focused on the relationship between gene duplicability and the structure of the primate protein-protein interaction network.

2.1.3 Protein-protein interaction networks

It is clear that most entities in the biosphere exist as components of complex pathways and networks. In a network, each vertex or point is defined as a “node”, with nodes connected to one another via “edges” (Freeman, 1971) (Figure 2.1). Many different biological networks exist, including those that describe protein interactions and metabolic events. PINs are undirected networks in which it is customary to consider the nodes as proteins, and the edges as physical interactions between proteins. PINs have previously provided insights into cell robustness to perturbation, protein function and the molecular

basis of disease (Brun et al., 2003, Han et al., 2004, Kim et al., 2006, Ideker and Sharan, 2008).

As described in the introductory paragraph to this chapter, the aim of this investigation is to understand the influence that a position of a protein in a PIN has on the duplicability of its encoding gene. This aim was divided into two distinct questions. First, we examined the relationship between gene duplicability and network centrality in primates. Then, we asked if the phylogenetic trees of physically interacting proteins were more similar than expected by chance, and if such similarity could be attributed to co-duplication. The reasoning behind each of these questions shall be described separately in the following sections.

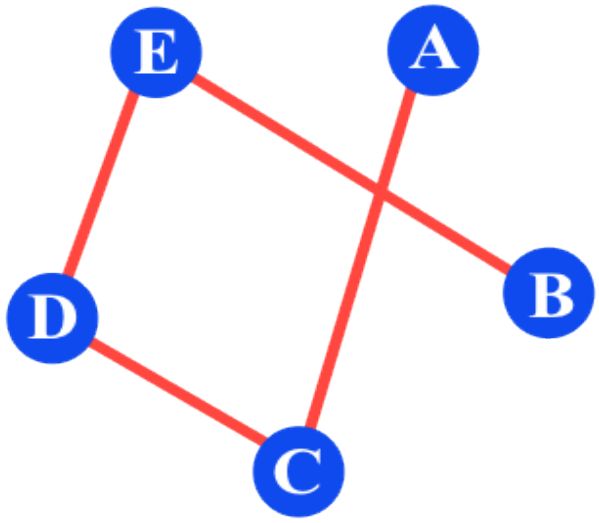


Figure 2.1 A network.

Each node (A, B, C, D and E) is blue and each edge is red. In a protein-protein interaction network, each node is a protein and each edge is a physical interaction between two proteins.

2.1.4 Analysis of the relationship between gene duplicability and network centrality in primates

The position of a protein in a network may be described by various characteristics, or "network centralities". Three of the most common network centralities are degree, betweenness and closeness. Node degree indicates how many links a node has to other nodes. For example, in Figure 2.1, node "A" has a degree 1, while node "E" has a degree 2. Betweenness is a measure of the number of shortest paths between protein pairs to which a protein belongs. Vertices that have a high probability of occurring on a randomly selected shortest path between two randomly selected nodes have a high betweenness. Finally, closeness is defined as the inverse of the sum of its distances to all other nodes. The more central a node is, the lower its total distance to all other nodes (Freeman, 1979).

Some aspects of gene evolution have been demonstrated to be affected by the centrality of their encoded products in the PIN. For instance, genes that occupy more central network positions tend to evolve more slowly (Fraser et al., 2002, Hahn and Kern, 2005). Although gene duplicability has also been demonstrated to be affected by the protein network centrality, this relationship has not remained constant over the evolution of eukaryotes. For example, in *C. elegans* and *S. cerevisiae*, singleton (i.e. unduplicated) genes tend to occupy more central positions in the network than duplicated genes (Hughes and Friedman, 2005, Prachumwat and Li, 2006). This is possibly because duplication of a gene may disrupt the dosage balance of the interactions in which it is involved, which may be more detrimental for the most highly connected genes (Papp et al., 2003). Conversely, duplicated genes tend to be more central than singleton genes in *Homo sapiens* (Liang and Li, 2007). Although this is known to be a derived character resulting from the high duplicability of metazoan-specific genes (D'antonio and Ciccarelli, 2011), it remains unclear why a different pattern is observed in *H. sapiens*. Liang and Li (2007) suggested that perhaps in mammals, a high connectivity of duplicated genes might confer a greater

chance of functional diversification (e.g. tissue specialization). Functional diversification may not be a major factor in yeast, due to the simplicity of such single-celled organisms. The contrasting observations between human and other species indicates that the relationship between network position and gene duplicability is dynamic. Thus, the first section of this investigation analyzed the relationship between centrality and duplication throughout the evolution of primates.

2.1.5 Comparison of phylogenetic tree topologies between physically interacting proteins

It has been demonstrated that genes encoding interacting proteins tend to co-evolve (for a review, see Lovell and Robertson, 2010). Co-evolution describes the process in which a change in one entity establishes a selective pressure for a change in another entity (Fraser et al., 2004). For example, interacting genes manifest more similar branch lengths in their phylogenies than would be expected from a random network (Fraser et al., 2004, Li and Rodrigo, 2009). Such similarities in branch lengths have usually been assessed using the mirrortree approach. In this approach, a distance matrix is constructed from a MSA for each homologous family of an interacting pair of proteins. The similarity between the distance matrices is calculated as a correlation coefficient (Goh et al., 2000, Pazos and Valencia, 2001, Pazos and Valencia, 2008).

It is less clear if phylogenetic trees inferred from interacting proteins are topologically more similar than expected at random. A number of correlated tree topologies among interacting genes have been reported (Fryxell, 1996, Koretke et al., 2000). However, on a whole-interactome level, Kelly and Stumpf (2010) found negligible evidence for an increased topological similarity between the trees of interacting proteins in yeast orthologous sequences. Unfortunately, both the mirrortree method, and the 1:1

ortholog approach undertaken by Kelly and Stumpf, does not address potential topological similarity between interacting phylogenetic trees that may be attributed to co-duplication.

Almost two decades ago, Fryxell hypothesized that the phylogenetic tree topologies belonging to physically interacting proteins may be similar due to co-duplication. This could occur because the successful duplication and divergence of one gene may alter the selective environment, thus facilitating the duplication and divergence of functionally interacting genes (Fryxell et al., 1996). Alternatively, the duplication of a gene may be deleterious unless its interacting partner duplicates at a similar time. For example, the duplication of a gene without its interacting partner may disrupt the balanced concentration of subcomponents in a protein complex. In turn, an imbalance of complex subcomponents may disrupt protein binding or form toxic precipitates. However, co-duplication on a whole interactome-scale is yet to be detected in primates. Thus, the second null hypothesis of this investigation stated that the phylogenetic trees of physically interacting proteins are no more similar than expected by chance, and that there is no evidence for co-duplication in primates.

2.1.6 Primates

With the exception of the outgroup, all of the species used in this investigation belong to the Order Primata. This order represents an interesting clade in the Class Mammalia, primarily due to their biomedical relevance and their phylogenetic position with respect to humans. Comparative primate genomic analyses have provided novel insights into various aspects of primate evolution (for a review, see Marques-Bonet et al., 2009). In particular, primates represent an interesting clade in which to examine the relationship between gene duplicability and network position of the encoded proteins for two reasons. First, primates exhibit an increased rate of gene duplication, and a decreased rate of nucleotide substitution compared to other mammals (Yi et al., 2002, Hahn et al.,

2007, Steiper and Seiffert, 2012). Incorporating interactome and gene duplication data may provide some interesting insights into primate evolution. Second, the human PIN is the highest-quality mammalian PIN currently available. Due to the scarcity of interactomic data, particularly for non-model organisms, it is commonplace to transfer interactions that occur in one organism to another organism (for example, Matthews et al. (2001), Huang et al. (2007) and Wiles et al. (2010) have transferred interaction data between *H. sapiens*, *S. cerevisiae* and *C. elegans*). The investigation described in this chapter assumes it is realistic to transfer such interaction data from humans to a set of closely related species, such as other primates. Taken together, primates represent a natural choice of data set for this study.

2.2 Materials and Methods

2.2.1 Genomic data collection and data set assembly

A data set consisting of six primates (human, chimpanzee, gorilla, orangutan, macaque and marmoset) and one rodent (mouse) was assembled. The phylogeny of these species may be found in Figure 2.2. For each taxon, the longest canonical transcript for all the protein-coding genes was retrieved from the Ensembl database (version 61; Feburary 2011) (Flicek et al., 2011). 1,906 human genes and 5 mouse genes that had not been assigned to an Ensembl gene family were removed. The initial data set consisted of 147,686 genes separated into 27,167 gene families. 571 sequences that were unlikely to encode functional proteins as their coding sequence was interrupted by a stop codon or their length was not a multiple of three were discarded. 147,115 genes divided into 26,932 gene families were retained. Finally, homologous gene families that contained strictly fewer than four sequences, the minimum number of sequences required for a phylogenetic tree to convey non-trivial information, were removed. Hence, the initial data set comprised 12,158 gene families and 125,909 genes (Table 2.1).

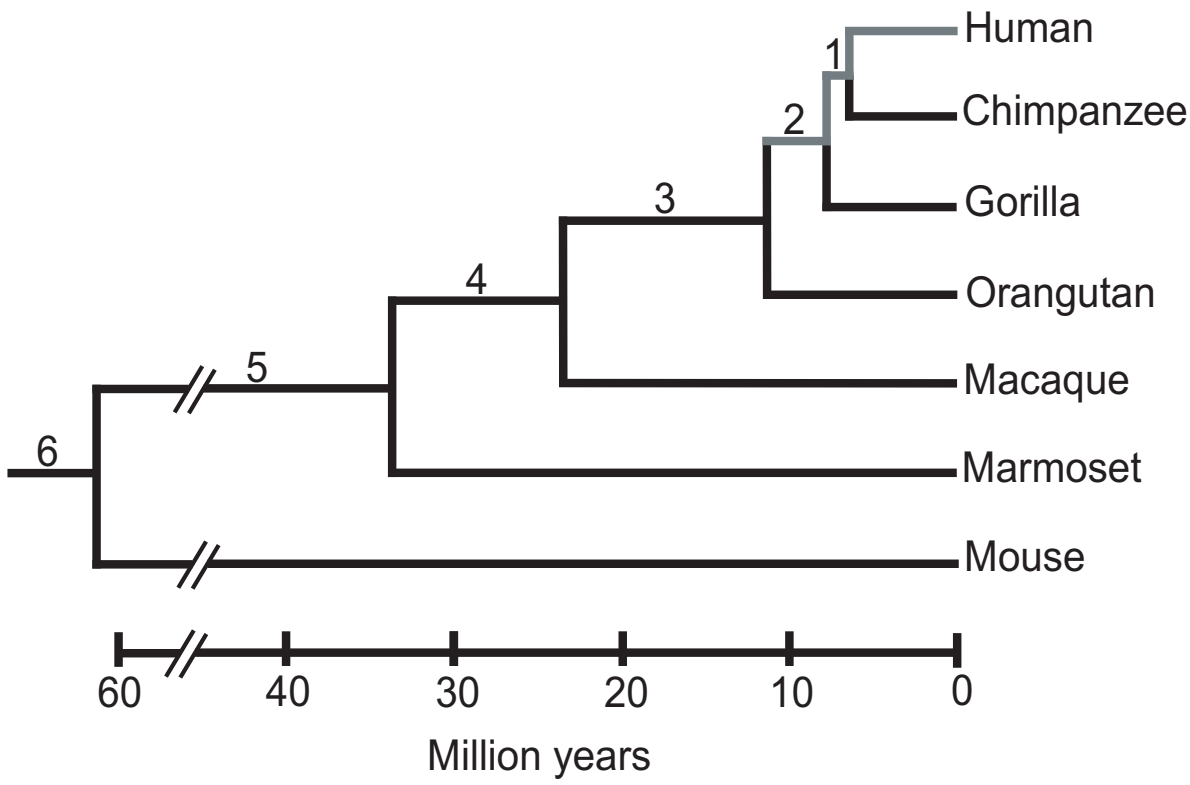


Figure 2.2 Primate phylogeny.

The numbers above the internal branches (1-6) represent the name that was assigned to a particular branch. Estimation of species divergence was retrieved from Benton et al. (2009; page 44).

Table 2.1 Summary of the data set preparation stage.

Species	Common name	No. of genes retrieved from Ensembl	No. of genes after filtering	# genes in fams with >4 seqs
Homo sapiens	Human	20,716	20,570	18,161
Pan troglodytes	Chimpanzee	19,829	19,758	17,321
Gorilla gorilla	Western Gorilla	20,962	20,934	17,836
Pongo abelii	Sumatran Orangutan	20,068	20,045	16,897
Macaca mulatta	Rhesus Macaque	21,905	21,890	18,302
Callithrix jacchus	Common Marmoset	21,168	21,150	18,244
Mus musculus	Mouse	23,038	22,768	19,148
Total		**147,686**	**147,115**	**125,909**

2.2.2 Network preparation

The human interactome was assembled from interactions available in the BioGRID database version 3.1.81 (Stark et al., 2011). Only non-redundant physical interactions among pairs of human proteins with an Ensembl ID were considered. The network (termed PIN0) contained 9,087 proteins connected by 39,883 interactions (Figure 2.3). As described in section 2.1.5, we addressed whether the phylogenetic trees of physically interacting proteins tend to be significantly more topologically similar than expected at random, and if this similarity could be attributed to co-duplication. For this purpose, it was necessary to produce a subnetwork solely comprising proteins belonging to gene families that portray non-trivial phylogenetic tree topologies. Proteins in the network belonging to homologous gene families that contained less than four sequences were removed. This sub-network (PIN1) comprised 8,650 proteins connected by 37,878 interactions (Figure 2.3).

There are a number of confounding factors that may influence results that are obtained in this analysis, and so must be ameliorated. Many functionally important genes encode proteins that self-interact. Such self-interaction confers several structural and functional advantages to proteins, including improved stability and specificity to active sites (reviewed in Marianayagam et al., 2004). PINs have been demonstrated to be enriched in such self-interactions (Ispolatov et al., 2005, Pereira-Leal et al., 2007). Thus, self-interacting pairs of proteins were removed from PIN1. Duplication of a gene encoding self-interacting proteins creates a pair of paralogous proteins that interact with each other. It is known that the PIN is enriched in proteins that are encoded by paralogous genes (Ispolatov et al., 2005, Pereira-Leal et al., 2007). As shall be described in the following sections, a set of genes that duplicated throughout the evolution of primates was identified. Thus, proteins that are encoded by paralogous genes were also removed. Once interactions between paralogous genes, and self-interactions were removed, what remained was a subnetwork of PIN1, which was named PIN2. As can be observed from Figure 2.3, PIN2

contained 8,518 proteins connected via 36,501 interactions.

Duplication events potentially affect large chromosomal regions. This involves the simultaneous duplication of multiple adjacent genes, which would consequently exhibit similar duplication histories. Furthermore, genes that encode functionally related proteins tend to cluster together in the genome (Lee and Sonnhammer, 2003, Makino and McLysaght, 2008). For each protein in PIN2, the chromosomal position of its encoding gene was retrieved from the Ensembl database. Pairs of interacting genes that are located on the same arm of the same chromosome were removed. Thus, PIN3 contained 8,426 proteins connected by 35,269 interactions (Figure 2.3).

It is important to gauge whether the observed outcome of an analysis is substantially different from expected. To achieve this, a null distribution of values under which the observed value should fall was obtained, and the difference between the expected and observed measure was examined. A set of random networks (either 250 or 10,000 random networks, depending on the analysis) was generated for each original network (PIN0, PIN1, PIN2 and PIN3) using a network rewiring approach. This algorithm functions by repeatedly selecting two edges at random (e.g., A–B and C–D) and swapping them (yielding A–D and C–B, or A–C and B–D). The operation was iterated $100 \times m$ times on each random network, where m is the number of edges. Therefore, each random network contains the same nodes, the same number of edges, and the same degree for each node as the original network. A measure that was calculated in this investigation was compared to the measure expected from a null distribution of random networks. P-values were computed as the proportion of random networks with a parameter value higher or equal to, or lower or equal to (depending on the section of the analysis) the observed network.

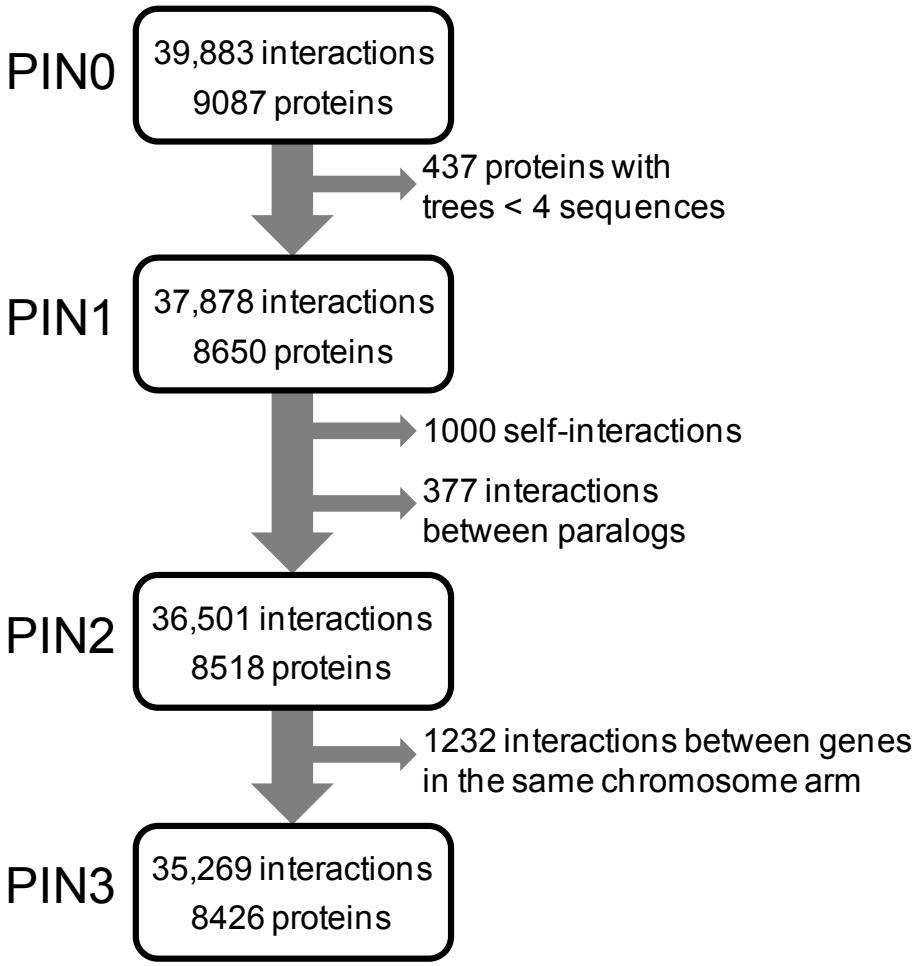


Figure 2.3 Description of subnetworks.

2.2.3 Phylogenetic tree reconstruction

The amino acid sequences of each primate homologous gene family were aligned using MUSCLE (Edgar, 2004) implemented in the TranslatorX program (Abascal et al., 2010). The resulting amino acid alignments guided the alignment of the corresponding nucleotide coding sequences. These nucleotide alignments were used to reconstruct phylogenetic trees using SPIMAP version 1.1 (Rasmussen and Kellis, 2011). SPIMAP is an empirically Bayesian algorithm that endeavours to accurately reconstruct gene trees using prior information that is learned from the genomes and species tree. The prior genomic information was captured in two parameters. First, a duplication and loss parameter aimed to learn gene duplication and loss rates across all of the homologous gene families. This was calculated by counting the number of genes per species in all 12,158 gene families using the "spimap-train-duploss" program of SPIMAP. A rate of gene birth of 0.007101 duplications/gene/million years and a rate of gene death of 0.002081 losses/gene/million years were estimated. The second parameter estimated substitution rates. 5,280 single orthologous gene families (i.e. gene families that contained at most one gene from each species) were extracted. For each of these gene families, phylogenetic trees were reconstructed using RAxML software (version 7.0.4) (Stamatakis, 2006) and the optimal model chosen by ModelGenerator (Keane et al., 2006). All other RaxML parameters were maintained at default settings. 1,732 gene trees with a topology congruent to the species tree were extracted. This set of gene trees was used to determine the substitution rate parameters using the "spimap-train-rates" program in SPIMAP. Once these parameters were calculated, SPIMAP reconstructed gene trees using the "spimap" program with default settings (that is, providing a gene family alignment as an input, a species tree, HKY85 as a model of DNA substitution and to allow base frequencies and transition/transversion ratios to be calculated empirically).

The setting of HKY85 as a universal model of DNA substitution for every gene family in the data set is obviously not ideal, as HKY85 may not accurately reflect the reality of how all the gene families evolved. However, it is currently the only model of DNA substitution that is implemented in SPIMAP. One may ask why not use an alternative method of phylogenetic tree reconstruction, in which more models of DNA substitution are implemented. Originally, phylogenetic trees were reconstructed using both SPIMAP (with the parameters described in the previous paragaph) and PhyML version 3.0 (Guindon and Gascuel., 2003). Using PhyML, the model of substitution was chosen by ModelGenerator.

An Approximately Unbiased (AU) test implemented in Consel v. 1.19 (Shimodaira and Hasegawa, 2001) was performed on each pair of phylogenetic trees (i.e. one gene family, with a phylogenetic tree reconstructed with both PhyML and SPIMAP). The AU test was successfully performed on 12,091 of the 12,158 pairs of phylogenetic trees. For 138 of these pairs of trees, the PhyML tree had a statistically significant higher likelihood value than the SPIMAP tree, according to the AU test. For 2,783 pairs of trees, the SPIMAP tree had a significantly higher likelihood value than the PhyML tree. In 9,170 pairs of trees, neither tree had a significantly higher likelihood than than the other. So in total, there were essentially only 138 of 12,091 cases in which PhyML proposed a significantly more likely tree topology than SPIMAP. As such, SPIMAP was selected as the method of phylogenetic tree reconstruction, in spite of the limitation of having to use HKY85 as a universal model of DNA substitution.

2.2.4 Gene duplication inference

This study used two approaches to identify gene duplication events. Given a species tree and a gene tree, the aim of gene/species tree reconciliation is to identify the topological discrepancies between a gene and species tree, and to interpret these discrepancies as duplication and loss events on the gene tree (Figure 2.4) (Goodman et al.,

1979, Page, 1994). For each of the 12,158 phylogenetic trees, gene duplications were inferred using gene/species tree reconciliation with the PhyloTree class in ETE package version 2.1 (Huerta-Cepas et al., 2010). However, an incorrectly reconstructed gene tree will inevitably invoke false inferences of duplication and loss using a gene/species tree reconciliation algorithm (Hahn, 2007). To address this possible confounding factor, duplications were independently inferred using a species overlap method, again with the PhyloTree class in ETE package version 2.1. In this procedure, a duplication is assigned if the same taxa are found either side of the node (Figure 2.4).

The timing of each duplication event was identified through examination of the species that were represented in the descendent leaves of a duplication node. The duplication event was assigned to the branch preceding the deepest node in the reference species tree whose descendants include all of the species involved in the duplication. For instance, if the taxa that descended from a duplication node included only great apes, the duplication event would be assigned to the branch immediately preceding the radiation of the great apes. This was conducted using the PhyloTree class in ETE package version 2.1, for both species overlap and reconciliation.

2.2.5 Mapping primate duplications to human interactome

Each duplication event was mapped to a human gene that encodes a protein in the human PIN, using the PhyloTree class in ETE package version 2.1. Briefly, the algorithm examined the descendant leaves of each duplication node. If there was at least one human gene in this set of leaves, the duplication event was assigned to this human gene (or set of human genes). Otherwise, the parental node of that node was systematically examined until the descendant leaves contained at least one human homolog. Thus, each duplication event was assigned to a human gene (or a set of human genes) that were either the result of this duplication or the closest human homolog(s) to the genes involved in this duplication.

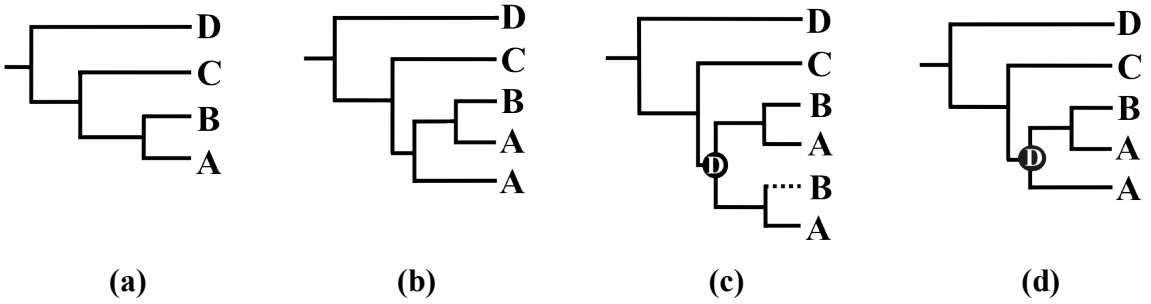


Figure 2.4. Gene duplication inference (a) Species tree (b) Gene tree (c) Reconciled gene tree (d) Duplication inferred with species overlap.

Duplication node is marked with a "D" in Figure 2.4c and 2.4d. A broken branch denotes a loss.

2.2.6 Biological enrichment analysis

A biological enrichment analysis was conducted to investigate if the set of human genes assigned to each branch of the species tree exhibited significant overrepresentation in certain biological characteristics compared to the rest of the human genome. This was conducted using the duplications that were inferred with gene/species tree reconciliation. FatiGO (Al-Shahrour et al., 2004) inputs two lists of genes (i.e. a list of human homologs that were assigned to a branch, and the rest of the human genome) and converts them into two lists of Gene Ontology (GO) terms (Ashburner et al., 2000). A Fisher's exact test for 2 × 2 contingency tables is subsequently implemented to identify significant over-representation of GO terms in one gene set with respect to the other. Duplicate genes within and between the two lists were removed using the "remove all duplicates" parameter. Each set of human genes was compared to the rest of the human genome. Enrichment was searched for in the "biological process" GO category using "direct transmission" for the GO levels 3-7. Fisher's exact test for "over-represented terms in list 1" was conducted, which is recommended when searching a set of genes against the rest of the genome (http://bioinfo.cipf.es/babelomicswiki/tool:fatigo). An adjusted P-value of less than 0.05 was deemed to be significant.

2.2.7 Exploration of the relationship between network structure and gene duplicability

As described in the introduction section, there were two questions posed in this investigation. The first question addressed the relationship between gene duplicability and network centrality. The second question asked if the phylogenetic tree topologies of physically interacting proteins were significantly more similar than expected from random pairs. Finally, we asked if any such similarity could be attributed to co-duplication. The methods used to address each of these questions shall be discussed in turn.

2.2.7.1 Analysis of the relationship between gene duplicability and network centrality in primates

It was hypothesized that there is no relationship between gene duplicability and protein centrality. First, this hypothesis was tested for all the proteins in the human interactome, and subsequently for the human genes assigned to each branch of the species tree individually.

For each protein in PIN0, degree, betweenness and closeness centralities were calculated using the "centrality" algorithm implemented in NetworkX package (http://networkx.lanl.gov/). To examine whether duplicated genes tend to occupy significantly more central positions in the PIN than non-duplicated genes, a Mann-Whitney test was implemented. The Mann-Whitney test is a non-parametric statistical test that assesses whether one of two samples of independent observations tend to have larger values than the other set of samples. The Mann-Whitney test was implemented in the "Wilcoxon-Rank Sum and Signed Rank Test" package (v. 2.16.0) in R (http://stat.ethz.ch/R-manual/R-patched/library/stats/html/wilcox.test.html). Subsequently, the centrality values for the set of human genes assigned to each branch of the phylogeny were obtained from the centralities that had been calculated for the proteins in the interactome. To examine whether each the genes assigned to each branch were more central than genes that had not duplicated in a particular branch, a Mann-Whitney test was once again performed, as described in previous paragraph.

Next, we examined the age profile of the sets of duplicated genes assigned to each branch of the phylogeny. A similarity search of each human gene against the GenBank NR database was conducted (downloaded on 12/10/2010; Pruitt et al., (2007)) using the BlastP algorithm (Altschul et al., 1990) implemented in Blast version 2.2.23, with an E-value cut-off set to 10^{-8}. Potential homologs were retained in the event that the database sequence

aligned to greater than 80% of the query sequence and met the E-value requirement. This step removes hits that were identified as putative homologs based on common domains or motifs. The taxonomic classification of each identified homolog was obtained from the NCBI Taxonomy database (downloaded on 12/10/2010). If all of the homologs to a particular human gene were classified as metazoan, the human gene was categorized as "metazoan-specific". Alternatively, if the human gene identified homologs in species that are more ancient than metazoans, the gene was classified as "ancient".

There was the possibility that the taxonomic classification of a single homolog could determine whether a human gene was metazoan-specific or ancient. For example, in the event that a particular human gene obtained 499 metazoan homologs, and a single *S. cerevisiae* homolog, this particular gene would be classified as "ancient". However, perhaps the single yeast homolog is not a true homolog of the human gene in question. To minimize the risk that a single homolog would determine the evolutionary origin of a human gene, the process of gene age assignment was repeated. In this case, a gene was defined as ancient if at least 5% of the homologs corresponded to non-metazoan genomes. In both cases, once every human gene was assigned as either "ancient" or "metazoan-specific", the proportion of ancient human genes in each branch was calculated.

As described in the introduction, one molecular mechanism by which gene duplication occurs is through a process called retroposition. As these "retrogenes" lack functional regulatory elements, it is thought that the majority of retrogenes will eventually become pseudogenized, creating "processed pseudogenes". Such genes do not encode functional proteins. We wanted to investigate the proportion of potential processed pseudogenes in each branch of the phylogeny. As transcription has occurred, processed pseudogenes do not contain introns. Thus, a gene was classified as a "suspected pseudogene" if it contained exactly one exon. The number of exons per human gene assigned to each branch of the phylogeny was retrieved from the Ensembl database

(version 61) and the proportion of intronless genes (i.e. suspected pseudogenes) was calculated for each branch in the phylogeny.

2.2.7.2 Comparison of phylogenetic tree topologies between physically interacting proteins

The first null hypothesis I tested is that phylogenetic tree topologies of physically interacting proteins are not any more similar than what would be expected from comparing tree topologies belonging to random pairs of proteins. Each pair of physically interacting proteins in PIN1 (i.e. the subnetwork comprising solely non-trivial tree topologies) was converted into a pair of phylogenetic trees using script 2.1 in script index 2.0. TreeKO is a duplication-aware algorithm that compares two tree topologies (regardless of the number of duplications present) and provides a Robinson-Foulds based distance measure of the similarity between two topologies (Marcet-Houben and Gabaldón, 2011). Using the "tree comparison" algorithm in TreeKO, a "strict distance" between every pair of phylogenetic trees in PIN1 was calculated. The strict distance is essentially a weighted Robinson-Foulds distance that penalizes differences in evolutionarily relevant events such as gene duplications and gene losses, returning a value between 0 and 1. If the strict distance is 0, it may be interpreted that the two trees are identical in terms of their topology (and as such duplication patterns). If the score is close to 1, the two trees are highly dissimilar in terms of their topology. Using the "root method" parameter, each phylogenetic tree was rooted in such a way that minimized the number of duplications assigned to each tree. Once a strict distance score was calculated for each pair of interacting phylogenetic trees in PIN1, an overall average topological distance was calculated for all the pairs of trees in PIN1.

It is important to evaluate whether the average topological distance that was calculated between the trees in PIN1 was to be expected. 250 randomized networks were assembled as described in section 2.2.2. Identical to the protocol explained in the previous

paragraph, each pair of interacting proteins in each of the 250 randomized PIN1s was converted into a pair of phylogenetic trees. A topological distance was calculated between each pair of phylogenetic trees and from this, an average topological distance was calculated. Subsequently, the average topological distance obtained for PIN1 was compared to the null distribution of average topological distances calculated from the 250 PIN1 randomized networks. A P-value was calculated as the proportion of randomized networks for which the average topological distance was lower than or equal to the observed network. Subsequently, the analysis described in this section was repeated for PIN2 (in which self-interactions and interactions between proteins encoded by paralogous genes were removed) and PIN3 (in which genes that are clustered on the same arm of the same chromosome were removed).

The second null hypothesis of this section states that the topological similarity between phylogenetic trees of physically interacting proteins is not attributed to co-duplication. First, it was considered whether the human interactome overall was enriched in interactions among proteins encoded by duplicated genes. The number of interactions involving genes that had undergone duplication in any branch of the phylogeny (N) in PIN1 was calculated. The network was randomized 10,000 times (see section 2.2.2 for protocol). For each of these randomized networks, the number of interactions involving genes that had undergone duplication in any branch of the phylogeny was re-calculated. A P-value was computed as the proportion of randomizations for which the simulated N-value was higher or equal to the observed value. The protocol described in this paragraph was repeated for PIN2 and PIN3.

Subsequently, the number of interactions in PIN1 between genes that underwent duplication in each branch of the phylogeny was computed as N_i. PIN1 was randomized 10,000 times (see section 2.2.2). For each random PIN1 network, the number of interactions between genes that underwent duplication in each branch was re-computed. A

P-value was calculated as the proportion of randomizations for which the simulated *N*-value was higher or equal to the observed value. The analysis described in this paragraph was repeated for subnetworks PIN2 and PIN3.

2.3 Results

2.3.1 Analysis of gene duplication

After filtering, there were 125,909 genes divided among 12,158 phylogenetic trees in the data set. The species/gene tree reconciliation approach inferred a total of 22,969 duplication events between all the phylogenetic trees, while the species overlap method inferred a total of 15,814 duplications. For duplications inferred with gene/species tree reconciliation, an overall gene duplication rate of 0.00348 duplications/gene/million years was estimated across the phylogeny of the studied species. However, this rate widely varied across the different branches of the tree, ranging from 0.0012 duplications/gene/million years (on the chimpanzee external branch) to 0.0252 duplications/gene/million years (on the internal branch 2; Figure 2.2). In agreement with previous reports (Hahn et al., 2007), an increased duplication rate was observed in the primate lineage (0.00388 duplications/gene/million years) compared to the mouse external branch (0.0018 duplications/gene/million years). Furthermore, an accelerated rate of duplication in great apes (0.0041 duplications/gene/million years) was observed compared to the average rate in primates (0.00388 duplications/gene/million years), once again consistent with previous observations (Hahn et al., 2007). Finally, there was a remarkably high gene duplication rate in the branch subtending the human, chimpanzee and gorilla clade (0.0252 duplications/gene/million years). Although this sudden burst of duplication has been previously described (Marques-Bonet et al., 2009), it is yet to be satisfactorily explained (Table 2.2, Table 2.3).

In total, 22,285 human genes were assigned to the phylogeny with gene tree/species tree reconciliation, and 16,339 human genes assigned with species overlap (Table 2.2; Table 2.3). The gene enrichment analysis identified 31 unique biological processes enriched among duplicated genes in the external branches of the phylogeny (Electronic Appendix 2.1). The enriched biological processes include immune/defense response,

sensory perception and xenobiotic metabolism. This corresponds to previous research that demonstrated that in primates, duplicated genes are enriched in GO categories associated with DNA metabolism, DNA recombination, DNA transposition, defense response, xenobiotic metabolism, sensory perception and signal transduction (Huerta-Cepas et al., 2007). Bailey and Eichler (2006) noted that the enrichment of duplicated genes in the immune system might reflect an increased sophistication in the primate immune, xenobiotic recognition and detoxification systems, thus facilitating changes to food sources or infectious agents. Additionally, significant sensory perception enrichment was clearly observed in the mouse external branch. 30.53% of genes assigned to this branch were annotated to the GO term "sensory perception of smell", compared with just 0.93% of the rest of the genome. This enrichment correlates with previous observations that olfactory transduction pathway genes have duplicated in the mouse lineage (Niimura and Nei, 2005).

Table 2.2 Gene duplications inferred (gene tree/species tree reconciliation).

Branch	Branch Length	# Duplications per branch	Duplication rate (Dups/gene/MY)	# Human genes assigned to branch
External branch:				
Human	6.5	495	0.0037	790
Chimpanzee	6.5	157	0.0012	217
Gorilla	8.0	424	0.0025	426
Orangutan	11.2	292	0.0013	342
Macaque	23.5	902	0.0018	731
Marmoset	33.7	1,526	0.0021	1,043
Mouse	61.5	2,584	0.0018	796
Internal branch:				
Branch 1	1.5	90	0.0030	179
Branch 2	3.2	1,655	0.0252	1,805
Branch 3	12.3	1,770	0.0071	1,906
Branch 4	10.2	2,127	0.0099	2,274
Branch 5	27.8	3,220	0.0055	3,108
Branch 6	-	7,727	-	8,668

Table 2.3 Gene duplications inferred (species overlap).

Branch	Branch Length	# Duplications per branch	Duplication rate (Dups/gene/MY)	# Human genes assigned to branch
External branch:				
Human	6.5	495	0.0037	790
Chimpanzee	6.5	157	0.0012	217
Gorilla	8.0	424	0.0025	426
Orangutan	11.2	292	0.0013	342
Macaque	23.5	902	0.0018	731
Marmoset	33.7	1,526	0.0021	1,043
Mouse	61.5	2,584	0.0018	796
Internal branch:				
Branch 1	1.5	90	0.0030	179
Branch 2	3.2	216	0.0033	344
Branch 3	12.3	331	0.0013	539
Branch 4	10.2	706	0.0033	1,029
Branch 5	27.8	967	0.0016	1,293
Branch 6	-	7,124	-	8,610

2.3.2 The relationship between network centrality and gene duplicability in the primate PIN

The relationship between the network centralities of a protein and the duplicability of its encoding gene was evaluated. Overall, duplicated genes occupied significantly more central positions in the human PIN0 than singleton genes (P-value = 2.89×10^{-13} for degree; P-value = 3.01×10^{-10} for betweenness; P-value = 2.11×10^{-14} for closeness). This is in agreement with previous observations in the human interactome (Liang and Li, 2007).

Examining each branch of the phylogeny individually, duplicated genes generally exhibited a higher average centrality than singleton genes in ten out of the thirteen branches of the species tree (Table 2.4, 2.5, 2.6). Unexpectedly, the opposite trend was observed in the three remaining branches (the human lineage, and internal branches 1 and 2; labelled in grey on Figure 2.2). This opposite trend is statistically significantly different in two of these branches for all three centralities and for duplications inferred with both reconciliation and species overlap (the human external branch and internal branch 1; Table 2.4, 2.5, 2.6). Therefore, we suggested that the relationship between centrality and duplicability has inverted during primate radiation.

D'Antonio and Ciccarelli (2011) demonstrated that human genes of ancient origin exhibit the same pattern as observed in *E. coli*, *S. cerevisiae* and *D. melanogaster* (i.e. duplicated genes tend to be less central). In contrast, human genes that originated within the metazoans exhibit the opposite trend. Thus, the varying relationship between centrality and duplicability could be explained if duplications in the human external branch and internal branch 1 and 2 primarily involved ancient genes. However, the proportion of ancient genes in these branches (on average, approximately 16% for duplications inferred with reconciliation) was generally lower than the proportion of

ancient genes assigned to the other branches of the phylogeny (approximately 21%) (Table 2.7, 2.8). Similar results are obtained for duplications inferred with species overlap, as the proportion of ancient genes inferred on the branches of interest was approximately 14%, compared to the rest of the branches, in which ~21% of the genes were classed as ancient. This indicates that the age of genes is not the factor responsible for the heterogeneity in the observed relationship between duplicability and centrality.

We hypothesized that the human external branch and internal branches 1 and 2 may be enriched in suspected pseudogenes. As pseudogenes are generally non-functional, the selective mechanisms that promote the positive association between centrality and duplicability may not operate on these genes. However, the branches of interest did not exhibit an extraordinarily different proportion of suspected pseudogenes to that observed in the other branches for duplications inferred with reconciliation or species overlap (Table 2.7, Table 2.8).

Table 2.4 Comparison of degree centrality between duplicated and non-duplicated genes for each branch of primate phylogeny.

Column 1 (entitled "Branch") indicates the name assigned to a branch. Column 2 is the number of duplicated genes that are present in a branch, and also present in PIN0. Column 3 is the mean degree for the genes that are present in that branch. Columns 4 and 5 are the number of genes, and the mean degree, for non-duplicated genes (i.e. every gene in the interactome, that has not duplicated in that particular branch). P-value indicated by Mann-Whitney test.

		Degree				
	Branch	**Duplicated**		**Non-duplicated**		**P – value**
		n	**mean**	**n**	**mean**	
External branches	**Human**	97	6.2	8,990	**8.81**	0.001
	Chimpanzee	26	**53.04**	9,061	8.65	0.3
	Gorilla	165	**11.89**	8,922	8.72	0.062
	Orangutan	110	**11.77**	8,977	8.74	0.218
	Macaque	374	**11.4**	8,713	8.67	0.001
	Marmoset	574	**11.38**	8,513	8.6	1.19E-08
	Mouse	229	**15.25**	8,858	8.61	0.174
Gene/species tree Reconciliation (Internal branch)	**Branch 1**	36	5.39	9,051	**8.79**	3.63E-04
	Branch 2	763	8.22	8,324	**8.83**	0.67
	Branch 3	909	**9.56**	8,178	8.69	1.56E-01
	Branch 4	977	**10**	8,110	8.63	0.015
	Branch 5	1,504	**9.61**	7,583	8.61	4.79E-05
	Branch 6	4,470	**9.92**	4,617	7.67	9.62E-11
Species overlap (Internal branch)	**Branch 1**	36	5.39	9,051	**8.79**	3.63E-04
	Branch 2	81	**9.75**	9,006	8.77	0.837
	Branch 3	141	**9.11**	8,946	8.77	0.818
	Branch 4	356	8.62	8,731	**8.78**	0.886
	Branch 5	476	**9.49**	8,611	8.74	0.266
	Branch 6	4,439	**9.92**	4,648	7.69	2.57E-10

Table 2.5 Comparison of closeness between duplicated and non-duplicated genes for each branch of the primate phylogeny.

Column 1 (titled "Branch") indicates the name assigned to a branch. Column 2 is the number of duplicated genes that are present in a branch, and also present in PIN0. Column 3 is the mean closeness for the genes that are present in that branch. Columns 4 and 5 are the number of genes, and the mean closeness, for non-duplicated genes (i.e. every gene in the interactome, that has not duplicated in that particular branch). P-value indicated by Mann-Whitney test.

		Closeness				
	Branch	**Duplicated**		**Non-duplicated**		**P-value**
		N	**Mean**	**N**	**Mean**	
External branches	**Human**	97	0.249	8,990	**0.266**	0.01
	Chimpanzee	26	**0.232**	9,061	**0.266**	0.105
	Gorilla	165	**0.267**	8,922	0.266	1.85E-04
	Orangutan	110	**0.271**	8,977	0.266	0.008
	Macaque	374	**0.274**	8,713	0.266	2.58E-16
	Marmoset	574	**0.273**	8,513	0.265	1.14E-20
	Mouse	229	**0.266**	8,858	0.266	0.043
Gene/species tree reconciliation (Internal branch)	**Branch 1**	36	0.238	9,051	**0.266**	0.003
	Branch 2	763	**0.271**	8,324	0.265	0.175
	Branch 3	909	0.265	8,178	**0.266**	0.328
	Branch 4	977	**0.267**	8,110	0.266	0.007
	Branch 5	1,504	**0.268**	7,583	0.265	1.19E-05
	Branch 6	4,470	**0.266**	4,617	0.266	1.01E-07
Species overlap (Internal branch)	**Branch 1**	36	0.238	9,051	**0.266**	0.003
	Branch 2	81	**0.273**	9,006	0.266	0.035
	Branch 3	141	**0.272**	8,946	0.266	0.084
	Branch 4	356	**0.267**	8,731	0.266	0.046
	Branch 5	476	**0.266**	8,611	0.266	0.013
	Branch 6	4,439	**0.266**	4,648	0.266	9.53E-08

Table 2.6 Comparison of betweenness between duplicated and non-duplicated genes for each branch of the primate phylogeny.

Column 1 (titled "Branch") indicates the name assigned to a branch. Column 2 is the number of duplicated genes that are present in a branch, and also present in PIN0. Column 3 is the mean betweenness for the genes that are present in that branch. Columns 4 and 5 are the number of genes, and the mean betweenness, for non-duplicated genes (i.e. every gene in the interactome, that has not duplicated in that particular branch). P-value indicated by Mann-Whitney test.

		Betweenness				
	Branch	**Duplicated**		**Non-duplicated**		**P-value**
		N	**Mean**	**N**	**Mean**	
External branches	**Human**	97	7849.72	8,990	**12922.32**	0.002
	Chimpanzee	26	**507787.62**	9,061	11448.03	0.26
	Gorilla	165	**21134.46**	8,922	12715.3	0.02
	Orangutan	110	**20053.03**	8,977	12780.13	0.097
	Macaque	374	**17163.64**	8,713	12683.79	1.71E-05
	Marmoset	574	**17281.96**	8,513	12570.56	1.10E-09
	Mouse	229	**72909.96**	8,858	11315.95	0.006
Gene/species tree Reconciliation (Internal branch)	**Branch 1**	36	7932.45	9,051	**12887.8**	0.005
	Branch 2	763	10162.68	8,324	**13116.16**	0.677
	Branch 3	909	**14722.74**	8,178	12662.03	0.130
	Branch 4	977	**14369.92**	8,110	12687.26	0.084
	Branch 5	1,504	**13455.87**	7,583	12751.61	2.09E-04
	Branch 6	4,470	**16462.57**	4,617	9388.22	1.07E-10
Species overlap (Internal branch)	**Branch 1**	36	7932.45	9,051	**12887.8**	0.005
	Branch 2	81	**14138.07**	9,006	12856.75	0.633
	Branch 3	141	**18260.83**	8,946	12783.18	0.549
	Branch 4	356	10864.04	8,731	**12949.89**	0.720
	Branch 5	476	**14686.19**	8,611	12767.67	0.223
	Branch 6	4,439	**16499.05**	4,648	9400.56	2.76E-10

Table 2.7 Age and pseudogene analysis (gene/species tree reconciliation).

Columns 1-3 indicate the suspected pseudogene content for each branch of the phylogeny.

Columns 4-6 indicate the proportion of genes in each branch that were classified as ancient.

Pseudogene Analysis				Age analysis			
				Gene classed as ancient, if 5% homologs are ancient		Gene classed as ancient, if one homolog is ancient	
Branch	**# Human genes in branch**	**# One-exon human genes in branch**	**% Intronless genes in branch**	**# Ancient genes in branch**	**% Ancient genes in branch**	**# Ancient genes in branch**	**% Ancient genes in branch**
External							
Human	790	147	18.61	75	9.49	115	14.56
Chimpanzee	217	46	21.20	23	10.60	33	15.21
Gorilla	426	93	21.83	94	22.06	120	28.17
Orangutan	342	53	15.50	76	22.22	96	28.07
Macaque	731	71	9.71	235	32.15	282	38.58
Marmoset	1,043	101	9.68	323	30.97	406	38.93
Mouse	796	334	41.96	102	12.81	160	20.10
Internal							
Branch 1	179	24	13.41	31	17.32	36	20.11
Branch 2	1,805	142	7.87	418	23.16	536	29.70
Branch 3	1,906	149	7.82	406	21.30	558	29.28
Branch 4	2,274	250	10.99	524	23.04	690	30.34
Branch 5	3,108	279	8.98	662	21.30	886	28.51
Branch 6	8,668	712	8.21	1796	20.72	2,530	29.19

Table 2.8 Age and pseudogene analysis (species overlap).

Columns 1-3 indicate the suspected pseudogene content for each branch of the phylogeny.

Columns 4-6 indicate the proportion of genes in each branch that were classified as ancient.

Pseudogene Analysis				Age analysis			
				Gene classed as ancient, if 5% homologs are ancient		Gene classed as ancient, if one homolog is ancient	
Branch	**# Human genes in branch**	**# One-exon human genes in branch**	**% Intronless genes in branch**	**# Ancient genes in branch**	**% Ancient genes in branch**	**# Ancient genes in branch**	**% Ancient genes in branch**
External							
Human	790	147	18.61	75	9.49	115	14.56
Chimpanzee	217	46	21.20	23	10.60	33	15.21
Gorilla	426	93	21.83	94	22.06	120	28.17
Orangutan	342	53	15.50	76	22.22	96	28.07
Macaque	731	71	9.71	235	32.15	282	38.58
Marmoset	1,043	101	9.68	323	30.97	406	38.93
Mouse	796	334	41.96	102	12.81	160	20.10
Internal							
Branch 1	179	24	13.41	31	17.31	36	20.11
Branch 2	344	46	13.37	51	14.83	68	19.77
Branch 3	539	96	17.81	104	19.29	128	23.75
Branch 4	1,029	204	19.83	210	20.41	284	27.60
Branch 5	1,293	233	18.02	251	19.41	337	26.06
Branch 6	8,610	712	8.27	1,779	20.66	2,511	29.16

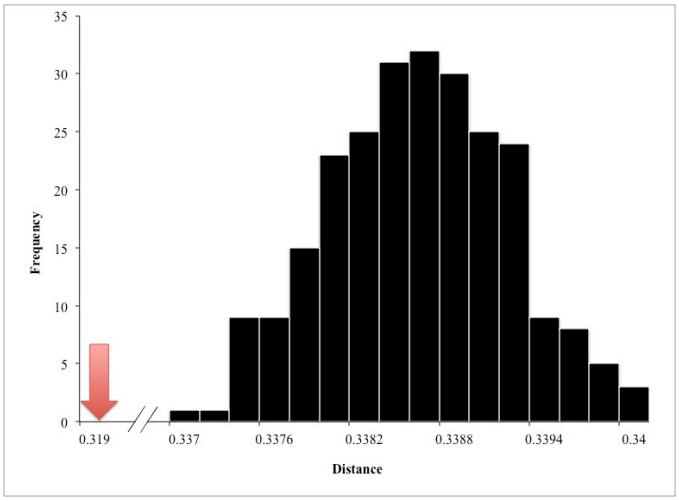
2.3.3 Comparison of the phylogenetic tree topologies of physically interacting proteins

The second null hypothesis in this investigation asserted that the tree topologies of physically interacting proteins are not more similar than what would be expected at random. In PIN1, each pair of interacting proteins was converted into a pair of interacting trees. The average topological distance between all the trees in a PIN was computed as 0.319. To establish whether the topologies between the trees of interacting proteins in PIN1 were more similar than expected at random, a null distribution of topological distances was calculated for 250 randomized PIN1 networks. None of the 250 randomized networks exhibited an average topological distance lower than or equal to the observed network (average value for the random networks, $D = 0.339$; $P < 0.004$; Figure 2.5a). This indicates that the phylogenetic trees belonging to physically interacting proteins were significantly more similar than expected at random.

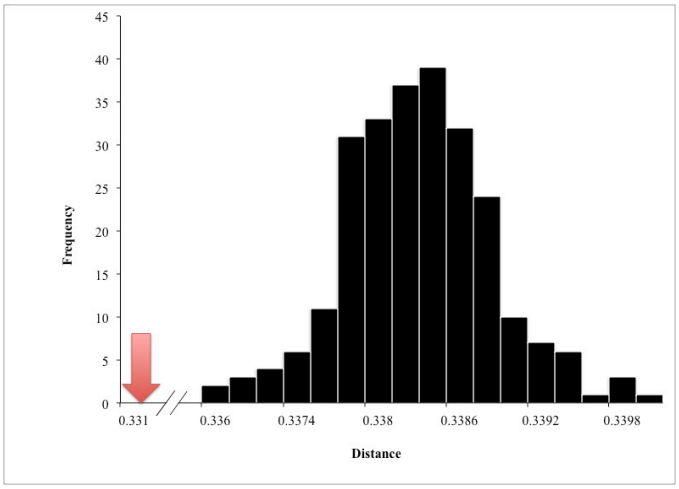
There are a number of structural features of PINs might also produce such a similarity and should be eliminated as potential sources of confounding bias. PIN2 is a subnetwork of PIN1, in which self-interactions and interactions among proteins that are encoded by paralogous genes have been removed. In PIN2, a significantly lower average topological distance between the trees of interacting proteins than expected in a random network was observed ($D = 0.331$; average value for the randomizations, $D = 0.338$; $P < 0.004$; Figure 2.5b). PIN3 is a subnetwork of PIN2 in which interactions between genes that are located on the same arm of a chromosome have been removed. Similarly in PIN3, the average topological distance between the trees of interacting proteins is still lower than expected at random ($D = 0.331$; average value for the simulations, $D = 0.338$; $P < 0.004$; Figure 2.5c). These results indicate that genes encoding interacting proteins manifest more similar tree topologies than expected from random pairs. In addition, the

observations suggest that this pattern is independent of the enrichment of the network in self-interactions, interactions among paralogous genes and interactions among genes that co-localize in the genome.

(a)



(b)



(c)

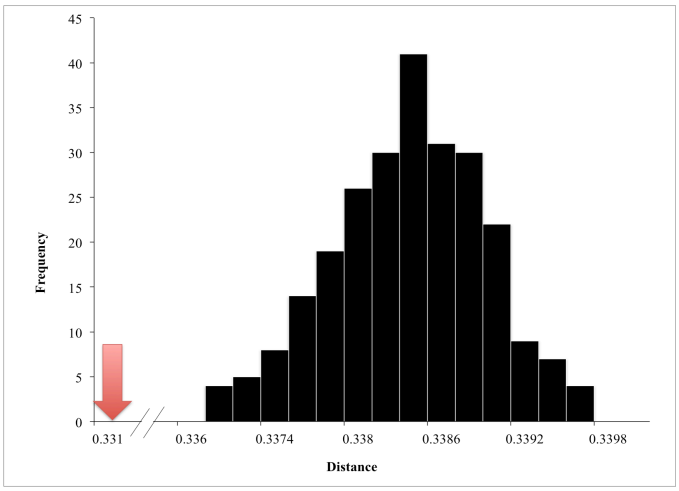


Figure 2.5 Distance between phylogenetic trees of interacting proteins in the human interactome. The observed average topological distance in the interactome is represented as a red arrow, and the average topological distribution inferred from 250 randomized networks is represented as a histogram. Figure (a) indicates distances for PIN1, (b) indicates distances for PIN2 and (c) indicates distances for PIN3.

2.3.4 Evidence for co-duplication in the primate phylogeny

The final null hypothesis stated that the observed similarity in phylogenetic tree topologies is not due to co-duplication. In this section, *N* indicates the number of interactions that occur between duplicated genes. First, it was found that the human interactome (PIN1) was enriched in interactions among proteins encoded by duplicated genes. None of the 10,000 random networks exhibited an *N* value higher than or equal to the observed one ($P < 0.0001$), indicating that duplicated genes tend to interact with each other in the interactome. This result holds when self-interactions and interactions among paralogs ($N = 21{,}872$; $P < 0.0001$ for PIN2), and interactions between genes locating in the same chromosome arm ($N = 21{,}152$; $P < 0.0001$ for PIN3) were removed from the analyses.

The number of interactions between duplicated genes were calculated for each branch of the primate phylogeny separately and compared to a null distribution of 10,000 randomized PIN1 networks. In PIN1, *N* values for each branch were significantly higher than expected from a random network in all 13 branches ($P < 0.05$), indicating that genes that have undergone duplication in each of these branches tend to interact with each other (Table 2.9). When self-interactions and interactions among paralogs are removed (PIN2), *N* values are higher than the average values for the random networks in 10 out of the 13 branches, with statistically significant differences in four of the branches (the external branches leading to gorilla, marmoset and mouse, and internal branch 6; Table 2.10). Almost equivalent results were obtained when interactions among genes in the same chromosome arm were removed from the analysis (PIN3; Table 2.11).

We concluded that although the tendency for genes that duplicated in a given branch to interact with each other is partially due to the enrichment of the network in self-interactions and interactions among paralogs, such features cannot completely account for the observed trend.

Table 2.9 Number of interactions between proteins encoded by genes duplicated in the phylogeny branch (PIN1).

N_i is the observed number of interactions (in PIN1) between proteins encoded by genes duplicated in branch *i*. M_i is the average *N* value across 10,000 network randomizations. Finally, *P* is the proportion of randomizations for which the simulated *N* value is higher or equal to the observed one.

Branch		**PIN1**		
		N	***M***	***P***
External Branches	**Human**	**14**	2.21	<0.0001
	Chimpanzee	**9**	3.73	0.001
	Gorilla	**55**	24.2	<0.0001
	Orangutan	**33**	10.52	<0.0001
	Macaque	**138**	114.28	0.0119
	Marmoset	**380**	269.35	<0.0001
	Mouse	**147**	65.73	<0.0001
Gene/species tree reconciliation	**Branch 1**	**2**	0.22	0.019
	Branch 2	**339**	244.43	<0.0001
	Branch 3	**555**	473.75	<0.0001
	Branch 4	**700**	601.32	<0.0001
	Branch 5	**1,502**	1313.07	<0.0001
	Branch 6	**13,658**	12340.78	<0.0001
Species overlap	**Branch 1**	**2**	0.22	0.019
	Branch 2	**14**	3.75	<0.0001
	Branch 3	**35**	10.24	<0.0001
	Branch 4	**98**	59.26	<0.0001
	Branch 5	**197**	127.93	<0.0001
	Branch 6	**13,509**	12166.9	<0.0001

Table 2.10 Number of interactions between proteins encoded by genes duplicated in the phylogeny branch (PIN2).

N_i is the observed number of interactions between proteins (in PIN2) encoded by genes duplicated in branch *i*. M_i is the average *N* value across 10,000 network randomizations. Finally, *P* is the proportion of randomizations for which the simulated *N* value is higher or equal to the observed one.

Branch		**PIN2**		
		N	***M***	***P***
External Branches	**Human**	1	**1.97**	0.8618
	Chimpanzee	**4**	2.6	0.2334
	Gorilla	**34**	23.21	0.0169
	Orangutan	**15**	9.79	0.0659
	Macaque	**96**	111.04	0.9443
	Marmoset	**299**	261.51	0.0059
	Mouse	**109**	61.92	<0.0001
Gene/species tree reconciliation	**Branch 1**	**1**	0.18	0.1668
	Branch 2	**249**	233.45	0.1421
	Branch 3	431	**452.42**	0.8816
	Branch 4	**582**	580.25	0.4765
	Branch 5	1,282	**1256.72**	0.1951
	Branch 6	**12,752**	11747.23	<0.0001
Species overlap	**Branch 1**	**1**	0.18	0.1668
	Branch 2	**5**	3.53	0.2763
	Branch 3	**18**	9.49	0.0078
	Branch 4	56	**56.75**	0.5569
	Branch 5	**130**	122.06	0.2353
	Branch 6	**12,605**	11576.51	<0.0001

Table 2.11 Number of interactions between proteins encoded by genes duplicated in the phylogeny branch (PIN3).

N_i is the observed number of interactions (in PIN3) between proteins encoded by genes duplicated in branch i. M_i is the average N value across 10,000 network randomizations. Finally, P is the proportion of randomizations for which the simulated N value is higher or equal to the observed one.

Branch		**PIN3**		
		N	***M***	***P***
Branches	**Human**	1	**1.82**	0.8407
	Chimpanzee	**3**	1.95	0.3054
	Gorilla	**33**	22.15	0.0153
	Orangutan	**14**	9.44	0.0921
	Macaque	89	**105.6**	0.9649
	Marmoset	**285**	252.35	0.0143
	Mouse	**103**	59.09	<0.0001
Gene/species tree reconciliation	**Branch 1**	**1**	0.16	0.1524
	Branch 2	**237**	223.89	0.1706
	Branch 3	415	**436.76**	0.8926
	Branch 4	**563**	562.98	0.5063
	Branch 5	**1,248**	1221.08	0.1776
	Branch 6	**12,332**	11359.67	<0.0001
Species overlap	**Branch 1**	**1**	0.16	0.1524
	Branch 2	**5**	3.38	0.2473
	Branch 3	**18**	9.03	0.0043
	Branch 4	54	**54.96**	0.5674
	Branch 5	**125**	117.24	0.2343
	Branch 6	**12,188**	11192.24	<0.0001

2.4 Discussion

The aim of this chapter was to combine comparative genomics and protein-protein interaction network data to explore the relationship between the structure of the primate PIN and the duplicability of the genes encoding its components. To achieve this, two distinct questions were asked. First, the relationship between protein network centrality and gene duplicability was evaluated. Second, we examined whether interacting proteins manifest topologically similar phylogenetic trees, and whether there is evidence for co-duplication among interacting proteins.

Consistent with previous observations in the *H. sapiens* genome (Liang and Li, 2007), it was observed that primate genes that duplicated in any branch of the species tree tend to be more central than singleton genes (Table 2.4, Table 2.5, Table 2.6). Unexpectedly, the opposite relationship between duplicability and centrality is observed in the external branch leading to humans, and in the internal branches subtending the human/chimpanzee (internal branch 1) and the human/chimpanzee/gorilla (internal branch 2). This resembles the pattern observed in *E. coli*, *S. cerevisiae* and *D. melanogaster* (Hughes and Friedman, 2005, Prachumwat and Li, 2006). Therefore, the relationship between duplicability and centrality seems to have undergone a reversal during the evolution of great apes.

The contrasting pattern observed among ancient and more recently evolved human genes could provide an explanation for the different relationship between centrality and duplicability that is observed in the different branches of the phylogeny (D'antonio and Ciccarelli, 2011). However, the proportion of ancient genes among genes that duplicated in the human external branch, and internal branches 1 and 2, is generally lower than genes that duplicated in the other branches of the phylogeny (Table 2.7, Table 2.8). This indicates that gene age is not the factor responsible for the heterogeneity in the

relationship between duplicability and centrality observed here. It was also examined whether these particular branches contained an unusually high proportion of suspected pseudogenes, as perhaps these genes would not be under the same selective pressure that promotes the relationship between gene duplicability and network centrality observed in the other branches. However, the proportion of suspected pseudogenes in these particular branches were not substantially different from that observed in other branches (Table 2.7, Table 2.8).

Next, it was observed that the phylogenetic trees of physically interacting proteins exhibit a higher similarity than expected at random. This was in spite of accounting for possible confounding factors such as self-interactions and interactions among paralogs, or interactions between functionally related genes. These observations contrast with those made by Kelly and Stumpf (2010), who recently found negligible evidence that pairs of interacting yeast proteins presented similar phylogenetic trees topologies. Three possible reasons might account for the different results obtained in both studies. First, Kelly and Stumpf analyzed the yeast interactome, whereas this study focused on primates. It is possible that both interactomes exhibit a different trend. Second, the data sets used by Kelly and Stumpf (2,528–5,109 proteins and 5,728–21,283 interactions) were remarkably smaller than those employed in the current study, which could have limited the statistical power in the analyses of Kelly and Stumpf. Lastly, Kelly and Stumpf inferred phylogenetic trees from 1:1 orthologous sets, which removes the effect of duplication and loss events in the tree topologies. In contrast, this investigation used entire homologous gene families, including paralogs. Thus, the different results obtained in the analysis by Kelly and Stumpf (2010) and this analysis may also potentially be the result, at least partially, of interacting genes exhibiting similar patterns of duplication and/or loss.

Finally, the number of interactions between genes that underwent duplication at

any branch of the phylogeny was found to be higher than expected from a random network. This observation indicated that duplicated human genes tend to interact with each other in the PIN, supporting the hypothesis that the duplication of a gene may increase the likelihood of duplication of its interacting partners. This trend holds true when genes that duplicated in each particular branch of the phylogeny are analyzed separately. The significance vanishes for most of the branches when self-interactions and interactions among paralogs are removed. However, the trend remains significant for four of the branches (the external branches leading to gorilla, marmoset and mouse, and internal branch 6; Table 2.11). Interestingly, these branches include the three longest branches to which the most duplications had been assigned (the external branches leading to mouse and marmoset, and internal branch 6). Perhaps the lack of significance in the remaining branches may be at least partially due to reduced statistical power in the shorter branches. Alternatively, the absence of significance in these branches might be a consequence of the reduced efficacy of selective mechanisms that favour the co-duplication of interacting genes in the same branches of the phylogeny. As one may expect that the selective advantage of duplicating the interacting partner of a protein would be small, the tendency of interacting genes to co-duplicate in the same branches of the species tree may only be observed in organisms in which natural selection is highly efficient. Primates have a lower effective population size (see section 4.1.2 for a description of effective population size) than rodents (Hughes and Friedman, 2009). Therefore, the evolutionary pressure that promotes the co-duplication of genes encoding interacting proteins may be less efficient in primates.

In summary, a major aim of the post-genomic era is to move from a reductionist approach of studying cell components to a more integrative view of the cell. A key element of this effort is to understand the interactions between the individual cellular

components, and to combine such information to produce models of entire biological systems. Taken together, the analyses describe in this chapter indicate that the primate PIN imposes constraints on the fate of genes encoding its components.

Chapter 3: An estimation of the timing of divergence between *A. melanoleuca* and *U. maritimus.*

3.1 Introduction

Ursidae is a taxonomic family comprising eight species assigned to three subfamilies, including Ailuropodinae (monotypic with the giant panda) and Ursinae (encompassing six species – one of which is the polar bear). The Ursinae experienced rapid radiation (Waits et al., 1999, Krause et al., 2008), and elucidation of the timing of speciation events in this clade has proved to be a challenging task (for example, Waits et al., 1999, Yu et al., 2004). In such investigations, a valuable node to calibrate is the divergence between polar bear and giant panda. This divergence is usually calibrated at 12 MYA (for example, Talbot and Shields (1996), Waits et al. (1999), Yu et al. (2004), Yu et al. (2007)), based on teeth morphology from a potential ancestor of the giant panda (Thenius, 1979, quoted in Krause et al., 2008) and genetic distances that were calculated over two decades ago (Wayne et al., 1991).

With the increased availability of molecular sequence data, the validity of the proposed 12 MY divergence between polar bear and giant panda has been questioned. For example, mitochondrial sequences have recently estimated that the two species may have diverged more anciently, approximately 19 MYA (Krause et al., 2008). The authors argue that as the fossil record for bear species is sparse and the oldest giant panda fossil, *Ailuropoda microta*, is dated at < 2.4 MYA (Jin et al., 2007), an early Miocene divergence between giant panda and polar bear is possible. To date, the largest study that has explored the Ursidae phylogeny using nuclear sequences consisted of merely 14 genes (Pagès et al., 2008). This data set did not attempt to estimate any divergence times among the Ursidae.

BGI-Shenzhen, China recently sequenced the full nuclear genomes of giant panda and polar bear, providing more data that could be used to address the divergence between these two species from a novel perspective. This chapter commences with an introduction to molecular phylogenetic dating and the Ursidae. Subsequently, a study is described that incorporated the largest amount of nuclear data available to date from 22 vertebrates, including two bear species, in a Bayesian framework, to accurately estimate the timing of divergence between the polar bear and giant panda. The acquisition of a more precise calibration point between these two species will aid future investigations that aim to resolve the enigmatic timing of divergence events that occurred among ursine bears.

3.1.1 An introduction to Ursidae

Ursidae is a taxonomic family encompassing eight species assigned to three subfamilies: Ailuropodinae (monotypic with *Ailuropoda melanoleuca*), Tremarctinae (monotypic with *Tremarctos ornatus*), and Ursinae (encompassing six species – *Ursus thibetanus*, *Ursus arctos*, *Ursus americanus*, *Ursus maritimus*, *Helarctos malayanus*, and *Ursus ursinus*) (Figure 3.1). Ursidae represent an interesting phylogenetic clade to investigate for a number of reasons. Ecologically, members of this family are present on most continents and occupy ecological niches ranging from the Arctic ice shelves to tropical rainforests. Conservationally, most bear genera are currently classified as threatened (http://www.redlist.org). For example, the polar bear (*Ursus maritimus*) has become emblematic of the decline of Arctic biodiversity as a result of global climate change (Hunter et al., 2010). In addition, multiple environmental risks threaten the sustainability of the giant panda (*Ailuropoda melanoleuca*) (Liu et al., 2001, Zhang, 2008). Finally, from an evolutionary perspective, the Ursinae subfamily experienced rapid radiation (Waits et al., 1999) and clarification of ursine bear speciation events has

proven to be a challenging task. In attempts to decipher the timing of such events, a valuable node to calibrate is the divergence between giant panda and polar bear (Figure 3.1).

Two primary sources of data have provided divergence estimates for the polar bear-giant panda speciation event. The first source of data is the examination of fossils. Thenius (1979) estimated that the polar bear-giant panda divergence occurred 12-15 MYA, based on teeth morphology of a potential ancestor of the giant panda (*Agriarctos*) (Krause et al., 2008). This estimate has been applied as a calibration point in numerous studies (for example, Waits et al. (1999), Yu et al. (2004)). More recently, the oldest giant panda fossil, *Ailuropoda microta*, was dated at < 2.4 MYA (Jin et al., 2007). Aside from a few recent discoveries (Ingólfsson and Wiig, 2008, Lindqvist et al., 2010), polar bear fossils are quite rare. This is thought to be because the animals live and die mostly over vast areas of sea ice. Thus, upon their death, it is likely that their remains are scavenged by other animals, or disappear into the ocean (Harington, 2008, Ingólfsson and Wiig, 2008, Laidre et al., 2008).

Molecular sequences have also been employed to explore the timing of speciation events in the Ursidae. Krause et al. (2008) used whole mitochondrial genomes to demonstrate that the giant panda-polar bear divergence may have occurred more anciently than suggested by fossils, approximately 19 MYA. Using 14 nuclear genes, the most comprehensive investigation that explored the Ursidae phylogeny using nuclear sequence data did not attempt to date the speciation events in the Ursidae phylogeny (Pagès et al., 2008).

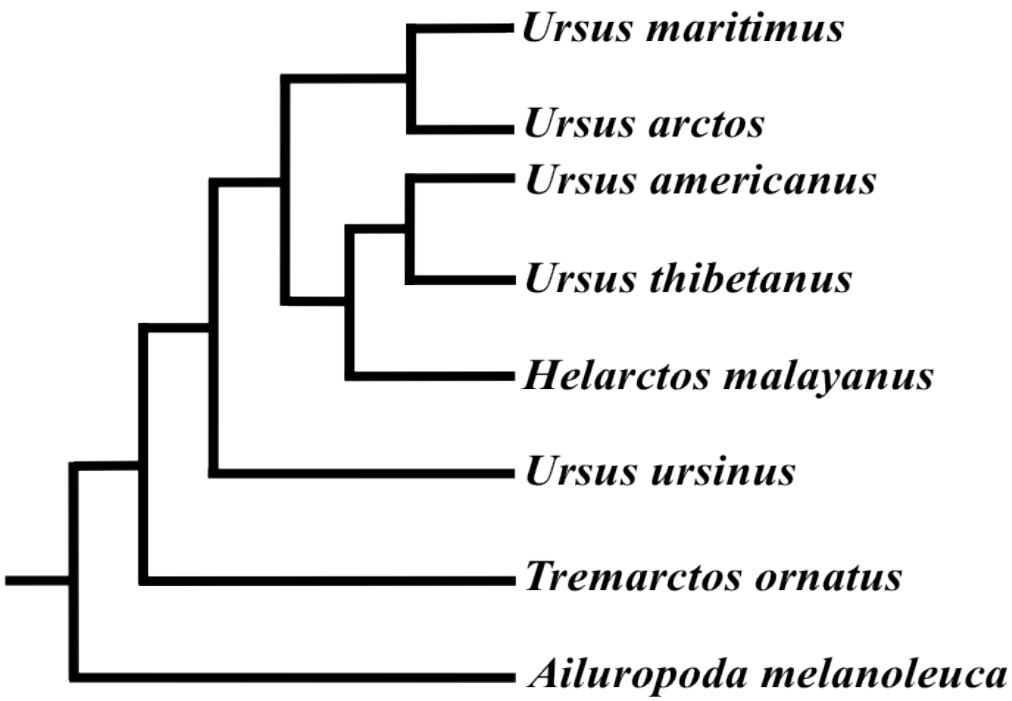


Figure 3.1 Ursidae phylogeny.

Different variations of the Ursidae phylogeny have been suggested, specifically with respect to the branching order of the ursine bears. The phylogeny presented here was adapted from Krause et al. (2008; Figure 1).

3.1.2 Estimation of divergence between polar bear and giant panda

The objective of the investigation described in this chapter is to establish an accurate estimate of the divergence between giant panda and polar bear using a large amount of nuclear sequence data in a Bayesian framework. In this section, the major components of a molecular phylogenetic dating analysis will be described.

3.1.3 The molecular clock hypothesis

Zuckerkandl and Pauling (1962) recognized that the similarity of protein sequences between different species could be indicative of the timing of speciation and gene duplication events, as molecular evolution appeared to occur at an approximately uniform rate over time. Soon after, the term molecular evolutionary clock (or "molecular clock" for short) was introduced (Zuckerkandl and Pauling, 1965). Recently, as the volume of molecular sequence data in the public domain has increased, the molecular clock concept has been applied to a range of evolutionary questions. For example, Korber et al. (2000) used the molecular clock to establish that the last common ancestor of the main pandemic strain of human immunodeficiency virus (HIV) diverged in the 1930s. This estimation was used to counter claims that the virus was originally spread through the distribution of a contaminated oral polio vaccine in Central Africa between 1957 and 1960 (Hooper, 2001). In addition, molecular clocks have been implemented to explore the timing of several controversial evolutionary events, such as the "Cambrian explosion" of the metazoan phyla (Bromham et al., 1998, Blair and Hedges, 2005, Erwin et al., 2011) and the proliferation of modern mammalian orders (Bininda-Emonds et al., 2007, Meredith et al., 2011).

3.1.4 Relaxed molecular clock models

Although the molecular clock hypothesis was originally postulated using empirical data, it soon received a theoretical backing. Kimura proposed the neutral theory of evolution, suggesting that a large proportion of mutations do not impact evolutionary fitness, so natural selection would neither favour nor disfavour them (Kimura, 1968, Kimura and Ohta, 1971). Eventually, each of these neutral mutations would either spread throughout a population and become fixed in all its members, or would be entirely lost through genetic drift. Kimura demonstrated that a molecular clock is expected if the rates at which mutations become fixed in a population (known as the substitution rate) is equivalent to the rate of appearance of new mutations in each member of the population (known as the mutation rate). This concept was defined as the "strict" molecular clock.

Subsequent research demonstrated that the null hypothesis of the molecular clock is too simplistic, as the rate of molecular evolution can vary significantly among genes and organisms (Wolfe et al., 1987, Steiper and Seiffert, 2012). To remedy this matter, "relaxed" molecular clocks were proposed, that retain some aspects of the original molecular clock hypothesis while relaxing the assumption of a strictly constant rate between all lineages of a phylogeny. Three probabilistic models of molecular clock relaxation were used in this investigation: Cox-Ingersoll-Ross (CIR) (Lepage et al., 2007), uncorrelated gamma (UGam) (Drummond et al., 2006) and lognormal (LogN) (Thorne et al., 1998). Primarily, each of these relaxed molecular clock models differ in whether autocorrelation occurs and the distribution used to model the rate of substitution.

Temporal autocorrelation, as implemented in LogN and CIR, limits the speed at which a substitution rate can vary from an ancestral to a descendent lineage (Sanderson, 1997). A descendent lineage may inherit an initial substitution rate from its ancestral lineage, and evolve new rates independently. The biological reasoning behind

autocorrelated relaxed molecular clocks can be summarized in the form of two key assumptions. The first assumption is that substitution rates are indirectly heritable because they are correlated with a variety of inherited characteristics. For example, among mammals, nucleotide substitution rates correlate with body size (Gillooly et al., 2005) and generation time (Nikolaev et al., 2007). The second assumption is that the rate of mutation and substitution are correlated. Unless evolution is proceeding in an effectively neutral manner, or closely related species are experiencing similar selection intensities, it is unlikely that substitution and mutation rates will be strictly correlated. As I cannot be certain of this second assumption for the species involved in this investigation, both autocorrelated (CIR and LogN) and an uncorrelated relaxed molecular clock model (UGam) were initially selected. Subsequently, Bayes factors were implemented to explore whether an autocorrelated or uncorrelated molecular clock model best suited the observed data.

Relaxed molecular clock also differ in terms of the distribution that describes the substitution rate variation between each ancestral and descendent branch. The LogN model assigns the ancestral substitution rate as the mean for a lognormal distribution, from which the descendent substitution rate is calculated. The CIR model operates slightly differently. It allows the rate of substitution to vary in descendent branches in a Brownian (spring-like) motion, in which small rate changes are permitted, while larger substitution rate changes are unlikely to be biologically accurate and so discouraged. Finally, regarding the UGam model, the rates of substitution are drawn independently of each other across the branches of the phylogeny, from a gamma distribution.

3.1.5 The fossil record

Fossils have been used for over 200 years to estimate the timing of speciation events. The fossil record describes the traces of organisms that remain in fossiliferous rock formations. Speciation divergence events can be estimated using a combination of molecular sequences and the fossil record. When the selected molecular clock is calibrated using external information about the geological ages of one or more nodes of the phylogeny, branch lengths that have been estimated from sequences can be converted into geological times.

Although the fossil record has provided many insights into divergence events for a wide range of species (for example, Douzery et al., 2004, Steiper and Young, 2006), it is regularly criticized as flawed (Graur and Martin, 2004). The imperfection of the fossil record has been known since Darwin's time. For example, he pointed out that many living organisms are composed solely of soft parts, and so are unlikely to be preserved. In addition, other species inhabit environments in which sedimentation does not accumulate. Finally, Darwin argued that the incompleteness of the fossil record may give the illusion of an explosive event, but with the eventual discovery of older and better preserved rock, the ancestors of taxa could potentially be identified.

There has been substantial discussion regarding how best to apply fossil constraints in a molecular phylogenetic dating analysis (Graur and Martin, 2004, Glazko et al., 2005). In essence, fossil calibration can either rely on a single or few "well documented" dates, or a large number of independent divergence estimates. Although the latter approach is by no means flawless, reliance on a single paleontological date has been strongly criticized (Lee, 1999, Graur and Martin, 2004). For this reason, a number of calibration points were selected for this analysis.

3.2 Materials and Methods

3.2.1 Data collection and data set assembly

A data set consisting of 21 vertebrate species (all of the species described in Table 3.1 with the exception of polar bear) was assembled. For each taxon, the longest canonical transcript isoform for all of the protein coding genes was retrieved from Ensembl BioMart database (version 62; April 2011) (Kinsella et al., 2011). This initial data set comprised 554,087 genes. Once genes that had not been assigned to an Ensembl gene family were removed, 421,430 genes remained in total. The protein coding genes of polar bear were retrieved directly from BGI Shenzhen, China and added to the data set.

From the updated data set that included polar bear sequences, genes that were unlikely to encode functional proteins as their coding sequence was interrupted by a stop codon or their length was not a multiple of three, were discarded. A total of 415,069 non-polar bear genes separated among 58,451 gene families, and 21,888 polar bear genes that are not yet deposited in the Ensembl database (and thus have not been assigned to an Ensembl gene family) were retained (Table 3.1).

Table 3.1 Genes retained at each stage of the data set assembly.

Polar bear protein coding genes are currently not available in the Ensembl database. The asterix (*) denotes that none of the polar bear genes had been assigned to Ensembl gene families. Assignment of polar bear genes to homologous gene families occurred in Section 3.2.2.

Taxonomic Name	Species	No. of genes retrieved from Ensembl	No. of genes in Ensembl families	No. of genes after filtering
Anolis carolinensis	Anole	22,269	18,939	17,792
Gallus gallus	Chicken	17,934	16,736	16,508
Pan troglodytes	Chimpanzee	27,166	19,829	19,758
Bos taurus	Cow	25,670	21,048	21,009
Canis familiaris	Dog	24,660	19,305	19,281
Loxodonta africana	Elephant	23,208	20,020	20,019
Taeniopygia guttata	Finch	18,581	17,475	17,470
Gorilla gorilla	Gorilla	29,216	20,962	20,934
Cavia porcellus	Guinea Pig	25,028	18,673	18,661
Equus caballus	Horse	26,954	20,436	20,381
Homo sapiens	Human	22,268	21,285	21,285
Macaca mulatta	Macaque	30,247	21,905	21,890
Callithrix jacchus	Marmoset	32,339	21,168	21,150
Mus musculus	Mouse	36,814	23,038	22,772
Monodelphis domestica	Opossum	22,038	19,466	19,448
Pongo abelii	Orangutan	28,087	20,068	20,045
Ailuropoda melanoleuca	Giant Panda	23,225	19,330	19,329
Sus scrofa	Pig	20,460	17,493	17,480
Ornithorhynchus anatinus	Platypus	22,369	17,951	17,934
Ursus maritimus	Polar Bear	22,673	*	21,888
Oryctolagus cuniculus	Rabbit	23,365	23,365	19,025
Rattus norvegicus	Rat	29,516	22,938	22,898
Total	**-**	**554,087**	**421,430**	**436,957**

3.2.2 Orthology assignment and gene family alignment

All of the sequences in the data set were translated from nucleotides into amino acids using the Transeq program implemented in the EMBOSS package version 6.5.0.0 (Rice et al., 2000). Codons were translated into the first open reading frame using the standard genetic code. A reciprocal BLAST strategy was employed to detect putative orthologs between polar bear and the other vertebrates. To achieve this, I implemented the BlastP algorithm (version 2.2.21) to compare the polar bear genome against all the other vertebrates (and vice versa) with an E-value cut-off set to 10^{-10}.

Each polar bear gene was inserted into the homologous gene family of its reciprocal best hit. To generate a meaningful phylogenetic hypothesis of species evolution, it was essential to solely align orthologous genes that diverged via speciation events. 4,993 single copy orthologous gene families, consisting of 99,533 genes in total, were extracted. Each of these single copy orthologous gene families contained at most one sequence per species in the data set. The sequences in each single copy orthologous gene family were aligned using Muscle 3.7, with default settings (i.e. providing the homologous gene family, and not invoking the use of any parameters to accelerate the alignment procedure).

3.2.3 Alignment improvement

Visual inspection of the homologous gene family alignments demonstrated that there was heterogeneity in multiple sequence alignment quality that may affect phylogenetic tree reconstruction. From this set of 4,993 alignments, three supermatrices were constructed and compared to examine the quality of the data used in the investigation. First, each single orthologous family was concatenated into a supermatrix that was 3,268,187 residues long. If a taxon was missing in a particular family, the sequence was represented by a series of question marks. However, as discussed in section 1.3, the amount of missing data that can be tolerated

in a supermatrix has been the source of debate. To ameliorate the impact that missing data could potentially have on the phylogenetic inferences generated in this investigation, partial single copy orthologous gene families (i.e. those that did not have a representative from each of the 22 species) were removed. 405 single orthologous gene families in this data set contained the full 22 taxa. These 405 gene families were concatenated into a supermatrix, comprising 303,480 residues. This is almost 10% of the original alignment, which was 3,268,187 residues in length. To clarify, at this point, there were two independent supermatrices: the original supermatrix of the concatenated data, and the supermatrix after partial single orthologous gene families were removed. Some regions of the aligned sequences may have been ambiguously aligned or extremely divergent, potentially biasing any inferred phylogenetic hypothesis. There are a number of algorithms available to automatically detect and remove potentially mis-aligned regions, such as trimAl version 1.4 (Capella-Gutiérrez et al., 2009), Block Mapping and Gathering with Entropy (BMGE) version 1.1 (Criscuolo and Gribaldo, 2010) and Gblocks version 0.9b (Castresana, 2000). Each of the 405 single copy orthologous gene family alignments (in which all 22 species were present) was individually improved using Gblocks with the parameters: gapped positions were eliminated, the minimum block length was set to 8 amino acid positions, while the maximum number of permitted consecutive non-conserved positions was set to 15. Subsequently, each gene family was concatenated into a supermatrix. The resultant supermatrix was 102,740 residues in length. Each single copy orthologous gene family (without partial families) was also independently improved two other automatic alignment improvement software programs (trimAl and BMGE), and supermatrices were constructed. However, in subsequent analyses, these additional supermatrices produced almost identical results to those described for Gblocks. Thus, Gblocks was arbitrarily selected as the alignment improvement program of choice, and data for the supermatrices derived using trimAl and BMGE are not shown. To be clear, at this

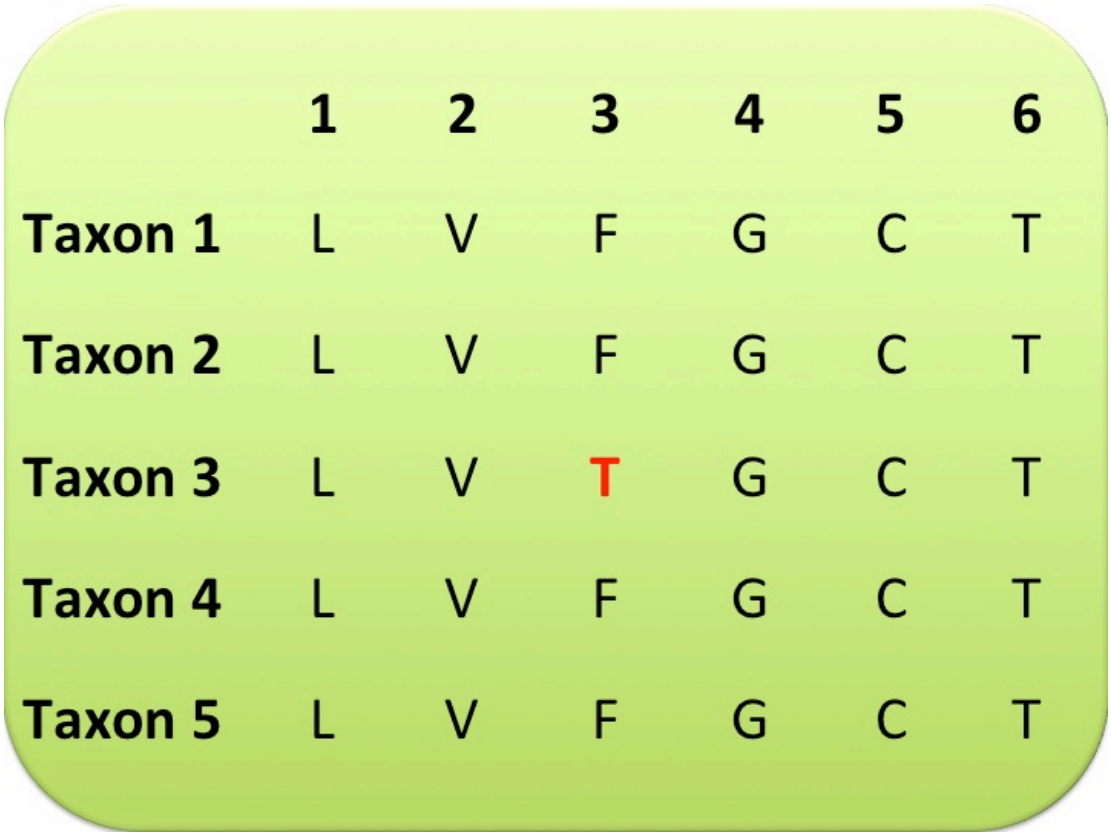
point, there were three supermatrices. These were:

(i) The supermatrix (3,268,187 residues in length) of 4,993 single orthologous families (full and partial families, that had not been subject to any alignment improvement procedure).

(ii) The supermatrix (303,480 residues in length) of 405 single orthologous families with partial families removed (that had not been subject to any alignment improvement procedure).

(iii) The supermatrix (102,740 residues in length) of 405 single orthologous families that had been improved with Gblocks, with partial homologous gene families removed.

A singleton may be defined as a unique residue in an otherwise completely conserved column of an alignment (Figure 3.2). This singleton may be the result of an evolutionary process such as positive selection, but could also be produced artificially as a result of sequencing or assembly errors. The number of singletons present per species in each supermatrix was calculated, using script 3.1 in script index 3.0. The supermatrix that had been improved with Gblocks was deemed to be the most sensible alignment for use for the remainder of the investigation.



	1	2	3	4	5	6
Taxon 1	L	V	F	G	C	T
Taxon 2	L	V	F	G	C	T
Taxon 3	L	V	T	G	C	T
Taxon 4	L	V	F	G	C	T
Taxon 5	L	V	F	G	C	T

Figure 3.2 An example of a singleton.

In column 3 of this alignment, taxon 3 has a singleton character state. It is denoted in red.

None of the other sites manifest a singleton.

3.2.4 Phylogenetic tree reconstruction

There are a number of considerations necessary for phylogenetic tree reconstruction. A phylogenetic outgroup may be defined as a species (or set of species) that is closely related to, but phylogenetically distinct from, the set of taxa under investigation. Three species formed the outgroup for this investigation: *Gallus gallus*, *Taeniopygia guttata* and *Anolis carolinensis*. Another important consideration is selection of an appropriate amino acid substitution model. ModelGenerator (Keane et al., 2006) selects the best-fitting model of substitution according to AIC, BIC and a corrected AIC. As the supermatrix was 102,740 amino acid residues in length, it encountered computer memory issues. Instead, a "sliding window" approach was implemented. Sections of the supermatrix that were 10,000 amino acids in length were extracted. These sections corresponded to supermatrix residues 10,000-20,000; 30,000-40,000; 50,000-60,000; 70,000-80,000 and 90,000-100,000. ModelGenerator was used on each section individually. The number of substitution rate categories was set to 4. ModelGenerator consistently found the same model (JTT) for all sections of the supermatrix. Equivalent results were obtained using ProtTest3 instead of ModelGenerator. Thus, JTT was the model of amino acid substitution selected for phylogenetic tree reconstruction.

RAxML (version 7.0.4) (Stamatakis, 2006) reconstructed a phylogenetic tree topology for the supermatrix using a "PROTGAMMAIJTTF" model of amino acid substitution. This model uses the Jones-Taylor-Thornton amino acid substitution matrix (Jones et al., 1992). The "GAMMA" parameter indicates that the rate variation among the sites follows a gamma distribution. A proportion of the sites are deemed invariant as denoted by the "I" parameter, and the amino acid frequencies are estimated from the data set, as indicated by the "F" parameter. The supermatrix was bootstrapped 100 times. The outgroup was specified as *Gallus gallus*, *Taeniopygia guttata* and *Anolis carolinesis* and a user starter tree was specified (phylogeny of tree provided from Benton et al. (2009); Figure 3).

3.2.5 Preparation for molecular phylogenetic dating analysis

There are a number of requirements for a molecular phylogenetic dating analysis. Some of these have already been described (i.e. an alignment, a phylogenetic tree, an outgroup and a model of amino acid substitution). Other factors necessary are a relaxed clock model and a reliable set of calibrations. The selection of each of these shall be discussed.

PhyloBayes version 3.3 (Lartillot et al., 2009) was used to conduct the molecular phylogenetic dating analysis. First, it was necessary to identify the most suitable relaxed clock model. A deconstrained model observes the level of signal that the data itself exhibits, in the absence of an amino acid substitution or relaxed molecular clock model. The deconstrained model for the supermatrix was computed using "estbranches" software implemented in the multidistribute package (version 09.25.03) (Thorne, 2003). The supermatrix was provided, and a "JTT" amino acid substitution model was selected. Subsequently, the "bf" program was implemented in PhyloBayes version 3.3 in which each molecular clock model (i.e. CIR, LogN and UGam) was compared to the deconstrained model using the "long" parameter. Subsequently, the ratio of each relaxed clock model to the deconstrained model was calculated. From these ratios, the ratio of each relaxed molecular clock model to every other relaxed clock model was computed. Bayes Factors were interpreted according to Kass and Raftery (1995) . A set of 13 calibrations was chosen from Benton et al. (2009) (Table 3.2). Each calibration is described as a range that extended from a minimum constraint to a maximum constraint.

Table 3.2 Calibrations selected for molecular dating analysis.

This set of calibrations was selected from Benton et al. (2009).

Species	Species	Soft Maximum (MY)	Hard minimum (MY)
Mouse	**Rat**	14	10
Chicken	**Zebra Finch**	86.5	66
Human	**Chimpanzee**	10	5.7
Human	**Orangutan**	33.7	11.2
Human	**Macaque**	34	23.5
Human	**Marmoset**	65.8	33.7
Cow	**Pig**	65.8	52.4
Platypus	**Anole**	330.4	312.3
Mouse	**Guinea Pig**	58.9	52.3
Anole	**Chicken**	299.8	255.9
Human	**Opossum**	171.2	124
Human	**Platypus**	191.1	162.9
Horse	**Dog**	131.5	62.5

3.2.6 Molecular dating analysis

3.2.6.1 Preparation for molecular dating analysis

In addition to the observed sequence data, Bayesian phylogenetic inference can incorporate other sources of knowledge through the application of prior probability distributions, or "priors" for short. The prior may be described as the probability distribution that would express one's uncertainty about a quantity, before data is taken into consideration. In turn, the prior is combined with a likelihood function to provide a posterior probability distribution.

In this investigation, a birth-death prior distribution for the divergence times of the internal nodes was selected. The birth-death prior permits the implementation of soft bounds, which in turn allows for the integration of uncertainty in the calibrations provided (Yang and Rannala, 2006). This is important because hard bounds, such as those imposed by a uniform prior, often provide good lower bounds, but problematic upper bounds. Thus, a more flexible distribution and soft bounds are preferable, so a birth-death prior and soft bounds were specified (using the "bd" and "sb" parameters in PhyloBayes).

There are cases in which the prior distribution on divergence times that is proposed without calibrations may be substantially different from the prior probability conditional on the calibrations. Thus, an MCMC chain was conducted using PhyloBayes version 3.3 "under the prior" to determine if the distribution of nodes of interest sampled by the MCMC is sufficiently wide (i.e. non informative), before proceeding with a posterior sampling. To achieve this, the "prior" parameter was selected that deactivates all maximum likelihood computations. In addition, the set of calibrations, supermatrix, and phylogenetic tree that have previously been described were provided. The "soft bounds" parameter was set to 0.3 meaning that a proportion of 0.3 of the total probability mass was allocated outside of the specified bound. A birth death process and a JTT amino acid substitution matrix were

selected. Two MCMC chains were constructed in parallel for 25 hours, creating 10,400 and 9,624 cycles each. Convergence was assessed by implementing the "tracecomp" program in PhyloBayes with a burn-in of 1,000 cycles. That is to say, I removed approximately 10% of the total number of cycles from subsequent analyses, as suggested in the PhyloBayes version 3.3 manual (page 11). I examined how often the bounds (i.e. the age limits set by the calibration file) were disturbed, and the dates that were assigned for the uncalibrated internal nodes. These dates should be wide in range, as they are not yet being constrained by the calibration dates that are imposed on surrounding nodes.

3.2.6.2 Molecular dating analysis

Once a set of parameters had been selected, the next stage was to perform the molecular dating analysis. The "pb" program in Phylobayes was conducted on a supermatrix in duplicate. A JTT amino acid substitution model was used, the soft bounds parameter was set to 0.3 and the set of calibrations were provided. As the Bayes factor analysis failed to clearly select an optimum relaxed clock model, CIR, LogN and UGam were selected in turn. Each pair of MCMC chains was constructed for approximately 400 hours (generating ~ 6,300 cycles), until convergence was reached. Convergence was assessed using the "tracecomp" program implemented in PhyloBayes. The burn in was set to 10% of the total number of cycles for each model.

3.2.6.3 Sensitivity analysis

There are a number of parameters throughout the course of a molecular dating analysis that, when altered, may impact an inferred estimation of divergence. Two of these factors are the amino acid substitution model selected and the set of calibrations provided. JTT was the optimum model selected by both ModelGenerator and ProtTest. But what impact does altering the substitution model have on the inferred estimation of species divergence? The molecular

dating analysis was repeated exactly as described in section 3.2.6.2. The only variable that altered each time was the sequential replacement of the amino acid substitution model with WAG, GTR or LG.

Due to the incompleteness of the fossil record, it is not always possible to accurately calibrate every node surrounding the node of interest. An example of this issue is clearly observed in this study. Benton et al. (2009) provides the most comprehensive overview of vertebrate calibration points currently available. However, using the Benton et al. guide to calibration and the phylogeny that RAxML reconstructed, there are six distinct nodes that have been assigned almost identical calibration bounds (61.5/62.5 – 131.5 MY) (Table 3.3). To identify the impact that calibration of any of these nodes had on the final divergence dates produced, five alternative sets of calibration points were composed. Each alternative set maintained 12 of the 13 original calibrations. The only bounds that varied was the 13th node (Table 3.4). Each time, the 13th node was sequentially selected from Table 3.3. For each set of alternative calibrations, the molecular dating analysis was repeated, exactly as described in 3.2.6.2.

Table 3.3 Identical calibration points

Six nodes that were described with almost identical calibration points in Benton et al. (2009). The effect that using any one of these calibration had on inferred divergence estimates was examined.

Calibration	Species	Species	Soft Maximum (MY)	Hard minimum (MY)
Original	**Horse**	**Dog**	131.5	62.5
2	**Cow**	**Dog**	131.5	62.5
3	**Human**	**Elephant**	131.5	62.5
4	**Human**	**Horse**	131.5	62.5
5	**Human**	**Rabbit**	131.5	61.5
6	**Rabbit**	**Guinea Pig**	131.5	61.5

Table 3.4 Alternative calibration selection.

The original molecular dating analysis was conducted a further five times, to test the effect of using different calibration combinations on the inferred date. In each of the five analyses, calibration #13 was sequentially altered in each round of molecular dating to one of the calibration points described in Table 3.3.

Calibration #	Species	Species	Soft Maximum (MY)	Hard minimum (MY)
1	**Mouse**	**Rat**	14.0	10.0
2	**Chicken**	**Zebra Finch**	86.5	66.0
3	**Human**	**Chimpanzee**	10.0	5.7
4	**Human**	**Orangutan**	33.7	11.2
5	**Human**	**Macaque**	34.0	23.5
6	**Human**	**Marmoset**	65.8	33.7
7	**Cow**	**Pig**	65.8	52.4
8	**Platypus**	**Anole**	330.4	312.3
9	**Mouse**	**Guinea pig**	58.9	52.3
10	**Anole**	**Chicken**	299.8	255.9
11	**Human**	**Opossum**	171.2	124
12	**Human**	**Platypus**	191.1	162.9
13	**X**	**X**	131.5	61.5/62.5

3.3 Results

3.3.1 Phylogenetic tree reconstruction

In total, 4,993 single copy orthologous gene families were extracted and concatenated into a supermatrix that was 3,268,187 residues in length. Once partial families were removed, 405 gene families remained. These genes were combined into a supermatrix that was 303,480 residues in length. After improvement with Gblocks, a supermatrix that was 102,740 residues in length remained.

The number of singletons in each of the supermatrices was compared. It is evident that the vertebrate genomes used in this analysis are heterogeneous in the number of singletons per species. For example, excluding dog, there are 12,287 singletons per species in the original supermatrix. The dog, however, has 138,953 singletons; this is over ten times this average. Although such singletons could indeed arise from biological processes such as positive selection, to possess over ten times the average number per species seems suspicious. Similar unusually high numbers are observed in the pig, platypus and polar bear (Table 3.5). Once partial families were removed, the average number of singletons per species reduced dramatically. Unusually high numbers of singletons were observed in some of the genomes that were sequenced at low coverage (for example, the anole) or are considered to be of poor quality (for example, the platypus; Dr. David Alvarez-Ponce, personal communication). The supermatrix that had been improved with Gblocks was divided into sections of 10,000 amino acids, and ModelGenerator unanimously selected JTT as the optimal model of amino acid substitution for all sections (Table 3.6, almost identical results were obtained using ProtTest instead of ModelGenerator – data not shown). Thus, "JTT + I + G + F" was the set of parameters selected for phylogenetic tree reconstruction. This model uses the Jones-Taylor-Thornton amino acid substitution matrix (Jones et al., 1992). The "GAMMA" parameter indicates that the rate variation among the sites follows a gamma distribution. A proportion of

the sites are deemed invariant as denoted by the "I" parameter, and the amino acid frequencies are estimated from the data set, as indicated by the "F" parameter.

A phylogenetic tree was reconstructed and the resulting topology was examined. Generally, 100% bootstrap support was found for all the nodes in the tree, with the exception of the nodes defining the human/chimpanzee/gorilla clade and the horse/dog clade (Figure 3.3).

The low bootstrap support in the human-chimpanzee-gorilla clade reflects the uncertainty that has surrounded this particular phylogeny for a long period of time (for example, O'hUigin et al., 2002, Langergraber et al., 2012). As the gorilla lineage is thought to have diverged shortly before the human-chimpanzee lineage, polymorphisms in the ancestral population of all three species could persist from the divergence of the first to the second species (O'hUigin et al., 2002). The random fixation of alleles in these three lineages led to a situation in which different nucleotide positions in the species are represented by different phylogenies. Thus, the human genome may be considered a mosaic of different regions that are differentially related to chimpanzees and gorilla, which is the likely cause of the low bootstrap support found at this node.

Low bootstrap support is also observed at the node separating the horse-dog lineage. Despite progress over the last decade (for example, Beninda-Emonds et al., 2007), there are portions of the mammalian phylogeny that remain unresolved. There are competing hypotheses regarding the phylogenetic relationships between Cetartiodactyla (cow), Carnivora (dog) and Perissodactyla (horse). Some studies support the ((cow, horse), dog) hypothesis (for example, Murphy et al., 2001a,) while other studies support the ((horse, dog), cow) hypothesis (for example, Murphy et al., 2001b). Other studies have encountered similar issues regarding the phylogeny of this clade (for example, Kullberg et al., 2006, Hou et al., 2009). Thus, it was unsurprising that this node obtained a low bootstrap support value.

Table 3.5 Number of singletons in each supermatrix.

Column 2 indicates the number of singletons in the original unperturbed supermatrix. Column 3 indicates the number of singletons in the supermatrix, after partial families have been removed. Column 4 indicates the number of singletons in the supermatrix, after partial families have been removed, and the supermatrix has been improved with Gblocks.

Species	Supermatrix, no partial fams removed	Supermatrix, partial fams removed	Supermatrix, partial fams removed and Gblocks
Anole	21,897	6,277	2,677
Chicken	13,095	3,117	683
Chimpanzee	2,990	482	32
Cow	8,265	2,495	599
Dog	138,953	2,064	249
Elephant	22,015	1,841	972
Finch	6,595	1,933	912
Gorilla	4,871	2,030	773
Guinea Pig	6,096	2,316	738
Horse	3,383	805	289
Human	4,053	117	21
Macaque	7,997	2,796	772
Marmoset	6,363	1,178	428
Mouse	7,141	942	203
Opossum	14,437	3,285	1,370
Orangutan	4,988	1,946	298
Giant Panda	6,308	864	128
Pig	27,296	8,740	518
Platypus	32,566	13,332	2,556
Polar Bear	37,896	3,965	162
Rabbit	6,937	3,048	545
Rat	12,847	2,355	711

Table 3.6 Optimal model of amino acid substitution as selected by ModelGenerator.

JTT was the optimum model of amino acid substitution selected for all sections of the alignment, regardless of model selection criteria (AIC, AICc or BIC). In addition, almost identical results were obtained using a second algorithm of model selection, ProtTest, or using supermatrices that had been improved using different improvement softwares (TrimAl or BMGE).

	10,000-20,000	**30,000-40,000**	**50,000-60,000**	**70,000-80,000**	**90,000-100,000**
AIC	JTT + I + G	JTT + I + G + F	JTT + I + G + F	JTT + I + G + F	JTT + I + G + F
AICc	JTT + I + G	JTT + I + G + F	JTT + I + G + F	JTT + I + G + F	JTT + G + F
BIC	JTT + I + G	JTT + I + G + F	JTT + I + G + F	JTT + I + G	JTT + G

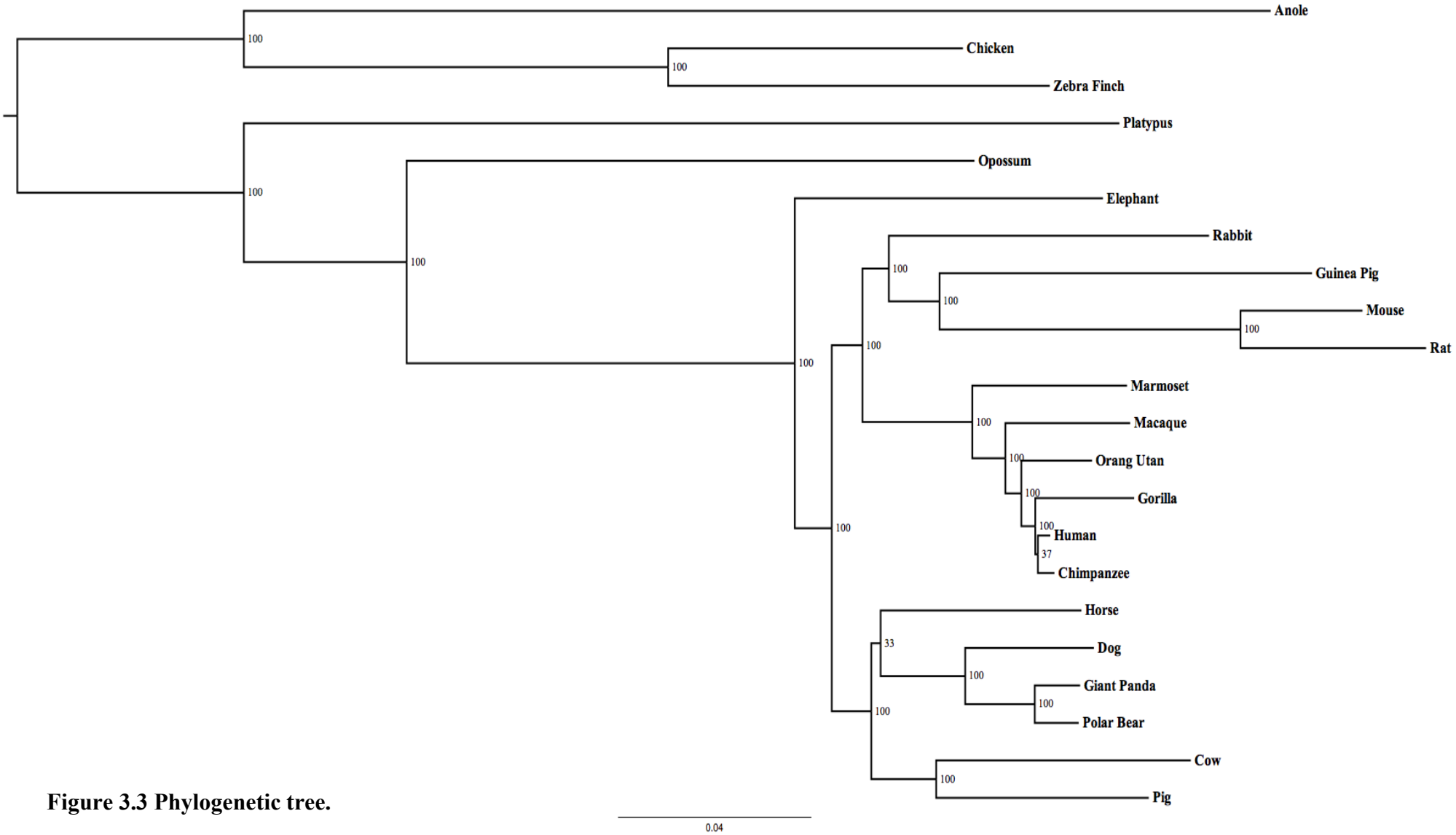


Figure 3.3 Phylogenetic tree.

The phylogeny reconstructed using the supermatrix. The value assigned to each node indicates a bootstrap support value.

3.3.2 Estimation of divergence times

Two MCMC chains were constructed "under the prior" to ensure that the distributions of the uncalibrated nodes were sufficiently wide before any posterior sampling was performed. Examining the resulting chronograms, all of the uncalibrated nodes exhibit an extremely wide time span, generally encompassing approximately 60-90 million years (Electronic Appendix 3.1; Table 3.7). In addition, I examined how often the bounds set by the calibrations were broken. Perhaps if any of the bounds were consistently broken (for example, 80-90% of the time), this may be indicative of incompatible calibration ranges. The upper bounds of the calibrations were generally broken 1.3-19.56% of the time, while the lower bounds were breached in the range of 6.2-17.62% of the time (Table 3.8).

Following the interpretation of Bayes Factors described by Kass and Raftery (section 1.5), none of the relaxed molecular clock models appeared to be particularly more suited to the data than any other model (Appendix 3.1). As such, all three relaxed clock models were used to calculate a divergence date between the species. From the chronograms, I examined the divergence estimates calculated for giant panda-polar bear node (Electronic Appendix 3.9). As can be observed from Table 3.9, the estimation is not narrow in range, spanning approximately 10-26 MY using the autocorrelated relaxed clock models, and 6-24 MY using the uncorrelated gamma model. This was disappointing, as the range estimated by these relaxed molecular clocks actually encompasses the estimates from fossil data (12 MY, Thenius, 1979, quoted in Krause et al., 2008) and mitochondrial sequence data (~19 MY, Krause et al., 2008). Next, I examined if the other uncalibrated nodes were similarly wide in range, and if they agreed with previous findings. Generally, the uncalibrated nodes encompassed ~59-93 MY (using a correlated molecular clock model) and ~58-112 MY (using an uncorrelated molecular clock model) (Table 3.10). Comparing these dates to previous observations proved to be a difficult task, due to the incompleteness of the fossil record, and

the debate between estimates obtained using molecular and fossil data. It is for these reasons that Benton et al. (2009) currently calibrates all of the nodes to a wide range of ~61.5-131.5 MY, encompassing ~70 MY. For example, the rabbit-human divergence is synonymous with the latest branching point between Primata and Rodentia. At present, fossil data for this branching point does not exceed 65.2 MY (Benton et al., 2009). However, molecular data have disputed this estimate (for example, Kitazoe et al., 2007). The divergence obtained for this particular node in this investigation (~69-85 MYA) agrees that the estimated timing of divergence appears to be slightly older than that calculated from the fossil record. The rabbit-mouse divergence is synonymous with the clade Glires that consists of the orders Rodentia and Lagomorpha. The date of divergence is traditionally estimated to have occurred ~65 MYA, or younger (Benton et al., 2009). Thus, the estimation of 50-70 MY for this particular node obtained in this investigation appears to broadly agree with the fossil record. However, such inferences should be made with extreme caution.

Table 3.7 Divergence estimate of uncalibrated nodes (under prior).

This table shows the divergence range for the uncalibrate nodes, estimated from the construction of MCMC chains under the prior.

Species	Species	Range of Estimate (MY)	Years encompassed (MY)
Rabbit	**Mouse**	53 – 129	~60
Rabbit	**Human**	58 – 147	~90
Rabbit	**Bear**	68 – 160	~90
Rabbit	**Elephant**	68 – 160	~90
Cow	**Dog**	62 – 150	~90
Dog	**Bear**	8 – 104	~70
Polar Bear	**Giant Panda**	0.9 – 71	~70

Table 3.8 Output from construction of MCMC chains under the prior.

This table shows the percentage of times that the bounds of the calibrations provided were broken, from constructing MCMC chains under the prior. The prior was conducted in parallel, producing "prior 1" and "prior 2" results.

			Prior 1 (calibration file)		Prior 2 (calibration file)	
	Upper bound	**Lower bound**	**% Overflow**	**% Underflow**	**% Overflow**	**% Underflow**
Mouse - Rat	14	10	14.9	14.74	14.05	14.13
Chicken - Finch	86.5	66	19.56	11.08	18.78	10.73
Human - Chimp	10	5.7	12.26	16.17	12.25	16.66
Human - Orangutan	33.7	11.2	1.3	7.83	1.48	8.41
Human - Macaque	34	23.5	18.74	9.26	17.99	9.76
Human - Marmoset	65.8	33.7	17.48	6.2	16.57	6.66
Pig - Cow	65.8	52.4	15.91	11.94	16.02	12.31
Lizard - Platypus	330.4	312.3	14.77	14.46	14.94	14.54
Mouse - Guinea Pig	58.9	52.5	12.32	17.05	12.74	17.62
Lizard - Chicken	299.8	255.9	13.36	15.39	13.12	15.02
Human - Opossum	171.2	124	7.94	14.26	8	14.19
Human - Platypus	191.1	162.9	17.6	10.48	17.71	10.87
Horse - Dog	131.5	61.5	6.01	15.72	6.06	14.98

Table 3.9 Molecular dating analysis – polar bear/giant panda node.

The first row indicates the molecular dating estimate for the polar bear-giant panda divergence once phylobayes is conducted "under the prior". The remaining rows indicate the molecular dating analysis, with JTT as the model of amino acid substitution for each of the relaxed clock models.

Model	Relaxed clock model	Run	Total # Cycles	Burn In	Min Age	Max Age
Under the prior	CIR	1	10,040	1,000	0.96	71.63
	CIR	2	9,624	1,000	0.85	71.76
JTT	CIR	1	7,526	700	10.33	23.31
	CIR	2	7,542	700	10.33	23.3
	Ln	1	6,440	600	12.74	26.73
	Ln	2	6,420	600	12.85	25.95
	UGam	1	5,038	500	6.29	24.35
	UGam	2	5,024	500	6.15	23.9

Table 3.10 Molecular dating analysis – other uncalibrated nodes.

Species	Species	Correlated mol. clock model (MY)	Uncorrelated mol. clock model (MY)
Rabbit	**Mouse**	59 – 70	58 – 78
Rabbit	**Human**	63 – 77	63 – 88
Rabbit	**Bear**	67 – 85	69 – 97
Rabbit	**Elephant**	75 – 93	76 – 112
Cow	**Dog**	60 – 75	60 – 84

There are a number of parameters throughout the course of this molecular dating analysis that, if altered, may impact the date of divergence inferred in this study. Two of these potential factors are the amino acid substitution model selected and the set of calibrations provided. Each of these factors was examined in turn, to see what impact, if any, alteration of the parameter had on the overall estimate of divergence.

JTT was unanimously selected as the model of amino acid substitution. In all sections of the alignment, variants of the JTT model composed the six best-fit models, according to AIC, AICc and BIC (Table 3.11). I asked how much the divergence estimates of the uncalibrated nodes of the phylogeny change, if the model of amino acid substitution is swapped for a less optimal model. However, the polar bear-giant panda divergence remained in the region of 10-27 MY with a correlated relaxed clock model, and 6-25 MY with an uncorrelated clock model, regardless of the model of amino acid substitution selected (Table 3.12). A similar pattern emerged from examination of the other calibrated nodes (Table 3.13). For example, the original dating analysis suggested that the rabbit-mouse divergence occurred approximately 59 to 70 MYA. From all the alternative amino acid substitution models selected, each correlated relaxed clock models estimate a divergence of ~58-70 MYA, while the uncorrelated model estimated a wider range of 58-78 MYA in all cases. (Table 3.13). Similarly, the date that is traditionally assigned to the rabbit-human node is approximately 65.2 MYA, based on fossils (Benton et al., 2009), and a variety of older and younger dates, based on molecular data. The estimates obtained from the correlated relaxed clock models display a wider range, encompassing 63-76 MY, while the uncorrelated relaxed clock models exhibit an even wider range again, in the region of 63-86 MYA (Table 3.13). Thus, I concluded that alteration of the model of amino acid substitution offers no extra insight into the divergence that the selection of the original model of substitution could not.

Accurate estimation of the divergence between polar bear and giant panda partially depended on the ability to accurately calibrate other surrounding nodes. However, there was an inherent limitation to this study, as six of the internal nodes are currently calibrated with the same wide range, encompassing almost 70 MY (61.5-131.5 MYA; Table 3.3). Thus, I repeated the molecular dating analysis five more times. Each time, all the parameters remained constant with the exception of a single calibration. The aim of this exercise was to understand what effect changing a single, potentially problematic, calibration had on the inferred estimates of divergence. The original estimate was approximately 10-29 MYA (using a correlated relaxed clock model) and 6-24 MYA (using an uncorrelated relaxed clock model). As can be observed from Table 3.13, the dates of divergence between giant panda and polar bear (or the other calibrated nodes) did not substantially alter, dependent upon the exact combination of calibrations used.

Table 3.11 First six models of amino acid substitution, for all sections of supermatrix, as selected by ModelGenerator, according to the AIC, AICc and the BIC criterion.

		AIC	AICc	BIC
10,000-20,000	**Model 1**	JTT+I +G + F	JTT+I+G	JTT+I+G
	Model 2	JTT+I+G	JTT+I+G+F	JTT+G
	Model 3	JTT+G+F	JTT+G	JTT+I+G+F
	Model 4	JTT+G	JTT+G+F	JTT+G+F
	Model 5	JTT+I+F	JTT+I	JTT+I
	Model 6	JTT+I	JTT+I+F	JTT+I+F
30,000-40,000	**Model 1**	JTT+I+G+F	JTT+I+G+F	JTT+I+G+F
	Model 2	JTT+G+F	JTT+G+F	JTT+G+F
	Model 3	JTT+I+G	JTT+I+G	JTT+I+G
	Model 4	JTT+G	JTT+G	JTT+G
	Model 5	JTT+I+F	JTT+I+F	JTT+I+F
	Model 6	JTT+I	JTT+I	JTT+I
50,000-60,000	**Model 1**	JTT+I+G+F	JTT+I+G+F	JTT+I+G+F
	Model 2	JTT+G+F	JTT+G+F	JTT+G+F
	Model 3	JTT+I+G	JTT+I+G	JTT+G
	Model 4	JTT+G	JTT+G	JTT+I+ G
	Model 5	JTT+I+F	JTT+I+F	JTT+I+F
	Model 6	JTT+I	JTT+I	JTT+I
70,000-80,000	**Model 1**	JTT+I+G+F	JTT+I+G+F	JTT+I+G
	Model 2	JTT+G+F	JTT+I+F	JTT+I+G+F
	Model 3	JTT+I+G	JTT+F+G	JTT+G
	Model 4	JTT+G	JTT+G	JTT+F+G
	Model 5	JTT+I+F	JTT+I+F	JTT+I
	Model 6	JTT+I	JTT+I	JTT+I+F
90,000-100,000	**Model 1**	JTT+I+G+F	JTT+G+F	JTT+G
	Model 2	JTT+G+F	JTT+I+G+F	JTT+I+G
	Model 3	JTT+G	JTT+G	JTT+ G+F
	Model 4	JTT+I+G	JTT+I+G	JTT+I+G+F
	Model 5	JTT+I+F	JTT+I+F	JTT+I
	Model 6	JTT+I	JTT+I	JTT+I+F

Table 3.12 Molecular dating analysis, conducted with alternative models of amino acid substitution.

This table describes the divergence estimates for the polar bear-giant panda node.

Model	Dating model	Run	Total # cycles	Burn In	Min Age	Max Age
WAG	CIR	1	6,284	600	10.33	23.96
	CIR	2	6,247	600	10.02	23.31
	Ln	1	6,439	600	12.81	26.37
	Ln	2	6,455	600	12.98	26.41
	UGam	1	9,431	900	6.15	23.82
	UGam	2	9,384	900	6.24	24.63
GTR	CIR	1	4,531	400	10.41	23.25
	CIR	2	4,472	400	10.28	23.69
	Ln	1	4,707	400	12.81	26.34
	Ln	2	4,706	400	12.91	25.6
	UGam	1	3,864	400	6.16	24.19
	UGam	2	3,709	400	6.13	24.07
LG	CIR	1	5,419	500	10.23	23.17
	CIR	2	5,376	500	10.45	23.32
	Ln	1	5,635	500	12.85	27.07
	Ln	2	5,609	500	12.71	26.4
	UGam	1	4,272	400	6.21	24.72
	UGam	2	4,463	400	6.22	24.54

Table 3.13 Molecular dating analysis, conducted with alternative models of amino acid substitution.

This table describes the divergence estimates for other uncalibrated nodes.

Species	Species	Correlated mol. clock model (MY)	Uncorrelated mol. clock model (MY)
Rabbit	**Mouse**	58 – 70	58 – 78
Rabbit	**Human**	63 – 77	63 – 86
Rabbit	**Bear**	68 – 85	70 – 98
Rabbit	**Elephant**	74 – 95	74 – 114
Cow	**Dog**	60 – 75	61 – 84

Table 3.14 Molecular dating analysis with alternative calibrations.

In column 1, each alternative calibration corresponds to a particular calibration that is described in Table3.3. Each alternative calibration was conducted in duplicate, generating run 1 and 2 (column 3). Column 3, 4 and 5 indicate the total number of PhyloBayes cycles, the burn in and the divergence estimate for the polar bear-giant panda node, for each model and calibration.

Calibration #	Dating model	Run	Total # Gens	Burn In	Min Age	Max Age
2 (Cow –Dog 131.5 – 61.5 MYA)	**CIR**	1	5,015	500	10.29	23.1
	CIR	2	5,008	500	10.55	17.7
	Ln	1	5,088	500	10.65	18.21
	Ln	2	5,121	500	10.66	18.35
	UGam	1	6,499	600	6.14	24.02
	UGam	2	6,468	600	6.2	23.82
3 (Human-Elephant 131.5-61.5 MYA)	**CIR**	1	5,026	500	10.48	17.96
	CIR	2	5,051	500	10.18	23.65
	Ln	1	5,143	500	13.01	27.41
	Ln	2	5,151	500	12.99	27.53
	UGam	1	6,549	600	6.2	24.2
	UGam	2	6,398	600	6.3	23.71
4 (Human-Horse 131.5 – 61.5 MYA)	**CIR**	1	4,818	1000	10.43	23.52
	CIR	2	4,786	1000	10.21	23.1
	Ln	1	4,872	500	12.85	26.49
	Ln	2	4,717	500	12.72	26.29
	UGam	1	6,194	600	6.13	24.62
	UGam	2	6,055	600	6.02	24.02
5 (Human-Rabbit 131.5-61.5 MYA)	**CIR**	1	4,744	500	10.39	23.31
	CIR	2	4,234	500	10.48	23.19
	Ln	1	4,699	500	12.64	26.23
	Ln	2	4,025	500	12.64	26.2
	UGam	1	5,658	600	6.04	24.4
	UGam	2	5,394	600	5.98	24.32
6 (Rabbit-Guinea Pig 131.5 – 61.5 MYA)	**CIR**	1	3,160	500	10.39	23.61
	CIR	2	3,178	500	10.42	23.86
	Ln	1	3,233	250	12.72	26.67
	Ln	2	3,203	250	12.89	26.34
	UGam	1	2,503	250	6.15	23.78
	UGam	2	2,464	250	6.19	24.36

Table 3.15 Molecular dating analysis with alternative calibrations (other nodes).

In column 1, each alternative calibration corresponds to a particular calibration that is described in Table 3.3. Columns 3 and 4 indicate the min and max divergence estimate calculated for the rabbit-mouse and rabbit-human node.

Species	Species	Correlated mol. clock model (MY)	Uncorrelated mol. clock model (MY)
Rabbit	**Mouse**	58 – 70	58 – 80
Rabbit	**Human**	61 – 78	63 – 88
Rabbit	**Bear**	66 – 86	70 – 98
Rabbit	**Elephant**	74 – 95	74 – 114
Cow	**Dog**	60 – 76	58 –8 6

3.4 Discussion

The objective of the investigation described in this chapter was to acquire an accurate estimation of the timing of divergence between polar bear and giant panda. In turn, this would provide a valuable calibration point in future explorations that aim to decipher the timing of ursine bear speciation events. Unlike previous molecular sequence analyses that depended on minute amounts of nuclear sequence data (for example, Pagès et al., 2008) or mitochondrial data (Krause et al., 2008), this study combined whole nuclear genomes into a Bayesian setting to understand the speciation event from a novel perspective.

Unfortunately, the divergence estimate that was eventually obtained for the node of interest is too wide to provide a useful calibration point between polar bear and giant panda (6-27 MYA). In fact, this range encompasses the 12-22 MY range that was obtained in previous studies (Thenius, 1979, Krause et al., 2008). I considered the factors that may have attributed to this result. First, I considered the data set itself. The meteoric rise in vertebrate genome sequence availability is commonly cited as a panacea for resolving difficult phylogenetic problems, including the resolution of the bear phylogeny (for example, Yu et al., 2004, Pagès et al., 2008). However, Milinkovitch (2010) argued that even with the inclusion of more data, it would remain difficult to differentiate between sequencing and assembly artifacts from true changes in the mode and tempo of evolution until there is better homogeneity in both taxon sampling and genome quality in all genomes. Data quality as a potential issue was considered prior to the commencement of this experiment, and a number of steps were undertaken in order to alleviate this issue as much as possible. For example, genes that were unlikely to encode functional proteins, as their sequence length was not a multiple of three or contained more than one stop codon, were removed. In addition, I used protein sequence data instead of nucleotides, to attempt to lessen the possibility of undetected saturation in the sequences. To alleviate potential problems with uneven taxon sampling in the

supermatrix, I removed all partial single copy orthologous gene families. Finally, Gblocks was performed on every homologous gene family alignment to remove ambiguously aligned or particularly divergent regions.

Singletons may indeed arise from processes such as positive selection. For example, a species may possess a unique amino acid at a particular point in a sequence that allows it to thrive in a particular environment, or evoke a particular immune response. In this analysis, singletons were used as a rough indication of genome sequence quality. As can be observed in Table 3.5, some species, including those suspected to be of low sequencing or assembly quality, exhibit suspiciously high numbers of singletons. Thus, it is possible that there was a data quality issue in this experiment.

Once a supermatrix was obtained, the next step was to combine the resulting phylogenetic tree with various parameters in a Bayesian setting to estimate the divergence between species. The phylogenetic tree itself contained two nodes with low bootstrap support, the human-chimpanzee-gorilla clade and the horse-cow-dog clade. Low support values for both of these nodes were not unexpected, as the phylogenetic relationships that comprise these clades are regularly the source of contention among molecular systematists. The incomplete and flawed fossil record could undoubtedly have played a role in the poor estimation of divergence obtained in this study, as could the improper selection of model of amino acid substitution or relaxed molecular clock model. However, systematic alteration of all of these parameters did not appear to drastically change the date of divergence between polar bear and giant panda, or between other species in the tree, such as the rabbit-mouse divergence or rabbit-human divergence.

In summary, in spite of the precautions take to alleviate any potential issues surrounding the estimation of the divergence between polar bear and giant panda, the

elucidation of an accurate estimation of timing of divergence between the bear species remains an elusive task.

Chapter 4: An investigation into the forces governing synonymous codon usage in vertebrates

4.1 Introduction

Synonymous codon usage (SCU) is maintained by a balance between selective and neutral processes (Bulmer, 1991). In prokaryotes and eukaryotic non-vertebrates, natural selection acting at the level of translation is deemed to be the dominant force shaping SCU patterns (Ikemura, 1985, Shields et al., 1988, Duret, 2000). However, the slight selective advantage offered by alternative synonymous codons was historically thought to be overcome by neutral processes (such as mutational bias and genetic drift) in species in which selection is less efficient, such as vertebrates (Rao et al., 2011).

Possible evidence for translational selection in some vertebrates has recently been detected (Musto et al., 2001, Romero et al., 2003, Urrutia and Hurst, 2003). However, a systematic examination into the cause of vertebrate synonymous codon usage has not been conducted to date. This chapter commences with an introduction to the genetic code and synonymous codon usage. Subsequently, a study is described that combined newly sequenced genomes and novel gene expression data to examine the null hypothesis that translational selection is unable to overcome mutational bias and random genetic drift in vertebrates.

4.1.1 The genetic code

The genetic code guides the conversion of nucleotide triplets into a sequence of amino acids (Crick et al., 1961). Although slight variations exist (Fox, 1987, Knight et al., 2001), the standard genetic code asserts that there are 20 amino acids unambiguously encoded by 61 sense codons, and three termination codons. The genetic code is degenerate, meaning that most of the amino acids are encoded by between two and six "synonymous" codons. In the past, the exact synonymous codon encoding a particular amino acid was thought to have little physiological effect on a cell. However, it has been recently demonstrated that the choice of synonymous codon at a particular position in a sequence affects various cellular mechanisms including protein folding (Zhou et al., 2009) and protein function (Hudson et al., 2011). Understanding synonymous codon usage also has biomedical and biotechnological applications. For example, synonymous mutations are implicated in the progression of various human diseases (Sauna and Kimchi-Sarfaty, 2011) and considerably alter transgene production rates (Angov et al., 2008). So, although synonymous codons encode the same amino acid, it is clear that it is vital to understand the causes and effects of synonymous codon usage variation.

4.1.2 The Selection-Mutation-Drift theory

The current accepted model explaining synonymous codon usage bias is the "Selection-Mutation-Drift" theory (Bulmer, 1991). This theory proposes that natural selection favours optimal codons over non-optimal codons for each amino acid, while mutational bias and genetic drift allow non-optimal codons to persist. The selectionist component of the theory asserts that natural selection maintains SCU bias to ensure translational accuracy, translational efficiency or both. The translational accuracy hypothesis posits that preferentially used codons are those that are most likely to ensure that the encoded protein

matches the sequence prescribed by the encoding gene and folds properly within a cell (Akashi, 1994, Stoletzki and Eyre-Walker, 2007, Drummond and Wilke, 2008, Zhou et al., 2009). Such translational accuracy is imperative in order to avoid wasting energy and cytotoxic misfolding, both of which may be detrimental to the cell (Komar et al., 1999, Drummond and Wilke, 2009).

Translational efficiency describes how effectively cellular material is used in protein translation. The translational efficiency hypothesis posits that different synonymous codons are translated at different speeds due to disparities in codon selection time (i.e. the time needed for each codon to find a suitable tRNA anti-codon) (Bulmer, 1991). Faster translation allows ribosomes to spend less time bound to mRNA. In turn, this elevates the number of free ribosomes available and increases the number of mRNAs that can potentially be translated per ribosome. Alternatively, Qian et al. (2012) recently suggested that preferred codons are not translated faster than unpreferred codons. Instead, the selection coefficient for synonymous mutations to achieve codon-tRNA balance is greater in highly expressed than in lowly expressed genes.

The mutational bias component of the Selection-Mutation-Drift theory refers to the systematic nonuniformity in mutations that arise from DNA replication and repair processes. In essence, mutational bias suggests that the codon composition of a gene is purely reflective of the location in a genome in which it is found. For example, base mismatches that are introduced into mammalian cell lines are preferentially repaired to guanine or cytosine, causing a mutational bias that would lead to differential patterns of synonymous codon usage (reviewed in Marais, 2003).

The final component of the Selection-Mutation-Drift theory is random genetic drift, which may affect synonymous codon usage patterns, depending on how effectively natural selection impacts a species. Effective population size (N_e) is a concept that was originally

introduced into population genetics by Wright (1931) (quoted in Wang et al., 2008), who defined it as:

"*the number of breeding individuals in an idealized population that would show the same amount of dispersion of allele frequencies under random genetic drift or the same amount of inbreeding as the population under consideration*"

The effectiveness of selection on alternative synonymous codon choice is determined by the product of N_e and the selective advantage (s) of a codon over its alternatives at the same site (Bulmer, 1991). This selective advantage is typically weak, so a large N_e is required for translational selection to be effective relative to genetic drift (Dos Reis and Wernisch, 2009).

4.1.3 Synonymous codon usage bias in vertebrates

In vertebrates, it is thought that base composition is the most pervasive influence on synonymous codon usage (Rao et al., 2011). There are a number of reasons suggested to explain why natural selection may be unable to overcome neutral processes in vertebrates. For example, in humans, the effective population size is approximately 10,000 (Zhao et al., 2000). Conversely, the effective population size of *E. coli* (in which translational selection has been detected) is thought to be orders of magnitude larger (Charlesworth and Eyre-Walker, 2006). As described in the previous section, efficacy of natural selection is a function of effective population size. Thus, perhaps the apparent absence of translational selection in vertebrates can at least partially be explained by the lower effective population sizes in humans.

Vertebrate chromosomal regions are compartmentalized based on GC content (i.e. isochores) (Bernardi and Bernardi, 1986, Bernardi, 1995), For example, α- and β- globin

genes show substantially different codon usage patterns, because they are located in isochores with different levels of GC content, in spite of the fact that they are both highly expressed and perform a similar function (Bernardi et al., 1985). From this, one could conclude that variation in vertebrate codon usage is the result of the isochore structure. However, evidence that translational selection may overcome random genetic drift has recently been detected in a handful of vertebrate species (Musto et al., 2001, Romero et al., 2003, Urrutia and Hurst, 2003, Yang and Nielsen, 2008).

4.1.4 Examination of synonymous codon usage bias in vertebrates

Synonymous codon usage patterns within and between species have been effectively explored using an array of methods. First, a multitude of indices have been developed specifically for codon usage investigations and implemented in various software packages, such as codonW (Peden, 1997). Second, a number of genomic characteristics have been observed to correlate with synonymous codon usage, and statistics to quantify these genomic characteristics have been developed. Finally, multivariate techniques have been successfully employed to explore how codon usage patterns co-vary with other genomic features, such as base composition or gene expression. The synonymous codon usage bias indices, genomic statistics and multivariate techniques that were implemented in this investigation will be described in this section.

4.1.4.1 Codon usage bias indices

Commonly implemented codon usage bias (CUB) indices may broadly be classified into two categories based on whether or not a prior knowledge of preferred codons is required. The three indices selected for this investigation were relative synonymous codon usage (RSCU), effective number of codons (N_c) and codon deviation co-efficient (CDC). It

was not necessary to supply a reference gene set for any of the codon usage bias indices selected for this analysis.

RSCU measurements have previously provided insights into synonymous codon usage patterns in a diverse array of organisms (Sharp et al., 1988, Stenico et al., 1994, Duret and Mouchiroud, 1999). The RSCU value of a particular codon is defined as the observed number of codon occurrences divided by the number of occurrences that would be expected if the synonymous codons for the encoded amino acid were used uniformly (Sharp et al., 1986). For unbiased synonymous codon usage, the RSCU value is 1. For synonymous codon usage that is more infrequent than expected codon usage for an amino acid, the RSCU is less than 1, while the inverse is true for codon usage that is more frequent than expected.

The "effective number of codons" (N_c) (Wright, 1990) is a second index of codon bias that estimates, for those codons with a synonymous alternative, how many synonymous codons are effectively used in a gene to encode a particular amino acid. In cases of extreme bias, in which one codon is exclusively used for each amino acid, the N_c value is 20. Conversely, if all synonymous alternatives are being used equally frequently, the N_c is 61. Genes that are more biased in their base composition are expected to exhibit a lower N_c. Thus, plotting N_c values against base composition parameters and exploring the relationship between the two can identify genes that are more restricted in their codon usage than might be expected given their base composition.

Recently, it has become apparent that it is essential to incorporate differential base composition parameters into CUB estimation (Zhang and Yu, 2010). The codon deviation coefficient (CDC) quantifies CUB accounting for heterogeneous background nucleotide composition not only in sequences, but at each position of a codon (Zhang et al., 2012). The calculation of the CDC score is based on two vectors. Each vector represents the expected and observed codon usage of one gene. A distance metric is calculated that is based on the cosine

of the angles between the observed and expected vectors. CDC calculates a distance coefficient ranging from 0 (i.e. the observed and expected codon vectors are pointing in the same direction in vector space, so there is no bias observed) to 1 (i.e. the observed and expected vectors are pointing at 90° to each other in vector space, resulting in maximum bias). CDC also employs a bootstrapping technique (see section 1.5) to assess the statistical significance of the distance coefficient that is obtained.

4.1.4.2 Genomic characteristics that correlate with codon usage bias

Often, the synonymous codon usage patterns observed in a genome correlate with particular genomic characteristics. For example, as described previously, there is a relationship between the base composition of a gene, and N_c of a gene. This is because as a gene's codon bias increases (i.e. is more AT- or GC-rich), the number of available codons decreases. In this investigation, GC content and GC content at the third synonymous codon position (GC_3) were included in the analysis to explore the relationship between the base composition of a gene, and the number of codons used by a gene. The third positions of codons are subject to less constraint than the other two positions. So, if there is any mutational bias, it is likely to impact the positions with the least constraint (Grocock and Sharp, 2002).

4.1.4.3 Correspondence analysis

Correspondence Analysis (CA) is a multivariate technique that is commonly implemented to explore how variation in codon usage co-varies with other biological traits, such as expression level or GC content (Shields et al., 1988, Musto et al., 2001). CA functions by representing each gene as a vector. Each co-ordinate of the vector indicates a gene's usage of a particular synonymous codon. A series of orthogonal axes are created by each gene's vector to identify trends that explain the data variation, with each subsequent axis explaining a

decreasing amount of variation. In this way, the fraction of the variation that is accounted for by each axis may be identified.

There is debate over whether it is more appropriate to perform CA on relative frequencies such as RSCU values or on raw codon counts (Perrière and Thioulouse, 2002, Zavala et al., 2002). Using RSCU values to perform a CA avoids possible biases linked to amino acid composition. However, while the desire to remove amino acid effects is justified, the action may infer other more subtle biases in the process (Perrière and Thioulouse, 2002). One alternative strategy is to perform a CA on raw codon counts and RSCU values in parallel, and to compare the results. This approach has already been successfully applied to studies devoted to *Heliobacter pylori* (Lafay et al., 2000) and transposable elements (Lerat et al., 2000) and was the approach undertaken in this study.

4.1.4.4 Significance tests used in investigation

There were three statistical significance tests implemented at various stages of this investigation, each of which shall be described. The Spearman correlation is a non-parametric measure of the linear relationship between two data sets. The co-efficient spans from -1 to +1, with 0 indicating that there is no correlation. Positive correlations imply that as x increases, so does y. Conversely, negative correlations imply that as x increases, y decreases. The Wilcoxon rank-sum test examines the null hypothesis that two sets of measurements are drawn from the same distribution. The alternative hypothesis is that one set of values tends to be larger than the values in the other sample. The Mann-Whitney test is also a non-parametric statistical test that assesses whether one of two samples of independent observations tend to have larger values than the other set of samples. However, in the Mann-Whitney test, the samples are assumed to be unpaired.

4.1.4.5 Exploration of synonymous codon usage in verterbates

At this point, we know that translational selection has been identified in various species, and that there are several synonymous codon usage indices and statistical tests commonly invoked to denote the level of codon bias present in a gene. But specifically how are these indices and gene sequences combined to identify evidence of translational selection? First, as the selective advantage offered by synonymous codons is expected to be quite weak, the selective pressure to reduce the cost of translation is expected to have the greatest impact on highly expressed genes. In agreement with this model, a positive correlation between CUB and expression level has been observed in several eukaryotes, such as *S. cerevisiae*, *C. elegans*, *Arabidopsis thaliana* and *D. melanogaster* (Duret and Mouchiroud, 1999, Castillo-Davis et al., 2002). Thus, the first null hypothesis considered is that highly and lowly expressed genes exhibit the same level of codon bias.

The selection on synonymous codon positions is thought to lead to a co-adaptation of codon usage and tRNA content to optimize translation. Such a correlation has previously been detected in prokaryotes (Ikemura, 1985, Kanaya et al., 1999) and eukaryotes (Ikemura, 1985, Kanaya et al., 2001). The second null hypothesis that I examined is that there is no correlation between preferred codons, and the most abundant cellular tRNAs. In addition, vertebrate synonymous codon usage was explored in other ways. Wright (1990) suggested that plotting N_c against GC_3 could be used to understand codon usage among genes. If GC_3 were the sole determinant of codon usage variation among genes, N_c values would fall on a continuous curve. This is because as the mutational bias in a gene becomes less pronounced, N_c becomes greater. In addition, such a plot would allow for the identification of genes that are more biased in their codon usage than expected, given their GC_3 content. Finally, correspondence analyses were conducted to establish if there were any additional patterns that affect synonymous codon usage that were not captured in the other analyses.

4.2 Methods

4.2.1 Genomic data collection and data set assembly

A data set comprising 38 vertebrates was assembled. For each taxon, the longest canonical transcripts for all the protein-coding genes were retrieved from Ensembl version 66 (Flicek et al., 2012). The initial data set comprised 712,902 sequences. Genes that contained non-translatable regions were removed. In total, 558,871 genes were retained (Table 4.1). The next stage was to assign a level of expression to each vertebrate gene. Su et al. (2004) described the patterns of gene expression for approximately 44,000 human transcripts across 84 tissues. Using this data, Dr. David-Alvarez Ponce assembled a data set consisting of expression data for 17,738 human genes across 84 tissues (Electronic Appendix 4.1). In total, 16,634 of the human genes that were retrieved from the Ensembl database were present in Dr. Alvarez-Ponce's data set. For each of these 16,634 genes, an expression level was assigned to each gene as the highest expression level a gene attained across all 84 tissues.

It was necessary to assign gene expression data for each non-human vertebrate and identify a set of highly expressed genes for each species. Compared to our knowledge of gene and protein sequence evolution, inter-species patterns of gene expression conservation are poorly understood. Recent research suggests that core gene expression level has remained relatively conserved between orthologous genes from different vertebrate species (Liao and Zhang, 2006, Chan et al., 2009). Thus, in this study, it was assumed that a highly expressed gene in humans was also likely to be relatively equally highly expressed in other vertebrates. Putative vertebrate orthologs were identified for each human gene using a reciprocal best-hit similarity search. The BLASTN algorithm (Altschul et al., 1990) compared the human genome to every other vertebrate genome (and vice versa). The E-value cut-off was set to e^{-10}. A total of 16,420 human genes found a recriprocal best hit in at least one other vertebrate species (Electronic Appendix 4.2). The expression level of each human gene was transferred

to its ortholog. Highly and lowly expressed genes were identified as the 5% highest and lowest expressed human genes and their orthologs in the other vertebrate species (Table 4.2, Table 4.3).

Highly expressed genes tend to produce shorter proteins to reduce the cost of protein production (Moriyama and Powell, 1998, Duret and Mouchiroud, 1999, Marais and Duret, 2001). For each species, the average gene length of each set of highly and lowly expressed genes were calculated (Table 4.2, Table 4.3). In general among all species, the average highly expressed gene length was shorter (1,081 nucleotides) than that of lowly expressed genes (1,381 nucleotides). However, a second pattern emerged when the gene lengths of highly expressed genes were examined in greater detail. In chimpanzee, for example, the average gene length was 1,179 base pairs. As expected, this length is shorter than the average for chimpanzee lowly expressed genes (1,539 base pairs). However, the four longest gene sequences in the chimpanzee highly expressed gene set were 20,010, 13,689, 10,716 and 10,191 base pairs in length. Such exceptionally long genes do not exhibit a typically short gene length expected of vertebrate highly expressed genes. There is the chance that these genes may have been erroneously categorized as highly expressed. Thus, an additional filtering step was performed, in which the 5% longest of each set of genes were removed (Table 4.2, Table 4.3).

Table 4.1 Number of genes retained at each stage of filtering process

Species	Initial number of genes	Number of genes after filtering	Number of orthologs
American Pika	15,993	7,652	5,684
Anole	17,800	17,735	9,118
Armadillo	14,839	5,285	3,346
Cat	15,047	5,850	4,412
Chicken	16,929	16,471	9,036
Chimpanzee	18,770	18,471	14,968
Cow	20,036	19,786	14,648
Dog	19,335	19,088	14,344
Elephant	20,310	19,798	14,339
Finch	17,501	17,285	9,076
Flying Fox	17,023	10,087	8,090
Galago	19,517	19,238	14,622
Gibbon	18,631	17,232	14,170
Gorilla	21,354	20,674	15,133
Guinea Pig	18,699	18,494	13,914
Hedgehog	14,591	5,777	3,981
Horse	20,459	20,291	14,539
Human	21,922	21,672	-
Hyrax	16,090	6,526	4,865
Kangaroo Rat	15,829	7,189	5,357
Macaque	22,137	21,644	15,148
Marmoset	21,703	20,784	15,094
Microbat	19,751	19,467	13,469
Mouse	23,043	22,938	16,460
Mouse Lemur	16,319	7,341	5,448
Opossum	23,225	19,417	12,460
Orangutan	19,793	19,512	14,803
Panda	19,609	18,443	14,045
Pig	20,192	17,097	11,769
Platypus	17,637	17,566	9,755
Rabbit	21,888	18,811	13,553
Rat	22,533	19,216	13,899
Sloth	23,360	4,561	3,036
Squirrel	12,425	5,187	3,543
Tarsier	14,828	5,791	4,254
Tasmanian Devil	18,302	13,646	11,745
Tenrec	18,920	7,016	4,287
Wallaby	16,562	5,833	3,735
Total	**712,902**	**558,871**	**394,745**

Table 4.2 Assembly of sets of highly expressed genes

Highly expressed genes				
	Original		Trimmed	
	# Genes	Gene Length (bp)	# Genes	Gene Length (bp)
American Pika	382	821.21	361	726.41
Anole	429	1348.97	399	1015.56
Armadillo	269	667.99	247	578.28
Cat	326	743.46	304	627.54
Chicken	431	1387.67	399	1033.08
Chimpanzee	793	1179.36	760	958.16
Cow	753	1239.03	722	998.37
Dog	732	1293.94	684	993.98
Elephant	736	1254.24	703	1023.03
Finch	430	1298.18	418	1085.75
Flying Fox	502	908.07	475	791.11
Galago	733	1259.02	703	1015.67
Gibbon	768	1199.58	722	966.42
Gorilla	810	1258.78	760	973.24
Guinea Pig	703	1254.9	665	983.31
Hedgehog	285	728.03	266	639.54
Horse	734	1229.86	703	985.59
Human	819	1191.52	779	946.63
Hyrax	339	728.82	323	653.9
Kangaroo Rat	357	794.64	342	705.96
Macaque	792	1201.6	741	950.58
Marmoset	796	1262.2	741	978.24
Microbat	694	1234.37	665	995.63
Mouse	741	1305.67	703	1032.6
Mouse Lemur	389	792.21	361	666.63
Opossum	642	1367.67	611	1018.81
Orangutan	800	1209.51	570	1049.16
Panda	798	1275.35	760	947.83
Pig	602	1136.46	570	919.54
Platypus	485	1224.14	456	936.76
Rabbit	704	1274.98	665	1011.18
Rat	714	1288.84	684	1034.62
Sloth	227	634.12	209	559.94
Squirrel	267	658.87	247	580.18
Tarsier	283	700.91	266	615.3
Tasmanian Devil	575	1305.33	551	1082.76
Tenrec	329	698.71	304	606.14
Wallaby	253	725.21	247	680.72
Average		**1081.14**		**878.11**

Table 4.3 Assembly of sets of lowly expressed genes

Lowly expressed genes				
	Original		Trimmed	
	# Genes	Gene Length (bp)	# Genes	Gene Length (bp)
American Pika	303	1219.07	285	1106.69
Anole	313	1656.2	304	1553.92
Armadillo	222	943.69	209	871.21
Cat	254	1001.63	247	955.28
Chicken	304	1686.07	285	1461.48
Chimpanzee	718	1539.91	684	1354.72
Cow	660	1558.61	627	1382.47
Dog	649	1530.52	608	1320.26
Elephant	668	1495.71	627	1302.88
Finch	311	1587.31	304	1492.93
Flying Fox	387	1314.71	361	1165.53
Galago	689	1574.06	646	1366.24
Gibbon	621	1567.41	589	1377.07
Gorilla	721	1552.7	684	1367.81
Guinea Pig	631	1524.35	608	1377.42
Hedgehog	234	1046.38	228	997.46
Horse	657	1476.52	627	1305.04
Human	818	1529.68	779	1346.81
Hyrax	271	1127.47	266	1078.34
Kangaroo Rat	288	1116.64	266	1000.68
Macaque	688	1513.5	646	1314.77
Marmoset	697	1536.8	665	1379.72
Microbat	570	1523.62	551	1374.27
Mouse	315	1077.57	304	1011.27
Mouse Lemur	675	1626.63	646	1431.77
Opossum	564	1559.8	538	1317.53
Orangutan	672	1484.58	513	1414.59
Panda	678	1520.42	646	1334.31
Pig	533	1417.57	513	1272.91
Platypus	365	1368.93	342	1171.18
Rabbit	594	1499.03	570	1355.66
Rat	633	1593.44	608	1428.07
Sloth	184	997.5	171	877.09
Squirrel	184	987.55	171	874.32
Tarsier	213	1034.58	209	1001.5
Tasmanian Devil	496	1572.94	475	1400.87
Tenrec	231	1022.34	228	990.66
Wallaby	205	1099.79	190	985.18
Average		**1381.19**		**1240**

4.2.2 Identification of translational selection in vertebrates

As discussed in the introduction, there are two genomic characteristics that are indicitive of translational selection in a species. First, I examined if each set of highly expressed genes was more biased in codon usage pattern than each set of lowly expressed genes. Subsequently, I asked if the preferred codons for an amino acid in each set of highly expressed genes correlated with the most abundant tRNA genes. In the following sections, each of these questions are addressed in turn.

4.2.2.1 Calculation of codon bias in highly and lowly expressed genes

This section examined whether codon bias for each set of highly expressed genes was more pronounced than the bias exhibited by each set of lowly expressed genes. The CUB of a gene was calculated using the CDC index that is implemented in the Compositional Analysis Toolkit v. 1.0 (http://cbrc.kaust.edu.sa/CAT/). As it has been previously demonstrated that certain CUB indices are affected by gene length (Urrutia and Hurst, 2001), the level of codon bias in each set of highly and lowly expressed genes was calculated in duplicate. In one case, gene length was explicitly strictly controlled for, and in the other case, it was not. Each case will be described separately.

First, for each species, I obtained a sub-set of highly and lowly expressed genes in which each highly expressed gene was paired to lowly expressed gene of exactly equal length. If there was more than one lowly expressed gene of equal length to a highly expressed gene (or vice versa), one of the genes of that particular length was selected at random. For each gene in each sub-set of highly and lowly expressed genes, a CDC index was calculated, using the standard genetic code and setting the bootstrap parameter to 1,000 bootstraps. If a gene obtained a P-value that was greater than 0.05, the gene was deemed to have exhibited a statistically insignificant codon bias. As Zhang et al. (2012) alluded to, and as observed in this

study, genes that displayed statistically insignificant P-values tended to be extremely short. Pairs of genes that did not obtain statistically significant CDC scores were removed. An average CDC score was calculated for the remaining significantly biased genes in each sub-set of highly and lowly expressed genes.

To investigate whether the CDC scores of each set of highly and lowly expressed genes for a species were significantly differently distributed from each other, a Wilcoxon signed rank test implemented in R package "Wilcoxon Rank Sum and Signed Rank Tests" was conducted (http://stat.ethz.ch/R-manual/R-patched/library/stats/html/wilcox.test.html). For this test, a set of CDC scores for highly and lowly expressed genes for a species were provided, and the "paired" parameter was set to true. The analysis described in the previous paragraph was subsequently repeated, with the exception that gene length was not explicitly controlled for. In this case, the full set of highly and lowly expressed gene sequences was provided to the CDC algorithm. For each gene, a CDC index was calculated. As in the previous paragraph, the standard genetic code was selected, and the bootstrap parameter was set to 1,000 bootstraps. Genes that did not obtain a statistically significant CDC index in the bootstrap procedure of the CDC algorithm were removed. For the remaining genes in each set of highly and lowly expressed genes, an average CDC index was calculated. A Wilcoxon rank sum test implemented in the R package "Wilcoxon Rank Sum and Signed Rank Tests" was conducted, using the same package as described in the previous paragraph. The set of CDC indices obtained for highly and lowly expressed genes were provided, and the "paired" parameter was set to "false". This is equivalent to the Mann-Whitney test.

4.2.2.2 Correlation of codon bias with tRNA abundance

This section examined if the preferred codons in a set of highly expressed genes from each species correlated with the most abundant tRNA genes in a genome. Codons encoding methionine, tryptophan and termination codons were excluded from this analysis. First, for each species, a set of preferred codons were identified. To achieve this, RSCU values for each amino acid were calculated for each set of highly expressed orthologs (and also in lowly expressed orthologs) using codonW version 1.4.4. The preferred codon used for each amino acid was defined as the one that obtained the highest RSCU value. Next, for 23 of the species in this study, the number of tRNA genes for each codon was retrieved from the Genomic tRNA database (retrieved April 18th 2012) (Chan and Lowe, 2009). Subsequently, using script 4.1 in script index 4.0, I examined if the preferred codon for each amino acid (i.e. the codon that obtained the highest RSCU value) matched the most abundant tRNA gene for that amino acid in each species.

4.2.3 Exploration of variation of synonymous codon usage in vertebrates

4.2.3.1 Examination of effective number of codons in genes

As suggested by Wright (1990), a plot of N_c versus GC_3 can be used as a preliminary exploration of codon heterogeneity. In such a plot, a normal distribution indicates the expected N_c value due to solely mutational bias. For the genes of each vertebrate species, N_c and GC_3 content were calculated using the codonW program. N_c was plotted against GC_3 using Python version 2.6.5 and the Pylab module implemented in matplotlib package version 1.1.1 (script 4.2 in the script index).

The distribution of genes in each plot was examined. Genes that were relatively balanced in their GC_3 content (45 – 55%), but biased in their codon usage ($N_c < 30$), were extracted for further investigation (using script 4.3 in script index 4.0). There were 193 genes

in total that displayed this pattern. For each of these genes, it's human ortholog was identified. A biological enrichment analysis was conducted using FatiGO (FatiGO is described in section 2.2.6). The set of 193 human orthologs was compared against the rest of the genome. Duplicate genes were removed within and between lists. Overrepresentation in cellular component, molecular function and biological process GO categories were searched for, using the "direct transmission" parameter. Fisher's exact test (for overrepresentation in list 1) was implemented, and an adjusted P-value of less than 0.05 was deemed significant. However, the 193 genes were not overrepresented in any GO terms. The N_c boundary was adjusted to 40. 1,326 genes were extracted (again, using script 4.3 in script index 4.0) that possessed a GC_3 content of 45-55% and an N_c less than 40. The FatiGO analysis described in this paragraph was repeated using almost identical parameters. The only exception was that enrichment of genes in biochemical pathways from the KEGG (Kanehisa, 2004) and Reactome (Joshi-Tope et al., 2005) databases were additionally searched for. For these 1,326 genes, the number of amino acids in each gene were calculated (in addition to the N_c and GC_3 values that had previously been calculated) and compared to the same genomic characteristics were calculated for the full set of genes from each species.

Table 4.4 Number of interesting genes per species extracted from GC_3 versus N_c plots

Species	# Genes
American Pika	13
Anole	89
Armadillo	26
Cat	16
Chicken	58
Chimpanzee	32
Cow	12
Dog	13
Elephant	19
Finch	45
Flying Fox	18
Galago	15
Gibbon	25
Gorilla	72
Guinea Pig	23
Hedgehog	26
Horse	14
Human	61
Hyrax	10
Kangaroo Rat	18
Macaque	60
Marmoset	58
Microbat	8
Mouse	59
Mouse Lemur	21
Panda	19
Opossum	24
Orangutan	68
Pig	39
Platypus	38
Rabbit	14
Rat	89
Sloth	21
Squirrel	25
Tarsier	25
Tasmanian Devil	110
Tenrec	18
Wallaby	25

4.2.3.2 Correspondence analysis

An extensive exploration into the source of codon usage variation among genes can be achieved using multivariate statistical analysis. Codons that encode methionine and tryptophan were removed as these two amino acids do not have synonymous alternative codons. In addition, the rarity of cysteine is thought to affect the outcome of CA and so these codons were removed (Perrière and Thioulouse, 2002). Finally, stop codons were removed. Thus, 57 codons were used in this analysis. A correspondence analysis was conducted in codonW for the full set of genes belonging to each species. This analysis was carried out in duplicate. In one case, RSCU values were used as the input data and in the other case, raw codon counts was used as the input data. The fraction of variation accounted for by each axis was identified from the ".summary.coa" output from codonW.

For each species, axis 1 versus axis 2 of each CA was plotted (using script 4.4 script index 4.0). Unusual clusters of non-highly expressed genes in anole, armadillo, horse and Tasmanian devil were identified on the resulting CA plots of axis 1 versus axis 2, using both RSCU values and raw codon counts. An example of such an unusual cluster may be observed in Figure 4.1. I wanted to examine what was unique about the codon usage of the genes composing these clusters, and why they were only present in certain genomes. To extract these genes for further investigation, it was necessary to locate their co-ordinates on the CA plot of axis 1 versus axis 2. This was achieved through visual inspection of the CA plots. The co-ordinates of the unusual genes from each plot (using RSCU data) of axis 1 versus axis 2 are described in Table 4.4. Using script 4.5 in script index 4.0, the sequences composing these clusters were extracted and visually inspected. Finally, two other sets of plots were constructed. First GC_3 was plotted against axis 1 for each species using an adapted version of script 4.2 in script index 4.0. Linear correlation was calculated using a Spearman correlation

co-efficient, implemented in SciPy. In addition, axis 2 was plotted against the expression level for each species, to explore any potential correlation between the two.

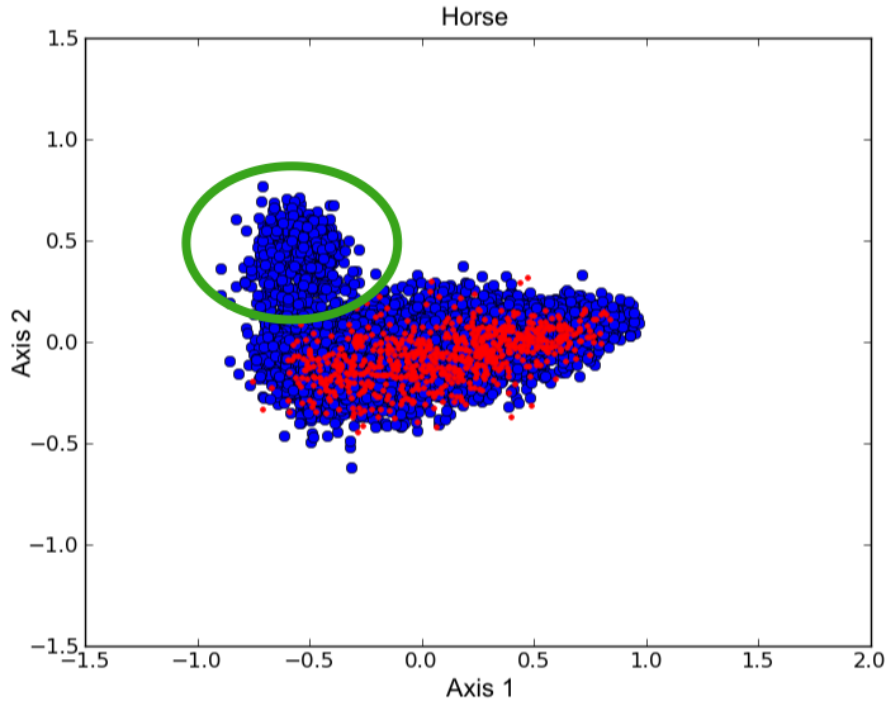


Figure 4.1 An example of an unusual cluster of genes.

Plot of axis 1 versus axis 2 of correspondence analysis for horse. Blue indicates non-highly expressed genes. Red indicates highly expressed genes. Genes that display an unusual pattern of codon usage are indicated with a green circle.

Table 4.5 Co-ordinates of genes composing unusual clusters that were extracted for further analysis.

Once axis 1 versus axis 2 of each CA was plotted, there were four species that exhibited unusual clusters of genes. To extract these genes for further analysis, it was necessary to define their location on the plot. Column 1 indicates the cut off for the co-ordinates of the genes that were extracted from the x-axis. Column 2 indicates the cut off for the co-ordinates of the genes that were extracted from the y-axis. For example, in horse, all genes that had an x-axis co-ordinate less than -0.2 and a y-axis co-ordinate greater than +0.2 was extracted for further analysis.

	X axis	**Y axis**	**No. Genes extracted**
Anole	**-0.1 – 0.5**	**Less than -0.3**	164
Armadillo	**0.2 – 0.5**	**Less than -0.3**	125
Horse	**Less than -0.2**	**Above +0.2**	350
Tasmanian devil	**Greater than -0.5, Less than 1**	**Greater than 0.5**	55

4.3 Results

4.3.1 Evidence for translational selection in vertebrates

This study employs four specific strategies to understand the interplay between mutation, selection and random genetic drift in 38 vertebrates, using a data set of 394,745 genes. First, I used a a novel algorithm (CDC index) that considers background nucleotide composition to examine whether codon usage bias is more pronounced in highly-expressed genes than in lowly-expressed genes. Second, I examined whether there was a correlation between the most abundant tRNA genes in a genome and the most preferred codons in each set of highly-expressed genes. Third, I compared base composition and N_c, and I investigated genes that appear to be using a more biased set of codons than might be expected from a simple mutational process. Finally, multivariate analyses were carried out to establish if there are additional patterns that substantially affect synonymous codon usage that might not be captured in these other analyses. Each of these results will be described in turn.

I examined whether highly expressed genes tended to be more biased in their codon usage pattern than lowly expressed genes. This experiment was conducted in duplicate. First, gene length was explicitly accounted for. The number of paired genes of identical length that were extracted for each species may be observed in Table 4.6. In all cases, highly expressed genes exhibited a higher average codon bias than lowly expressed genes. In general, the average codon bias for each set of highly expressed genes was in the range of ~0.14 – 0.15 (on average, across all species, 0.143), while the average codon bias for each set of lowly expressed genes was in the range of ~0.12 – 0.13 (on average, across all species, 0.135) (Table 4.6). Using the Wilcoxon signed-rank test and setting a significant P-value cut-off as less than or equal to 0.05, the difference in codon bias distributions between highly and lowly expressed genes was statistically significant in 21 of the 38 species (Table 4.6).

The calculation of CDC scores for highly and lowly expressed genes was repeated, not accounting for gene length. In this case, average codon bias was similar to that observed when gene length was strictly accounted for, ~0.14 – 0.15 in highly expressed genes (on average across all species, 0.156), versus ~0.12 – 0.13 in lowly expressed genes (on average across all species, 0.128) (Table 4.7). In this instance, the difference in codon bias levels between highly and lowly expressed genes was statistically significant in all 38 species, using the Mann-Whitney test and setting a significant P-value cut-off as less than 0.05 (Table 4.7). Although the difference in average CDC index between highly and lowly expressed genes is quite small (generally ~0.01 difference in CDC scores between highly and lowly expressed genes), a slightly larger difference (~0.02) was observed in rodents, marsupials and monotremes. Rodents and marsupials have previously been demonstrated to display a higher effective population size than other mammals, such as primates (Goodstadt et al., 2007, Hughes and Friedman, 2009). This corresponds favourably with the scenario in which species with a higher effective population size are more likely to display evidence of translational selection than those with smaller effective population sizes.

Next, I asked whether there was a correlation between the most abundant tRNA genes in a genome and the preferred codons in each set of highly expressed genes. For 23 species, I asked how often the most abundant tRNA gene matched the preferred synonymous sense codon (i.e. excluding termination codons, tryptophan and methionine). A match between tRNA gene and preferred codon was regularly observed in those amino acids that are encoded by two codons (Table 4.8). A more interesting observation is matches between codons and tRNA genes for amino acids that are encoded by more than two codons. In such cases, matches are still observed in all 23 species in leucine, valine, glycine, proline, serine and arginine (Table 4.8). An equivalent analysis of matching tRNA gene to mRNA codon abundance was conducted for lowly expressed genes. Almost identical results to those

described in the previous paragraph for highly expressed genes were obtained for lowly expressed genes (Table 4.9).

Table 4.6 CDC scores for each set of highly expressed (HE) and lowly expressed (LE) genes, once gene length is explicitly controlled.

Species	Number of genes	CDC (HE)	CDC (LE)	P-value
American Pika	59	0.15	0.14	0.01
Anole	69	0.13	0.12	0.02
Armadillo	40	0.15	0.14	0.47
Cat	38	0.14	0.15	0.48
Chicken	72	0.13	0.12	0.29
Chimpanzee	177	0.14	0.13	0.12
Cow	161	0.14	0.13	0.01
Dog	166	0.14	0.13	0
Elephant	172	0.14	0.13	0
Finch	68	0.13	0.13	0.35
Flying Fox	81	0.14	0.13	0.06
Galago	176	0.14	0.13	0.05
Gibbon	157	0.14	0.14	0.05
Gorilla	195	0.14	0.13	0.04
Guinea Pig	151	0.14	0.13	0.41
Hedgehog	39	0.15	0.14	0.1
Horse	171	0.14	0.13	0.02
Human	182	0.14	0.13	0.02
Hyrax	52	0.15	0.15	0.16
Kangaroo Rat	58	0.16	0.13	0
Macaque	175	0.15	0.14	0
Marmoset	161	0.14	0.13	0.02
Microbat	145	0.14	0.13	0.01
Mouse Lemur	172	0.14	0.15	0.59
Mouse	61	0.15	0.13	0
Panda	156	0.14	0.13	0
Opossum	133	0.14	0.12	0
Orangutan	156	0.14	0.14	0.01
Pig	130	0.14	0.14	0.31
Platypus	90	0.14	0.12	0.01
Rabbit	143	0.13	0.12	0.04
Rat	141	0.15	0.13	0
Sloth	19	0.16	0.15	0.42
Squirrel	38	0.16	0.16	0.03
Tarsier	36	0.15	0.15	0.15
Tasmanian Devil	107	0.14	0.14	0
Tenrec	44	0.13	0.13	0.66
Wallaby	35	0.15	0.15	0.29

Table 4.7 CDC scores for highly (HE) and lowly (LE) expressed genes, gene length not explicitly considered.

Species	CDC (HE)	CDC (LE)	P-value
American Pika	0.16	0.13	5.54E-13
Anole	0.14	0.11	2.16E-15
Armadillo	0.16	0.13	2.79E-05
Cat	0.17	0.13	5.19E-10
Chicken	0.15	0.12	5.61E-09
Chimpanzee	0.16	0.13	2.20E-16
Cow	0.15	0.12	2.20E-16
Dog	0.15	0.12	2.20E-16
Elephant	0.15	0.12	2.20E-16
Finch	0.15	0.12	2.20E-16
Flying Fox	0.15	0.13	4.69E-12
Galago	0.15	0.13	2.20E-16
Gibbon	0.15	0.13	2.20E-16
Gorilla	0.15	0.13	2.20E-16
Guinea Pig	0.15	0.13	2.30E-14
Hedgehog	0.17	0.13	1.15E-12
Horse	0.15	0.12	2.20E-16
Human	0.16	0.13	2.20E-16
Hyrax	0.17	0.13	3.04E-07
Kangaroo Rat	0.17	0.13	2.20E-16
Macaque	0.15	0.13	2.20E-16
Marmoset	0.15	0.12	2.20E-16
Microbat	0.15	0.12	2.20E-16
Mouse Lemur	0.15	0.14	1.86E-09
Mouse	0.16	0.12	2.20E-16
Panda	0.15	0.12	5.86E-16
Opossum	0.15	0.12	2.20E-16
Orangutan	0.16	0.13	2.20E-16
Pig	0.15	0.13	8.25E-16
Platypus	0.14	0.11	1.49E-08
Rabbit	0.14	0.12	2.20E-16
Rat	0.15	0.12	2.20E-16
Sloth	0.18	0.14	1.11E-07
Squirrel	0.17	0.13	1.28E-08
Tarsier	0.18	0.14	1.30E-07
Tasmanian Devil	0.15	0.12	2.20E-16
Tenrec	0.16	0.13	1.17E-09
Wallaby	0.17	0.13	3.81E-16

Table 4.8 Correlation between preferred codons and tRNA abundance (highly expressed genes).

Column 1 indicates the total number of cases (out of 20 amino acids) in which the preferred codon matched the most abundant tRNA gene for an amino acid. Column 2 demonstrates the number of these cases in which the amino acid was encoded by exactly 2 codons. The remaining columns indicate, for those amino acids encoded by more than 2 codons, precisely which codons matched the most abundant tRNA genes.

	Total No.	**AA encoded by 2 codons**	**Amino acids encoded by more than 2 codons**					
			Pro	**Leu**	**Ser**	**Val**	**Arg**	**Gly**
Anole	10	8	CCA	CUG				
Cat	8	8						
Chicken	10	9		CUG				
Chimpanzee	10	8				GUG		GGC
Cow	8	7				GUG		
Dog	10	8		CUG		GUG		
Finch	12	9		CUG	AGC			GGC
Gibbon	10	8				GUG		GGC
Gorilla	9	7				GUG		GGC
Guinea Pig	12	8		CUG	AGC	GUG		GGC
Horse	12	9			AGC	GUG		GGC
Human	10	8				GUG		GGC
Macaque	9	7				GUG		GGC
Marmoset	11	8		CUG		GUG		GGC
Mouse	12	9		CUG		GUG		GGC
Mouse Lemur	12	8		CUG	AGC	GUG		GGC
Opossum	13	7	CCU	CUG	UCU	GUG	AGA	GGC
Orangutan	11	8		CUG		GUG		GGC
Panda	10	8			AGC			GGC
Pig	9	7				GUG		GGC
Platypus	6	5					CGG	
Rabbit	13	9		CUG	AGC	GUG		GGC
Rat	11	8		CUG			AGG	GGC

Table 4.9 Correlation between preferred codons and tRNA abundance (lowly expressed genes).

Column 1 indicates the total number of cases (out of 20 amino acids) in which the preferred codon matched the most abundant tRNA gene for an amino acid. Column 2 demonstrates the number of these cases in which the amino acid was encoded by exactly 2 codons. The remaining columns indicate, for those amino acids encoded by more than 2 codons, precisely which codons matched the most abundant tRNA genes.

	Total No.	**AA encoded by 2 codons**	**Amino acids encoded by more than 2 codons**					
			Pro	**Leu**	**Ser**	**Val**	**Arg**	**Gly**
Anole	10	8	CCA	CUG				
Cat	8	8						
Chicken	10	9		CUG				
Chimpanzee	10	8				GUG		GGC
Cow	8	7				GUG		
Dog	10	8		CUG		GUG		
Finch	13	9	CCU	CUG	AGC			GGC
Gibbon	10	8				GUG		GGC
Gorilla	9	7				GUG		GGC
Guinea Pig	12	8		CUG	AGC	GUG		GGC
Horse	12	9				GUG		GGC
Human	10	8				GUG		GGC
Macaque	9	7				GUG		GGC
Marmoset	11	8		CUG		GUG		GGC
Mouse	12	9		CUG		GUG		GGC
Mouse Lemur	11	8		CUG		GUG		GGC
Opossum	11	7		CUG	UCU	GUG		GGC
Orangutan	11	8		CUG		GUG		GGC
Panda	10	8			AGC			GGC
Pig	9	7				GUG		GGC
Platypus	6	5					CGG	
Rabbit	13	9		CUG	AGC	GUG		GGC
Rat	11	8		CUG			AGG	GGC

4.3.2 Further exploration of synonymous codon usage variation in vertebrates

In each species, GC_3 values of genes ranged from 0 to 100%, while N_c values ranged from 20 to 61 (Table 4.10). As suggested by Wright, a plot of N_c against GC_3 can be used as a preliminary exploration of such heterogeneity. A normal distribution indicates the expected N_c values if bias is due to solely GC_3 because genes that are more biased in their base composition are expected to exhibit a lower number of effective codons As can be observed from the plots (Electronic Appendix 4.3), most genes in each species fall into a restricted cloud of genes, and highly expressed genes are not confined to a particular section of each plot.

A limited number of genes in each species contained a small number of effective codons ($N_c < 40$), in spite of the fact that their GC_3 content was relatively balanced (45-55%) (Table 4.4). These genes were extracted for further investigation. First, I examined whether this combined set of genes was enriched in any GO category or metabolic pathway. Genes were enriched in the GO category "defence response to bacterium" (GO:0042742), the Reactome pathways "REACT_1505" (integration of energy metabolism) and "REACT_15380" (diabetes pathway) and the KEGG pathways "hsa00190" (oxidative phosphorylation) and "hsa05012" (Parkinson's disease) (Electronic Appendix 4.4). These observations do not offer particular insight into why these genes may be using a smaller number of effective codons. The only noticeable difference between the genomic characteristics of these sets of genes and the rest of the orthologous genes was is a slightly lower GC content, GC_3 content and lower number of amino acids than the average for a given genome (Electronic Appendix 4.5). However, there is an inherent statistical limitation in the small number of genes that were extracted for most species, thus, such observations should be interpreted with caution.

Table 4.10 Minimum and maximum GC_3 content and N_c for each species.

Species	GC_3		N_c	
	Min	**Max**	**Min**	**Max**
American Pika	0	0.992	20	61
Anole	0.042	1	20	61
Armadillo	0.176	1	23.06	61
Cat	0.133	0.995	24.15	61
Chicken	0.153	1	21.5	61
Chimpanzee	0	1	23	61
Cow	0.198	0.994	20	61
Dog	0.125	1	22.73	61
Elephant	0.018	1	22.01	61
Finch	0.03	1	20	61
Flying Fox	0.083	1	20	61
Galago	0.077	0.978	20	61
Gibbon	0.059	0.976	25.09	61
Gorilla	0	1	20	61
Guinea Pig	0.029	1	22.14	61
Hedgehog	0.158	0.996	25.17	61
Horse	0.179	1	23	61
Human	0	1	20	61
Hyrax	0.2	1	25.03	61
Kangaroo Rat	0.125	1	23.35	61
Macaque	0.09	1	20.17	61
Marmoset	0.024	1	21.96	61
Microbat	0.091	1	21.91	61
Mouse	0	0.97	20.97	61
Mouse Lemur	0.161	1	24.15	61
Opossum	0.1	0.997	22.14	61
Orangutan	0	0.971	20	61
Panda	0.14	1	23.14	61
Pig	0	1	20.79	61
Platypus	0.112	1	21.91	61
Rabbit	0.06	1	21.5	61
Rat	0	1	20	61
Sloth	0.071	0.995	22.98	61
Squirrel	0.05	0.98	22.75	61
Tarsier	0.067	1	22.14	61
Tasmanian Devil	0.051	0.994	20.96	61
Tenrec	0.182	1	21.83	61
Wallaby	0.118	1	25.48	61
Min	0	0.97	20	61
Max	0.2	1	25.48	61

A correspondence analysis was conducted for each species to establish if there were additional patterns that greatly affect synonymous codon usage that may not have been previously uncovered. Axis 1 versus axis 2 of each correspondence analysis was plotted for each species. Most of the genes fall in a single cloud around the origin, and highly expressed genes are scattered throughout this cloud. The relative inertia of the first two axes was examined (Table 4.11; Electronic Appendix 4.6; Electronic Appendix 4.7). In the CA of RSCU values, it was observed that the first axis accounts for ~37.36% and the next highest axis accounts for ~4.22% of the relative inertia. Similarly, for the CA carried out with raw codon counts, the first axis accounts for ~30.83% of the relative inertia, while the next highest axis accounts for ~6.97%. Thus, it may be concluded that there is a single major trend governing synonymous codon usage.

A usefulness of correspondence analyses is that the co-ordinates of the genes on the major axes may be compared with other statistics (such as base composition) to investigate the meaning of the observed trends. There was no correlation detected between gene expression and axis 1 in any species, in agreement with previous observations (Musto et al., 2001) (Electronic Appendix 4.8). For each species, axis 1 co-ordinates strongly correlated with GC_3 (according to a Spearman correlation, ~0.96-0.99; Electronic Appendix 4.8 and 4.9). While the genomes are overall balanced in their GC content (~52%), the average GC content of the 1st position of the codons is 55%, of the 2nd positions is 42% and of the 3rd positions is 60% (Table 4.12). If vertebrate genomes were subject to a mutational bias towards GC-richness, the impact is most likely to be most pronounced at sites where there is little constraint. As third codon positions are subject to less constraint than first or second positions, the observation that GC_3 is higher than the overall GC-content is what is to be expected if the bias were due to mutational bias (Grocock and Sharp, 2002). Combining the evidence that the first axis contains most of the codon usage variation, and that the first axis

strongly correlates with GC_3 content suggests that mutational bias is a major factor that governs vertebrate synonymous codon usage.

From visual inspection of CA axis 1 versus axis 2 plots, a number of unusual clusters were identified in anole, armadillo, horse and tasmanian devil (Figure 4.2). It was observed that all the clusters comprise sequences that tend to be extremely short and repetitive in nature, and obviously represent among the most poorly sequenced and/or annotated sequences of the data sets.

Table 4.11 Correspondence analysis of RSCU values, and raw codon counts.

	RSCU values		Raw codon counts	
Species	**Axis 0 inertia**	**Max inertia of other axes**	**Axis 0 inertia**	**Max inertia of other axes**
American Pika	30.98	3.88	25.33	6.9
Anole	42.63	3.83	29.51	7.53
Armadillo	34.89	4.04	28.6	7.39
Cat	27.96	4.16	22.92	7
Chicken	41.83	3.85	30.54	6.41
Chimpanzee	42.19	4.13	36.12	5.99
Cow	44.58	3.84	37.04	7.04
Dog	44.44	4.25	36.85	6.91
Elephant	39.54	4.37	32.71	8.93
Finch	44.67	3.69	31.68	6.15
Flying Fox	37.13	4.35	31.55	6.39
Galago	37.6	4.78	31.65	6.99
Gibbon	41.08	4.01	35.91	5.81
Gorilla	42.68	4.14	34.93	6.03
Guinea Pig	40.58	4.27	34.46	7.08
Hedgehog	31.87	4.44	26.43	7.36
Horse	41.96	3.75	35.04	7.5
Human	42.35	4.21	34.07	6.12
Hyrax	29.5	4.55	25.92	6.66
Kangaroo Rat	30.74	6.88	27.32	7.04
Macaque	41.55	4.05	33.73	5.78
Marmoset	41.55	4.29	33.97	5.96
Microbat	44.75	3.62	37.23	6.2
Mouse	28.69	5.1	22.59	8.43
Mouse Lemur	33.34	4.06	26.75	6.52
Panda	42.89	3.54	36.14	6.83
Opossum	38.36	4.75	29.25	7.27
Orangutan	41.71	4.14	34.8	5.87
Pig	41.1	3.76	33.8	7.36
Platypus	41.8	3.23	32.73	7.01
Rabbit	48.12	3.21	39.49	6.66
Rat	26.71	4.95	20.1	8.51
Sloth	27.79	4.4	21.64	8.23
Squirrel	26.77	4.38	22.43	7.54
Tarsier	28.77	4.34	23.19	6.54
Tasmanian Devil	40.47	4.04	27.93	7.93
Tenrec	31.56	3.95	25.07	6.68
Wallaby	24.75	5.22	20.15	8.51

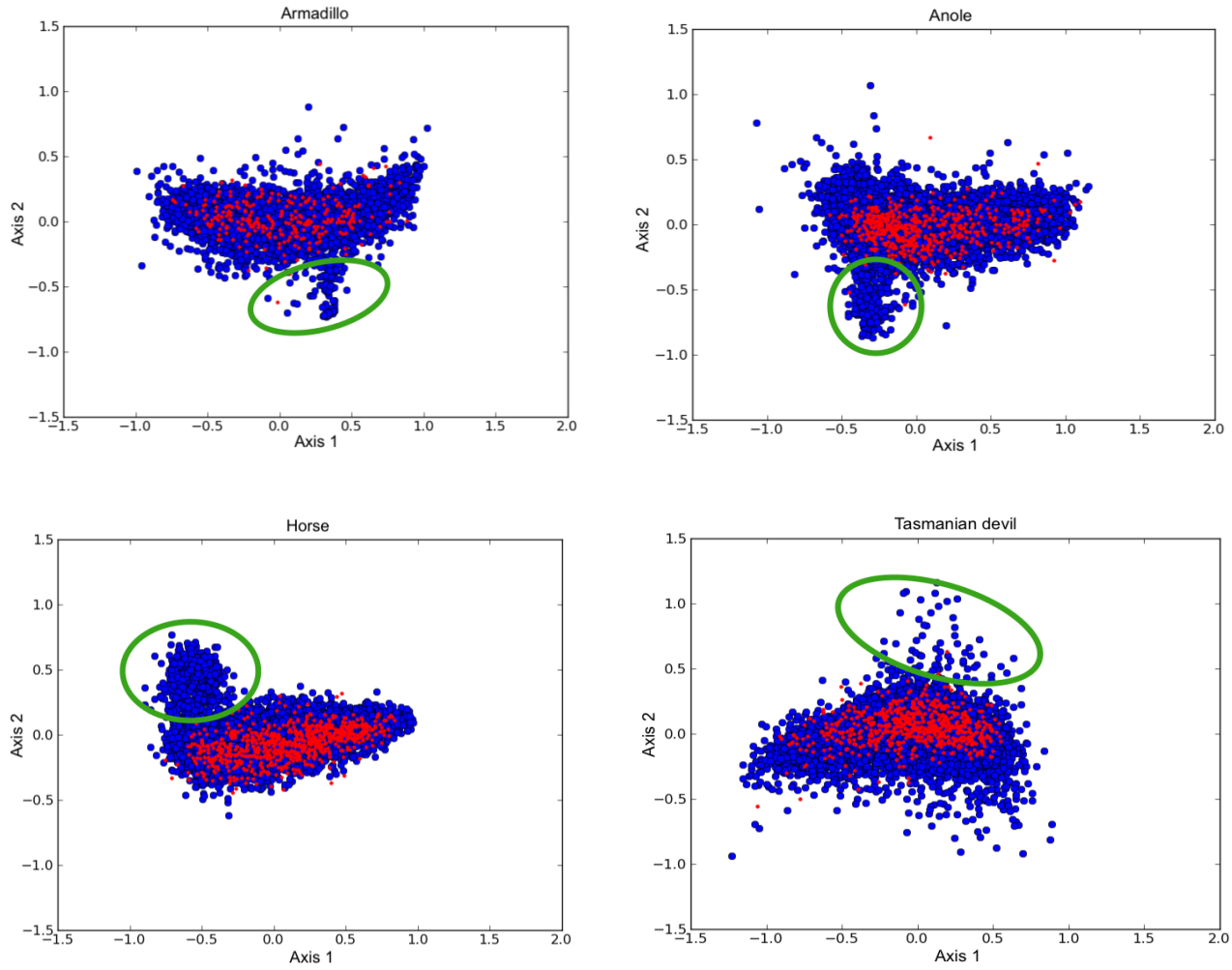


Figure 4.2 Clusters of genes on CA plots.

These clusters were observed from plotting axis 1 versus axis 2 of the correspondence analysis of each species.

Table 4.12 Average GC content in total, and at each codon position for each species.

Species	Total GC content	1st codon position	2nd codon position	3rd codon position
American Pika	55.89	57.19	43.55	66.94
Anole	49.03	53.02	40.36	53.73
Armadillo	53.07	55.21	42.64	61.37
Cat	55.53	56.89	43.87	65.83
Chicken	51.19	54.39	42.4	56.77
Chimpanzee	52.69	55.81	42.97	59.29
Cow	53.48	55.88	42.37	62.19
Dog	52.95	55.86	42.58	60.42
Elephant	51.81	54.63	41.63	59.17
Finch	52.01	54.61	41.13	60.28
Flying Fox	53.55	56.11	43.19	61.39
Galago	51.89	55.38	42.01	58.29
Gibbon	52.3	55.61	42.87	58.14
Gorilla	52.35	55.42	42.99	58.64
Guinea Pig	52.87	55.71	42.14	60.76
Hedgehog	52.74	54.72	42.57	60.95
Horse	51.63	54.48	41.15	59.26
Human	53.08	56.08	43.34	59.8
Hyrax	53.98	56.4	43.44	62.11
Kangaroo Rat	54.05	56.11	43.78	62.75
Macaque	52.25	55.47	42.89	58.38
Marmoset	52.17	55.46	42.56	58.48
Microbat	53.56	55.87	42.37	62.43
Mouse	51.6	54.32	42.25	58.24
Mouse Lemur	54.81	56.4	43.88	64.15
Opossum	48.89	53.13	40.84	52.68
Orangutan	52.54	55.61	43.1	58.92
Panda	53.02	55.73	42.3	61.03
Pig	52.29	55.81	42.41	61.65
Platypus	54.16	55.46	42.37	64.65
Polar Bear	52.16	54.75	42.02	59.88
Rabbit	54.34	56.62	42.75	64.1
Rat	51.49	54.24	42.05	58.16
Sloth	49.21	52.18	41.04	54.39
Squirrel	51.92	54.33	42.63	58.81
Tarsier	49.46	52.8	41.36	54.23
Tasmanian Devil	48.4	52.69	40.67	51.85
Tenrec	54.05	55.43	43.57	63.16
Wallaby	51.18	54.31	41.79	57.44
Average	52.40	55.13	42.41	59.76

4.4 Discussion

The aim of this chapter was to combine the recent surge in genome sequence availability with methodological advancements to further the understanding of the enigmatic relationship between mutation, selection and genetic drift in vertebrates. Detection of translational selection in vertebrates has proven to be a challenging and elusive task, obscured by low long-term effective population sizes and the existence of isochores (Musto et al., 2001; Urrutia and Hurst, 2001; Romero et al., 2003). There are two premises commonly invoked to be evident of translational selection. First, if selection acts to enhance protein translation, such selection should be particularly pronounced in highly expressed genes. Significantly stronger bias was observed in all cases when gene length was not controlled for, and in a substantial proportion of species when gene length was strictly controlled for. This pattern agrees with previous observations in *S. cerevisiae*, *C. elegans*, *A. thaliana* and *D. melanogaster* (Duret and Mouchiroud 1999; Castillo-Davis and Hartl 2002). The second sequence characteristic indicitive of the presence of natural selection is a correlation between the set of preferred codons used in highly expressed genes with the most abundant cellular tRNAs. In the 23 species for which tRNA gene data was available, a correlation between tRNA abundance and codon preference was regularly observed. To my knowledge, this is the first systematic detection of such a pattern in vertebrates, although isolated cases of such a correlation has previously been detected in prokaryotes (Ikemura, 1985, Yamao et al., 1991, Kanaya, 1999) and in eukaryotes (Ikemura, 1985, Kanaya et al., 2001). Taken together, these observations indicate that there may be an extremely weak selective force that affects synonymous codon usage choice in vertebrates.

I explored whether mutational bias could be detected as an additional factor governing synonymous codon usage in vertebrates. In plots that compare the effective number of codons to base composition, a normal distribution is expected if mutational bias solely

governs synonymous codon usage in vertebrates. For each species, most genes follow such a pattern. In addition, highly expressed genes do not tend to cluster in any one section of each plot. A more extensive analysis into the source of variation among synonymous codon usage choice may be obtained through a CA plot. From the examination of the level of variation contained in each axis, it became apparent that there was a single source of variation in synonymous codon usage that did not correlate with gene expression in any species, but strongly correlated with GC_3 content in all species. I also examined if axis 2 correlated with gene expression, but this was not the case. If mutational bias towards GC-richness were the sole force governing synonymous codon usage, one may expect to observe a higher GC content in the third codon position, which is under less selective constraint (Grocock and Sharp, 2002). In this study, the average GC content of each species was approximately 52%, while the average GC_3 content was 60%.

In summary, this chapter presents the first systematic study into the cause of synonymous codon usage in vertebrates. Although neutral processes undoubtedly play a substantial role in synonymous codon selection, Yang and Nielson (2008) suggested that there was possible evidence for an extremely weak selective force governing synonymous codon usage in humans. This investigation agrees that, in addition to mutational bias, translational selection appears to be an extremely weak force governing synonymous codon usage in vertebrates.

Chapter 5: Concluding remarks

The aim of this thesis was to combine increased genomic sequence and expression data availability with complementary methodological advancements to address questions pertaining to vertebrate molecular evolution from a novel perspective.

Chapter 2 coupled comparative genomics and protein-protein interaction network data to explore the relationship between the structure of the primate PIN and the duplicability of the genes encoding its components. A key methodological advancement exploited in this study was the development of an algorithm that allowed for the comparison of phylogenetic tree topologies in the presence of gene duplications (Marcet Houbon and Gabaldon, 2010). This novel algorithm may at least partially explain why this study found evidence for significant topological similarity between physically interacting proteins, contrary to a recent similar investigation in yeast (Kelly and Stumpf, 2010). This study also exploited the availability of high quality genomes, an essential factor given that Milinkovitch et al. (2010) recently demonstrated that low-coverage genomes tend to generate striking artifacts in gene duplication events.

In the third chapter, an attempt is made to estimate the timing of the speciation event between polar bear and giant panda, in the hope that an accurate divergence estimate would aid future studies that explore the radiation of ursine bears. The investigation benefitted from almost 30 times more nuclear sequence data per bear species than the current largest bear phylogeny study using nuclear data. Even this increase in data, combined with our current knowledge of the fossil record in a Bayesian framework, was not enough to understand the estimation of divergence between these two species. This should spur the quest to sequence more bear genomes, which should finally resolve the enigmatic divergence between these two species. In addition, the implementation of a singleton analysis was an effective method to

demonstrate the heterogeneity that exists between the sequences of the data set. Careful consideration of sequence heterogeneity was also necessary in chapter 4. Once CA plots had been constructed, it became clear that some species exhibited an unusual pattern of codon usage. However, on further examination, I observed that those genes were among the poorly sequenced in the data set. Thus, it is vital to be aware of sequence heterogeneity in any molecular evolutionary study.

The search for evidence that translational selection governs synonymous codon usage in vertebrates has proven to be challenging, and until now, widespread detection of such selection in vertebrates has been evasive. The investigation described in chapter 4 presents, to my knowledge, the first systematic study demonstrating that mutational bias is a key force affecting synonymous codon usage in vertebrates. In addition, the observations indicate that translational selection may be an additional widespread, albeit weak, selective force that governs the use of synonymous codons in vertebrates.

This thesis is a testimony to an obvious progression in the field of molecular evolution. For example, a decade ago, the studies in this thesis could not have been achieved, due to a lack of vertebrate genome sequence data, expression data and protein-protein interaction network data. In addition, some of the crucial methodological concepts, such as the CDC index and the treeKO algorithm, were yet to be devised. A progression of concept development of this extent is undoubtedly promising for the next decade, and future in general, of molecular evolutionary investigations.

Chapter 6: Bibliography

ABASCAL, F., ZARDOYA, R. & POSADA, D. (2005) ProtTest: selection for best-fit models of protein evolution. *Bioinformatics*, 21, 2104-2105.

ABASCAL, F., ZARDOYA, R. & TELFORD, M. J. (2010) Translator X: Multiple alignment of nucleotide sequences guided by amino acid translations. *Nucleic Acids Research,* 38, W7-W13.

AKAIKE, H. (1974) A new look at the statistical model identification. *IEEE Transactions of Automatic Control,* 19, 716-723.

AKASHI, H. (1994) Synonymous codon usage in *Drosophila melanogaster*: Natural selection and translational accuracy. *Genetics,* 136, 927-935.

AKASHI, H. (1997) Codon bias evolution in *Drosophila*. Population genetics of mutation-selection drift. *Gene,* 205, 269-278.

AL-SHAHROUR, F., DÍAZ-URIARTE, R. & DOPAZO, J. (2004) FatiGO: A web tool for finding significant associations of Gene Ontology terms with groups of genes. *Bioinformatics,* 20, 578-580.

ALFARO, M. E. & HUELSENBECK, J. P. (2006) Comparative performance of Bayesian and AIC-based measures of phylogenetic model uncertainty. *Systematic Biology,* 55, 89-96.

ALTSCHUL, S. F., GISH, W., MILLER, W., MYERS, E. W. & LIPMAN, D. J. (1990) Basic local alignment search tool. *Journal of Molecular Biology,* 215, 403-410.

AMOUTZIAS, G. & VAN DE PEER, Y. (2010) Single-gene and whole-genome duplications and the evolution of protein–protein interaction networks. *Evolutionary Genomics and Systems Biology,* 107, 413-429.

ANGOV, E., HILLIER, C. J., KINCAID, R. L. & LYON, J. A. (2008) Heterologous protein expression is enhanced by harmonizing the codon usage frequencies of the target gene with those of the expression host. *PloS ONE,* 3, e2189.

ANSORGE, W. J. (2009) Next-generation DNA sequencing techniques. *New Biotechnology,* 25, 195-203.

ASHBURNER, M., BALL, C. A., BLAKE, J. A., BOTSTEIN, D., BUTLER, H., CHERRY, J. M., DAVIS, A. P., DOLINSKI, K., DWIGHT, S. S. & EPPIG, J. T. (2000) Gene Ontology: Tool for the unification of biology. *Nature Genetics,* 25, 25-29.

AVERY, O. T., MACLEOD, C. M. & MCCARTY, M. (1944) Studies on the chemical nature of the substance inducing transformation of pneumococcal types. *The Journal of Experimental Medicine,* 79, 137-158.

BAELE, G., RAES, J., VAN DE PEER, Y. & VANSTEELANDT, S. (2006) An improved statistical method for detecting heterotachy in nucleotide sequences. *Molecular Biology and Evolution,* 23, 1397-1405.

BAILEY, J. A. & EICHLER, E. E. (2006) Primate segmental duplications: crucibles of evolution, diversity and disease. *Nature Reviews Genetics*, 7, 552-564.

BENINDA-EMONDS, O. (2004) Trees versus characters and the supertree/supermatrix paradox. *Systematic Biology*, 53, 356-359.

BENINDA-EMONDS, O., CARDILLO, M., JONES, K., MACPHEE, R., BECK, R., GRENYER, R., PRICE, S., VOS, R., GITTLEMAN, J. L. & PURVIS, A. (2008) The delayed rise of present-day mammals. *Nature*, 446, 507-512.

BENTON, M., DONOGHUE, P. & ASHER, R. (2009) Calibrating and constraining molecular clocks. In: *The Timetree of Life*. Cambridge University Press, Hedges, S., 35-86.

BERGSTEN, J. (2005) A review of long-branch attraction. *Cladistics,* 21, 163-193.

BERNARDI, G. (1995) The human genome: Organization and evolutionary history. *Annual Review of Genetics,* 29, 445-476.

BERNARDI, G. & BERNARDI, G. (1986) Compositional constraints and genome evolution. *Journal of Molecular Evolution,* 24, 1-11.

BLAIR, J. E. & HEDGES, S. B. (2005) Molecular clocks do not support the cambrian explosion. *Molecular Biology and Evolution,* 22, 387-390.

BOX, G. E. P. (1976) Science and statistics. *Journal of the American Statistical Association,* 791-799.

BRINKMANN, H. & PHILIPPE, H. (1999) Archaea sister group of bacteria? Indications from tree reconstruction artifacts in ancient phylogenies. *Molecular Biology and Evolution,* 16, 817-825.

BROMHAM, L., RAMBAUT, A., FORTEY, R., COOPER, A. & PENNY, D. (1998) Testing the Cambrian explosion hypothesis by using a molecular dating technique. *Proceedings of the National Academy of Sciences of the United States of America,* 95, 12386-12389.

BROWN, W. M., PRAGER, E. M., WANG, A. & WILSON, A. C. (1982) Mitochondrial DNA sequences of primates: Tempo and mode of evolution. *Journal of Molecular Evolution,* 18, 225-239.

BRUN, C., CHEVENET, F., MARTIN, D., WOJCIK, J., GUÉNOCHE, A. & JACQ, B. (2003) Functional classification of proteins for the prediction of cellular function from a protein-protein interaction network. Genome Biology, 5:R6.

BUCKLEY, T. R. & CUNNINGHAM, C. W. (2002) The effects of nucleotide substitution model assumptions on estimates of nonparametric bootstrap support. *Molecular Biology and Evolution,* 19, 394-405.

BUCKLEY, T. R., SIMON, C. & CHAMBERS, G. K. (2001) Exploring among-site rate variation models in a maximum likelihood framework using empirical data: Effects of model

assumptions on estimates of topology, branch lengths, and bootstrap support. *Systematic Biology,* 50, 67-86.

BULMER, M. (1988) Are codon usage patterns in unicellular organisms determined by selection-mutation balance? *Journal of Evolutionary Biology,* 1, 15-26.

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

CAMIN, J. H. & SOKAL, R. R. (1965) A method for deducing branching sequences in phylogeny. *Evolution,* 311-326.

CAPELLA-GUTIÉRREZ, S., SILLA-MARTÍNEZ, J. M. & GABALDÓN, T. (2009) TrimAl: A tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics,* 25, 1972-1973.

CASTILLO-DAVIS, C. I. & HARTL, D. L. (2002) Genome evolution and developmental constraint in *Caenorhabditis elegans*. *Molecular Biology and Evolution,* 19, 728-735.

CASTILLO-DAVIS, C. I., MEKHEDOV, S., HARTL, D. L. & KOONIN, E. (2002) Selection for short introns in highly expressed genes. *Nature*, 31, 415-418.

CASTRESANA, J. (2000) Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Molecular Biology and Evolution,* 17, 540-552.

CAVALLI-SFORZA, L. L. & EDWARDS, A. W. F. (1967) Phylogenetic analysis. Models and estimation procedures. *American Journal of Human Genetics,* 19, 233-257.

CHAMARY, J., PARMLEY, J. L. & HURST, L. D. (2006) Hearing silence: Non-neutral evolution at synonymous sites in mammals. *Nature Reviews Genetics,* 7, 98-108.

CHAN, E. T., QUON, G. T., CHUA, G., BABAK, T., TROCHESSET, M., ZIRNGIBL, R. A., AUBIN, J., RATCLIFFE, M., WILDE, A. & BRUDNO, M. (2009) Conservation of core gene expression in vertebrate tissues. *Journal of Biology,* 8, 33.

CHAN, P. P. & LOWE, T. M. (2009) GtRNAdb: A database of transfer RNA genes detected in genomic sequence. *Nucleic Acids Research,* 37, D93-D97.

CHANG, B. S. W. & CAMPBELL, D. L. (2000) Bias in phylogenetic reconstruction of vertebrate rhodopsin sequences. *Molecular Biology and Evolution,* 17, 1220-1231.

CHARLESWORTH, J. & EYRE-WALKER, A. (2006) The rate of adaptive evolution in enteric bacteria. *Molecular Biology and Evolution*, 23, 1348-1356.

CHATTERJEE, H. J., HO, S. Y. W., BARNES, I. & GROVES, C. (2009) Estimating the phylogeny and divergence times of primates using a supermatrix approach. *BMC Evolutionary Biology,* 9, 259.

CONANT, G. C. & WAGNER, A. (2002) GenomeHistory: a software tool and its application to fully sequenced genomes. *Nucleic Acids Research*, 30, 3378-3386.

CRICK, F., BARNETT, L., BRENNER, S. & WATTS-TOBIN, R. J. (1961) General nature of the genetic code for proteins. *Nature*, 192, 1227-1232.

CRISCUOLO, A. & GRIBALDO, S. (2010) BMGE (Block Mapping and Gathering with Entropy): A new software for selection of phylogenetic informative regions from multiple sequence alignments. *BMC Evolutionary Biology,* 10, 210.

CUMMINS, C. A. & MCINERNEY, J. O. (2011) A method for inferring the rate of evolution of homologous characters that can potentially improve phylogenetic inference, resolve deep divergence and correct systematic biases. *Systematic Biology,* 60, 833-844.

D'ANTONIO, M. & CICCARELLI, F. D. (2011) Modification of gene duplicability during the evolution of protein interaction network. *PLoS Computational Biology,* 7, e1002029.

DARWIN, C. (1859) On the Origin of Species. London, John Murray.

DAYHOFF, M. O. & ECK, R. V. (1968) *A model of evolutionary change in proteins*. In: Atlas of protein sequence and structure: 1967-68. National Biomedical Research Foundation, Silver Spring, Maryland.

DAYHOFF, M. O., SCHWARTZ, R. & ORCUTT, B. (1972) *A model of evolutionary change in proteins*. In: Atlas of Protein Sequence and Structure*,* 5**,** 345-352. National Biomedical Research Foundation, Silver Spring, Maryland.

DE QUEIROZ, A. & GATESY, J. (2007) The supermatrix approach to systematics. *Trends in Ecology and Evolution,* 22, 34-41.

DELSUC, F., BRINKMANN, H. & PHILIPPE, H. (2005) Phylogenomics and the reconstruction of the tree of life. *Nature Reviews Genetics,* 6, 361-375.

DOHERTY, A., ALVAREZ-PONCE, D. & MCINERNEY, J. O. (2012) Increased genome sampling reveals a dynamic relationship between gene duplicability and the structure of the primate protein-protein interaction network. *Molecular Biology and Evolution* (published online July 19, 2012).

DOS REIS, M. & WERNISCH, L. (2009) Estimating translational selection in eukaryotic genomes. *Molecular Biology and Evolution,* 26, 451-461.

DOUZERY, E., SNELL, E., BAPTESTE, E., DELSUC, F. & PHILLIPE, H (2004) The timing of eukaryotic evolution: does a relaxed molecular clock reconcile proteins and fossils? *Proceedings of the National Academy of Sciences of the United States of America,* 101, 15386-15391.

DRUMMOND, A. J., HO, S. Y. W., PHILLIPS, M. J. & RAMBAUT, A. (2006) Relaxed phylogenetics and dating with confidence. *PLoS Biology,* 4, e88.

DRUMMOND, D. A. & WILKE, C. O. (2009) The evolutionary consequences of erroneous protein synthesis. *Nature Reviews Genetics,* 10, 715-724.

DRUMMOND, D. A. & WILKE, C. O. (2008) Mistranslation-induced protein folding as a dominant constraint on coding sequence evolution. *Cell*, 134, 341-352.

DUNN, C. W., HEJNOL, A. C., MATUS, D. Q., PANG, K., BROWNE, W. E., SMITH, S. A., SEAVER, E., ROUSE, G., W., OBST, M., EDGECOMBE, G. D., SORENSON, M. V., HADDOCK, S. H. D., SCHMIDT-RHAESA, A., OKUSU, A., KRISTENSEN, R. M., WHEELER, W. C., MARTINDALE, M. Q., GIRIBET, G. (2008) Broad phylogenomic sampling improves resolution of the animal tree of life. *Nature*, 452, 745-749.

DURET, L. (2000) tRNA gene number and codon usage in the *C. elegans* genome are co-adapted for optimal translation of highly expressed genes. *Trends in Genetics,* 16**,** 287-289.

DURET, L. & MOUCHIROUD, D. (1999) Expression pattern and, surprisingly, gene length shape codon usage in *Caenorhabditis*, *Drosophila*, and *Arabidopsis*. *Proceedings of the National Academy of Sciences of the United States of America,* 96, 4482-4487.

DUTHEIL, J. Y. (2012) Detecting coevolving positions in a molecule: Why and how to account for phylogeny. *Briefings in Bioinformatics, 13,* 228-243.

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

EDWARDS, A. W. F. (1984) Likelihood. In: *Likelihood* City: Cambridge University Press.

EDWARDS, A. W. F. & CAVALLI-SFORZA, C. L. L. (1964) Reconstruction of evolutionary trees. In: *Phenetic and Phylogenetic Classification,* Ed. V. H. Haywood and J. McNeill, Systematic Association Publ. No. 6, London, 67-76.

EDWARDS, A. W. F. & CAVALLI-SFORZA, C. L. L. (1963) The reconstruction of evolution. *Heredity,* 18.

EFRON, B. (1979) Bootstrap methods: Another look at the jackknife. *The Annals of Statistics,* 7, 1-26.

EISEN, J. A. & FRASER, C. M. (2003) Phylogenomics: Intersection of evolution and genomics. *Science,* 300, 1706-1707.

ERWIN, D., LAFLAMME, M., TWEEDT, S., SPERLING, E., PISANI, D. & PETERSON, K. (2011) The Cambrian Conundrum: Early divergence and later ecological success in the early history of animals. *Science*, 334, 1091-1097.

ERWIN, T. (1991) An evolutionary basis for conservation strategies. *Science,* 253, 750-752.

FABRE, P. H., HAUTIER, L., DIMITROV, D. & DOUZERY, E. J. P. (2012) A glimpse on the pattern of rodent diversification: A phylogenetic approach. *BMC Evolutionary Biology,* 12, 88.

FARRER, M. J. (2006) Genetics of Parkinson's disease: paradigm shifts and future prospects. *Nature Reviews Genetics*, 7, 306-318.

FELSENSTEIN, J. (1978) Cases in which parsimony or compatibility methods will be positively misleading. *Systematic Biology,* 27**,** 401-410.

FELSENSTEIN, J. (1981) Evolutionary trees from DNA sequences: A maximum likelihood approach. *Journal of Molecular Evolution,* 17, 368-376.

FELSENSTEIN, J. (1985) Confidence limits on phylogenies: An approach using the bootstrap. *Evolution,* 39, 783-791.

FELSENSTEIN, J. (2004) Inferring phylogenies. *Sunderland, Massachusetts: Sinauer Associates.*

FISHER, R. A. (1912) On an absolute criterion for fitting frequency curves. *Messenger of Mathematics*, 41, 155-160.

FISHER, R. A. (1922) On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character,* 222, 309-368.

FITCH, W. M. (2000) Homology: A personal view on some of the problems. *Trends in Genetics,* 16, 227-231.

FITCH, W. M. & MARGOLIASH, E. (1967) Construction of phylogenetic trees. *Science,* 155, 279-284.

FITCH, W. M. & MARKOWITZ, E. (1970) An improved method for determining codon variability in a gene and its application to the rate of fixation of mutations in evolution. *Biochemical Genetics*, 4, 579-593.

FLICEK, P., AMODE, M. R., BARRELL, D., BEAL, K., BRENT, S., CARVALHO-SILVA, D., CLAPHAM, P., COATES, G., FAIRLEY, S. & FITZGERALD, S. (2012) Ensembl 2012. *Nucleic Acids Research,* 40, D84-D90.

FLICEK, P., AMODE, M. R., BARRELL, D., BEAL, K., BRENT, S., CHEN, Y., CLAPHAM, P., COATES, G., FAIRLEY, S. & FITZGERALD, S. (2011) Ensembl 2011. *Nucleic Acids Research,* 39, D800-D806.

FORTNA, A., KIM, Y., MACLAREN, E., MARSHALL, K., HAHN, G., MELTESEN, L., BRENTON, M., HINK, R., BURGERS, S. & HERNANDEZ-BOUSSARD, T. (2004) Lineage-specific gene duplication and loss in human and great ape evolution. *PLoS Biology,* 2, e207.

FOSTER, P. G., COX, C. J. & EMBLEY, T. M. (2009) The primary divisions of life: A phylogenomic approach employing composition-heterogeneous methods. *Philosophical Transactions of the Royal Society B: Biological Sciences,* 364**,** 2197-2207.

FOSTER, P. G., JERMIIN, L. S. & HICKEY, D. A. (1997) Nucleotide composition bias affects amino acid content in proteins coded by animal mitochondria. *Journal of Molecular Evolution,* 44, 282-288.

FRASER, H. B., HIRSH, A. E., STEINMETZ, L. M., SCHARFE, C. & FELDMAN, M. W. (2002) Evolutionary rate in the protein interaction network. Science, 296, 750-752.

FRASER, H. B., HIRSH, A. E., WALL, D. P. & EISEN, M. B. (2004) Co-evolution of gene expression among interacting proteins. *Proceedings of the National Academy of Sciences of the United States of America,* 101, 9033-9038.

FREEMAN, L. C. (1979) Centrality in social networks conceptual clarification. *Social Networks,* 1, 215-239.

FRYXELL, K. J. (1996) The co-evolution of gene family trees. *Trends in Genetics,* 12, 364-369.

GABALDÓN, T. (2008) Large-scale assignment of orthology: Back to phylogenetics. *Genome Biology,* 9, 235.

GADAGKAR, S. R. & KUMAR, S. (2005) Maximum likelihood outperforms maximum parsimony even when evolutionary rates are heterotachous. *Molecular Biology and Evolution,* 22**,** 2139-2141.

GAUT, B. S. & LEWIS, P. O. (1995) Success of maximum likelihood phylogeny inference in the four-taxon case. *Molecular Biology and Evolution,* 12, 152-162.

GEE, H. (2003) Ending incongruence. *Nature*, 425, 782.

GILLOOLY, J. F., ALLEN, A. P., WEST, G. B. & BROWN, J. H. (2005) The rate of DNA evolution: Effects of body size and temperature on the molecular clock. *Proceedings of the National Academy of Sciences of the United States of America,* 102, 140-145.

GLAZKO, G. V., KOONIN, E. V. & ROGOZIN, I. B. (2005) Molecular dating: Ape bones agree with chicken entrails. *Trends in Genetics,* 21, 89-92.

GLAZKO, G. V. & NEI, M. (2003) Estimation of divergence times for major lineages of primate species. *Molecular Biology and Evolution*, 20, 424-434.

GOH, C. S., BOGAN, A. A., JOACHIMIAK, M., WALTHER, D. & COHEN, F. E. (2000) Co-evolution of proteins with their interaction partners. *Journal of Molecular Biology,* 299, 283-293.

GOJOBORI, T., LI, W. H. & GRAUR, D. (1982) Patterns of nucleotide substitution in pseudogenes and functional genes. *Journal of Molecular Evolution,* 18, 360-369.

GOODMAN, M., CZELUSNIAK, J., MOORE, G. W., ROMERO-HERRERA, A. & MATSUDA, G. (1979) Fitting the gene lineage into its species lineage, a parsimony strategy illustrated by cladograms constructed from globin sequences. *Systematic Biology,* 28, 132-163.

GOODMAN, S. N. (1999) Toward evidence-based medical statistics. 2: The Bayes factor. *Annals of Internal Medicine,* 130, 1005-1013.

GOODSTADT, L., HEGER, A., WEBBER, C. & PONTING, C. P. (2007) An analysis of the gene complement of a marsupial *Monodelphis domestica*: Evolution of lineage-specific genes and giant chromosomes. Genome Research, 17, 969-981.

GOODSTADT, L. & PONTING, C. P. (2006) Phylogenetic reconstruction of orthology, paralogy, and conserved synteny for dog and human. *PLoS Computational Biology,* 2, e133.

GRAUR, D. & MARTIN, W. (2004) Reading the entrails of chickens: Molecular timescales of evolution and the illusion of precision. *Trends in Genetics,* 20, 80-86.

GREEN, P. (2007) 2× genomes - does depth matter? *Genome Research,* 17, 1547-1549.

GROCOCK, R. J. & SHARP, P. M. (2002) Synonymous codon usage in *Pseudomonas aeruginosa* pa01. *Gene,* 289, 131-139.

GUINDON, S. & GASCUEL, O. A. (2003) A simple, fast and accurate algorithm to estimate large phylogenies by maximum likelihood. *Systematic Biology*, 52, 696-704.

HAHN, M. W. (2009) Distinguishing among evolutionary models for the maintenance of gene duplicates. *Journal of Heredity*, 100, 605-617.

HAHN, M. W. (2007) Bias in phylogenetic tree reconciliation methods: Implications for vertebrate genome evolution. *Genome Biology,* 8, R141.

HAHN, M. W., DEMUTH, J. P. & HAN, S. G. (2007) Accelerated rate of gene gain and loss in primates. *Genetics,* 177, 1941.

HAHN, M. W. & KERN, A. D. (2005) Comparative genomics of centrality and essentiality in three eukaryotic protein-interaction networks. *Molecular Biology and Evolution*, 22, 803-806.

HAKES, L., PINNEY, J. W., LOVELL, S. C., OLIVER, S. G. & ROBERTSON, D. L. (2007) All duplicates are not equal: The difference between small-scale and genome duplication. *Genome Biology,* 8, R209.

HAN, J. D. J., BERTIN, N., HAO, T., GOLDBERG, D. S., BERRIZ, G. F., ZHANG, L. V., DUPUY, D., WALHOUT, A. J. M., CUSICK, M. E. & ROTH, F. P. (2004) Evidence for dynamically organized modularity in the yeast protein–protein interaction network. *Nature,* 430, 88-93.

HARINGTON, C. (2008) The evolution of Arctic marine mammals. *Ecological Applications,* 18, 23-40.

HASEGAWA, M., KISHINO, H. & YANO, T. (1985) Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *Journal of Molecular Evolution,* 22, 160-174.

HASTINGS, W. K. (1970) Monte Carlo sampling methods using Markov chains and their applications. *Biometrika,* 57, 97-109.

HEATH, T. A., HEDTKE, S. M. & HILLIS, D. M. (2008) Taxon sampling and the accuracy of phylogenetic analyses. *Journal of Systematics and Evolution,* 46, 239-257.

HERSHEY, A. D. & CHASE, M. (1952) Independent functions of viral protein and nucleic acid in growth of bacteriophage. *The Journal of General Physiology,* 36, 39-56.

HILLIS, D. M. (1996) Inferring complex phylogenies. *Nature,* 383, 130.

HOFFMANN, M., HILTON-TAYLOR, C., ANGULO, A., BÖHM, M., BROOKS, T. M., BUTCHART, S. H. M., CARPENTER, K. E., CHANSON, J., COLLEN, B. & COX, N. A. (2010) The impact of conservation on the status of the world's vertebrates. *Science,* 330**,** 1503-1509.

HOLDER, M. & LEWIS, P. O. (2003) Phylogeny estimation: Traditional and Bayesian approaches. *Nature Reviews Genetics,* 4, 275-284.

HOOPER, E. (2001) Experimental oral polio vaccines and acquired immunodeficiency syndrome. *Philosophical Transactions of the Royal Society of London, Series B*, 356, 803-814.

HOU, Z-C., ROMERO, R & WILDMAN, D. E. (2009) Phylogeny of the Ferungulata (Mammalia: Laurasiatheria) as determined from phylogenomic data. *Molecular Phylogenetics and Evolution*, 52, 660-664.

HRDY, I., HIRT, R. P., DOLEZAL, P., BARDONOVÁ, L., FOSTER, P. G., TACHEZY, J. & EMBLEY, T. M. (2004) Trichomonas hydrogenosomes contain the NADH dehydrogenase module of mitochondrial complex. *Nature,* 432, 618-622.

HUANG, T. W., LIN, C. Y. & KAO, C. Y. (2007) Reconstruction of human protein interolog network using evolutionary conserved network. *BMC Bioinformatics,* 8, 152.

HUBBARD, T., BARKER, D., BIRNEY, E., CAMERON, G., CHEN, Y., CLARK, L., COX, T., CUFF, J., CURWEN, V. & DOWN, T. (2002) The Ensembl genome database project. *Nucleic Acids Research,* 30**,** 38-41.

HUDSON, N. J., GU, Q., NAGARAJ, S. H., DING, Y. S., DALRYMPLE, B. P. & REVERTER, A. (2011) Eukaryotic evolutionary transitions are associated with extreme codon bias in functionally-related proteins. *PloS ONE,* 6, e25457.

HUELSENBECK, J. P. (1995) Performance of phylogenetic methods in simulation. *Systematic Biology*, 44, 17-48.

HUELSENBECK, J. P., RANNALA, B. & MASLY, J. P. (2000) Accomodating phylogenetic uncertainty in evolutionary studies. *Science*, 288, 2349-2350.

HUELSENBECK, J. P., RONQUIST, F., NIELSEN, R. & BOLLBACK, J. P. (2001) Bayesian inference of phylogeny and its impact on evolutionary biology. *Science,* 294**,** 2310-2314.

HUERTA-CEPAS, J., DOPAZO, H., DOPAZO, J. & GABALDÓN, T. (2007) The human phylome. *Genome Biology,* 8, R109.

HUERTA-CEPAS, J., DOPAZO, J. & GABALDÓN, T. (2010) ETE: A python environment for tree exploration. *BMC Bioinformatics,* 11, 24.

HUGHES, A. L. & FRIEDMAN, R. (2005) Gene duplication and the properties of biological networks. *Journal of Molecular Evolution,* 61, 758-764.

HUGHES, A. L. & FRIEDMAN, R. (2009) More radical amino acid replacement in primates than in rodents: support for the evolutionary role of effective population size. *Gene*, 440, 50-56.

HUNTER, C. M., CASWELL, H., RUNGE, M. C., REGEHR, E. V., AMSTRUP, S. C. & STIRLING, I. (2010) Climate change threatens polar bear populations: A stochastic demographic analysis. *Ecology,* 91, 2883-2897.

IDEKER, T. & SHARAN, R. (2008) Protein networks in disease. *Genome Research,* 18, 644-652.

IKEMURA, T. (1985) Codon usage and tRNA content in unicellular and multicellular organisms. *Molecular Biology and Evolution,* 2, 13-34.

INGÓLFSSON, Ó. & WIIG, Ø. (2008) Late Pleistocene fossil find in Svalbard: The oldest remains of a polar bear ever discovered. *Polar Research,* 28, 455-462.

ISPOLATOV, I., YURYEV, A., MAZO, I. & MASLOV, S. (2005) Binding properties and evolution of homodimers in protein–protein interaction networks. *Nucleic Acids Research,* 33, 3629-3635.

JACOB, H. J & KWITEK, A. E. (2001) Rat genetics: Attaching physiology and pharmacology to the genome. *Nature Reviews Genetics*, 3, 33-42.

JEFFROY, O., BRINKMANN, H., DELSUC, F. & PHILIPPE, H. (2006) Phylogenomics: The beginning of incongruence? *Trends in Genetics,* 22, 225-231.

JI, Q., LUO, Z. X., YUAN, C. X., WIBLE, J. R., ZHANG, J. P. & GEORGI, J. A. (2002) The earliest known eutherian mammal. *Nature,* 416, 816-822.

JIN, C., CIOCHON, R. L., DONG, W., HUNT, R. M., LIU, J., JAEGER, M. & ZHU, Q. (2007) The first skull of the earliest Giant Panda. *Proceedings of the National Academy of Sciences of the United States of America,* 104, 10932-10937.

JONES, D. T., TAYLOR, W. R. & THORNTON, J. M. (1992) The rapid generation of mutation data matrices from protein sequences. *Computational and Applied Biosciences*, 8, 275-282.

JOSHI-TOPE, G., GILLESPIE, M., VASTRIK, I., D'EUSTACHIO, P., SCHMIDT, E., DE BONO, B., JASSAL, B., GOPINATH, G., WU, G., MATTHEWS, L, LEWIS, S., BIRNEY, E. & STEIN, L. (2005) Reactome: a knowledgebase of biological pathways. *Nucleic Acids Research*, 33, D428-D432.

JUKES, T. H. & CANTOR, C. R. Evolution of protein molecules. In: Mammalian Protein Metabolism III, New York. Academic Press. Ed: H. N. Munro, 1969.

KANAYA, S., YAMADA, Y., KINOUCHI, M., KUDO, Y. & IKEMURA, T. (2001) Codon usage and trna genes in eukaryotes: Correlation of codon usage diversity with translation efficiency and with cg-dinucleotide usage as assessed by multivariate analysis. *Journal of Molecular Evolution,* 53, 290-298.

KANAYA, S., YAMADA, Y., KUDO, Y. & IKEMURA, T. (1999) Studies of codon usage and tRNA genes of 18 unicellular organisms and quantification of *Bacillus subtilis* tRNAs: Gene expression level and species-specific diversity of codon usage based on multivariate analysis. *Gene,* 238, 143-155.

KANEHISA, M., GOTO, S., KAWASHIMA, S., OKUNO, Y & HATTORI, M. (2002) The KEGG resource for deciphering the genome. *Nucleic Acids Research*, 32, D277-D280.

KASAHARA, M. (2007) The 2R hypothesis: An update. *Current Opinion in Immunology,* 19, 547-552.

KASS, R. E. & RAFTERY, A. E. (1995) Bayes Factors. *Journal of the American Statistical Association*, 90, 773-795.

KEANE, T. M., CREEVEY, C. J., PENTONY, M. M., NAUGHTON, T. J. & MCLNERNEY, J. O. (2006) Assessment of methods for amino acid matrix selection and their use on empirical data shows that ad hoc assumptions for choice of matrix are not justified. *BMC Evolutionary Biology,* 6, 29.

KELLY, W. P. & STUMPF, M. P. H. (2010) Trees on networks: resolving statistical patterns of phylogenetic similarities among interacting proteins. *BMC Bioinformatics*, 11.

KIDD, K. & SGARAMELLA-ZONTA, L. (1971) Phylogenetic analysis: Concepts and methods. *American Journal of Human Genetics,* 23, 235.

KIM, P. M., LU, L. J., XIA, Y. & GERSTEIN, M. B. (2006) Relating three-dimensional structures to protein networks provides evolutionary insights. *Science's STKE,* 314, 1938.

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

KIMURA, M. (1980) A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution,* 16, 111-120.

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

KIMURA, M. & OHTA, T. (1971) On the rate of molecular evolution. *Journal of Molecular Evolution,* 1, 1-17.

KINSELLA, R. J., KÄHÄRI, A., HAIDER, S., ZAMORA, J., PROCTOR, G., SPUDICH, G., ALMEIDA-KING, J., STAINES, D., DERWENT, P. & KERHORNOU, A. (2011) Ensembl BioMart: A hub for data retrieval across taxonomic space. *Database (Oxford) 2011 doi: 10.1093/database/bar030.*

KITAZOE, Y., KISHONO, H., WADDELL, P., NAKAJIMA, N., OKABAYASHI, T., WATABE, T. & OKUHURA, Y. (2007) Robust time estimate reconciles views of the antiquity of placental mammals. *PLoS ONE*, 4, e384.

KLUGE, A. G. (1989) A concern for evidence and a phylogenetic hypothesis of relationships among *Epicrates* (Boidae, Serpentes). *Systematic Zoology*, 38, 7-25.

KOLACZKOWSKI, B. & THORNTON, J. W. (2004) Performance of maximum parsimony and likelihood phylogenetics when evolution is heterogeneous. *Nature,* 431, 980-984.

KOMAR, A. A., LESNIK, T. & REISS, C. (1999) Synonymous codon substitutions affect ribosome traffic and protein folding during in vitro translation. *Febs Letters,* 462, 387-391.

KOONIN, E. V. (2005) Orthologs, paralogs, and evolutionary genomics. *Annual Review of Genetics,* 39, 309-338.

KOONIN, E. V. (2009) Darwinian evolution in the light of genomics. *Nucleic Acids Research,* 37, 1011-1034.

KORBER, B., MULDOON, M., THEILER, J., GAO, F., GUPTA, R., LAPEDES, A., HAHN, B., WOLINSKY, S. & BHATTACHARYA, T. (2000) Timing the ancestor of the HIV-1 pandemic strains. *Science,* 288, 1789-1796.

KORETKE, K. K., LUPAS, A. N., WARREN, P. V., ROSENBERG, M. & BROWN, J. R. (2000) Evolution of two-component signal transduction. *Molecular Biology and Evolution,* 17, 1956-1970.

KOSIOL, C., VINAŘ, T., DA FONSECA, R. R., HUBISZ, M. J., BUSTAMANTE, C. D., NIELSEN, R. & SIEPEL, A. (2008) Patterns of positive selection in six mammalian genomes. *PLoS Genetics,* 4, e1000144.

KRAUSE, J., UNGER, T., NOÇON, A., MALASPINAS, A. S., KOLOKOTRONIS, S. O., STILLER, M., SOIBELZON, L., SPRIGGS, H., DEAR, P. H. & BRIGGS, A. W. (2008) Mitochondrial genomes reveal an explosive radiation of extinct and extant bears near the Miocene-Pliocene boundary. *BMC Evolutionary Biology,* 8, 220.

KUHNER, M. K. & FELSENSTEIN, J. (1994) A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. *Molecular Biology and Evolution,* 11, 459-468.

KULLBERG, M., NILSSON, M. A., ARNASON, U., HARLEY, E. H. & JANKE, A. (2006) Housekeeping genes for phylogenetic analysis of Eutherian relationships. *Molecular Biology and Evolution*, 23, 1493-1503.

LAFAY, B., ATHERTON, J. C. & SHARP, P. M. (2000) Absence of translationally selected synonymous codon usage bias in *Helicobacter pylori*. *Microbiology,* 146, 851-860.

LAIDRE, K. L., STIRLING, I., LOWRY, L. F., WIIG, Ø., HEIDE-JØRGENSEN, M. P. & FERGUSON, S. H. (2008) Quantifying the sensitivity of Arctic marine mammals to climate-induced habitat change. *Ecological Applications,* 18, 97-125.

LAMARCK, J. B. D. (1809) Philosophie zoologique. Dentu, Paris.

LANDER, E. S. (2011) Initial impact of the sequencing of the human genome. *Nature,* 470, 187-197.

LANDER, E. S., LINTON, L. M., BIRREN, B., NUSBAUM, C., ZODY, M. C., BALDWIN, J., DEVON, K., DEWAR, K., DOYLE, M. & FITZHUGH, W. (2001) Initial sequencing and analysis of the human genome. *Nature,* 409, 860-921.

LANGERGRABER, K. E., PRÜFER, K., ROWNEY, C., BOESCH, C., CROCKFORD, C., FAWCETT, K., INOUE, E., INOUE-MURUYAMA, M., MITANI, J. C., MULLER, M. N., ROBBINS, M. M., SCHUBERT, G., STOINSKI, T. S., VIOLA, B., WATTS, D., WITTIG, R., WRANGHAM, R. W., ZUBERBÜHLER, K., PÄÄBO, S. & VIGILANT, L. (2012) Generation times in wild chimpanzees and gorillas suggest earlier divergence times in Great Ape and human evolution. *Proceedings of the National Academy of Sciences of the United States of America*, 39, 15716-15721.

LARTILLOT, N., LEPAGE, T. & BLANQUART, S. (2009) PhyloBayes 3: A Bayesian software package for phylogenetic reconstruction and molecular dating. Bioinformatics, 25, 2286-2288.

LEE, J. M. & SONNHAMMER, E. L. L. (2003) Genomic gene clustering analysis of pathways in eukaryotes. Genome Research, 13, 875-882.

LEE, M. S. Y. (1999) Molecular clock calibrations and metazoan divergence dates. *Journal of Molecular Evolution,* 49, 385-391.

LEE, M. S. Y. & CAMENS, A. B. (2009) Strong morphological support for the molecular evolutionary tree of placental mammals. *Journal of Evolutionary Biology,* 22, 2243-2257.

LEMMON, A. R & MORIARTY, E. C. (2004) The importance of proper model assumptions in Bayesian phylogenetics. *Systematic Biology*, 54, 265-277.

LEPAGE, T., BRYANT, D., PHILIPPE, H. & LARTILLOT, N. (2007) A general comparison of relaxed molecular clock models. *Molecular Biology and Evolution,* 24**,** 2669-2680.

LERAT, E., BIÉMONT, C. & CAPY, P. (2000) Codon usage and the origin of P-elements. *Molecular Biology and Evolution,* 17**,** 467-468.

LI, W. L. S. & RODRIGO, A. G. (2009) Covariation of branch lengths in phylogenies of functionally related genes. *PloS ONE,* 4**,** e8487.

LIANG, H. & LI, W. H. (2007) Gene essentiality, gene duplicability and protein connectivity in human and mouse. *Trends in Genetics,* 23**,** 375-378.

LIAO, B. Y. & ZHANG, J. (2006) Evolutionary conservation of expression profiles between human and mouse orthologous genes. *Molecular Biology and Evolution,* 3**,** 530-540.

LINDQVIST, C., SCHUSTER, S. C., SUN, Y., TALBOT, S. L., QI, J., RATAN, A., TOMSHO, L. P., KASSON, L., ZEYL, E. & AARS, J. (2010) Complete mitochondrial genome of a Pleistocene jawbone unveils the origin of polar bear. *Proceedings of the National Academy of Sciences of the United States of America,* 107**,** 5053.

LIPINSKI, K. J., FARSLOW, J. C., FITZPATRICK, K. A., LYNCH, M., KATJU, V. & BERGTHORSSON, U. (2011) High spontaneous rate of gene duplication in *Caenorhabditis elegans*. *Current Biology,* 21, 306-310.

LIU, J., LINDERMAN, M., OUYANG, Z., AN, L., YANG, J. & ZHANG, H. (2001) Ecological degradation in protected areas: The case of Wolong nature reserve for giant pandas. *Science,* 292**,** 98-101.

LOCKHART, P., STEEL, M., HENDY, M., WADDELL, P. J. & PENNY, D. (1996) Evolution of chlorophyll and bacteriochlorophyll: the problem of invariant sites in sequence analysis. *Proceedings of the National Academy of Sciences of the United States of America*, 93, 1930-1934.

LOOMIS, W. & SMITH, D. (1990) Molecular phylogeny of *Dictyostelium discoideum* by protein sequence comparison. *Proceedings of the National Academy of Sciences of the United States of America*, 87, 9093-9097.

LOVELL, S. C. & ROBERTSON, D. L. (2010) An integrated view of molecular coevolution in protein–protein interactions. *Molecular Biology and Evolution,* 27**,** 2567-2575.

LÖYTYNOJA, A. & GOLDMAN, N. (2005) An algorithm for the progressive multiple alignment of sequences without insertions. *Proceedings of the National Academy of Sciences of the United States of America*, 102, 10557-10562.

LUO, Z., YUAN, C., MENG, Q. & JI, Q. (2011) A Jurassic Eutherian mammal and divergence of marsupials and placentals. *Nature*, 476, 442-445.

MAERE, S., DE BODT, S., RAES, J., CASNEUF, T., VAN MONTAGU, M., KUIPER, M. & VAN DE PEER, Y. (2005) Modeling gene and genome duplications in eukaryotes. *Proceedings of the National Academy of Sciences of the United States of America,* 102**,** 5454-5459.

MAKINO, T., HOKAMP, K. & MCLYSAGHT, A. (2009) The complex relationship of gene duplication and essentiality. *Trends in Genetics,* 25**,** 152-155.

MAKINO, T. & MCLYSAGHT, A. (2008) Interacting gene clusters and the evolution of the vertebrate immune system. *Molecular Biology and Evolution*, 25, 1855-1862.

MARAIS, G. (2003) Biased gene conversion: implications for genome and sex evolution. *Trends in Genetics*, 19, 330-338.

MARAIS, G. & DURET, L. (2001) Synonymous codon usage, accuracy of translation, and gene length in caenorhabditis elegans. *Journal of Molecular Evolution,* 52**,** 275-280.

MARCET-HOUBEN, M. & GABALDÓN, T. (2011) Treeko: A duplication-aware algorithm for the comparison of phylogenetic trees. *Nucleic Acids Research,* 39**,** e66-e66.

MARDIS, E. R. (2008) The impact of next-generation sequencing technology on genetics. *Trends in Genetics,* 24**,** 133-141.

MARGUSH, T. & MCMORRIS, F. R. (1981) Consensus *n*-trees. *Bulletin of Mathematical Biology,* 43**,** 239-244.

MARQUES-BONET, T., RYDER, O. A. & EICHLER, E. (2009) Sequencing primate genomes: what have we learnt? *Annual Review of Genomics and Human Genetics*, 10, 355-386.

MARIANAYAGAM, N., SUNDE, M. & MATTHEWS, J. (2004) The power of 2: protein dimerization in biology. *Trends in Biochemical Sciences*, 29, 618-625.

MARLAND, E., PRACHUMWAT, A., MALTSEV, N., GU, Z. & LI, W. H. (2004) Higher gene duplicabilities for metabolic proteins than for nonmetabolic proteins in yeast and *E. coli*. *Journal of Molecular Evolution,* 59**,** 806-814.

MARQUES-BONET, T., KIDD, J. M., VENTURA, M., GRAVES, T. A., CHENG, Z., HILLIER, L. D. W., JIANG, Z., BAKER, C., MALFAVON-BORJA, R. & FULTON, L. A. (2009) A burst of segmental duplications in the genome of the african great ape ancestor. *Nature,* 457**,** 877-881.

MATTHEWS, L. R., VAGLIO, P., REBOUL, J., GE, H., DAVIS, B. P., GARRELS, J., VINCENT, S. & VIDAL, M. (2001) Identification of potential interaction networks using sequence-based searches for conserved protein-protein interactions or "interologs". *Genome Research,* 11**,** 2120-2126.

MCKENNA, M. C., BELL, S. K. & SIMPSON, G. G. (1997) Classification of mammals above the species level. In: *Classification of mammals above the species level* City: Columbia Univ Press.

MEREDITH, R. W., JANEČKA, J. E., GATESY, J., RYDER, O. A., FISHER, C. A., TEELING, E. C., GOODBLA, A., EIZIRIK, E., SIMÃO, T. L. L. & STADLER, T. (2011) Impacts of the Cretaceous Terrestrial revolution and K-PG extinction on mammal diversification. *Science,* 334**,** 521-524.

METROPOLIS, N., ROSENBLUTH, A. W., ROSENBLUTH, M. N., TELLER, A. H. & TELLER, E. (1953) Equation of state calculations by fast computing machines. *The Journal of Chemical Physics,* 21**,** 1087.

METZKER, M. L. (2009) Sequencing technologies - the next generation. *Nature Reviews Genetics,* 11**,** 31-46.

METZKER, M. L., MINDELL, D. P., LIU, X. M., PTAK, R. G., GIBBS, R. A. & HILLIS, D. M. (2002) Molecular evidence of HIV-1 transmission in a criminal case. *Proceedings of the National Academy of Sciences of the United States of America,* 99**,** 14292.

MILINKOVITCH, M. C., HELAERS, R., DEPIEREUX, E., TZIKA, A. C. & GABALDÓN, T. (2010) 2× genomes-depth does matter. *Genome Biology,* 11**,** R16.

MORIYAMA, E. N. & POWELL, J. R. (1998) Gene length and codon usage bias in *Drosophila melanogaster*, *Saccharomyces cerevisiae* and *Escherichia coli*. *Nucleic Acids Research,* 26**,** 3188-3193.

MURPHY, W. J., EIZIRIK, E., JOHNSON, W., ZHANG, Y., RYDER, O. & O'BRIEN, S. (2001a) Molecular phylogenetics and the origins of placental mammals. *Nature*, 409, 614-618.

MURPHY, W. J., EIZIRIK, E., O' BRIEN, S. J., MADSEN, O., SCALLY, M., DOUADY, C. J., TEELING, E., RYDER, O., STANHOPE, M. J., DE JONG, W. W. and SPRINGER, M. (2001b) Resolution of the early placental mammal radiation using Bayesian phylogenetics. *Science*. 294, 2348-2351.

MUSTO, H., CRUVEILLER, S., D'ONOFRIO, G., ROMERO, H. & BERNARDI, G. (2001) Translational selection on codon usage in *Xenopus laevis*. *Molecular Biology and Evolution,* 18**,** 1703-1707.

NIIMURA, Y. & NEI, M. (2005) Evolutionary changes of the number of olfactory receptor genes in the human and mouse lineages. *Gene,* 346**,** 23-28.

NIKOLAEV, S. I., MONTOYA-BURGOS, J. I., POPADIN, K., PARAND, L., MARGULIES, E. H. & ANTONARAKIS, S. E. (2007) Life-history traits drive the evolutionary rates of mammalian coding and noncoding genomic elements. *Proceedings of the National Academy of Sciences of the United States of America,* 104**,** 20443-20448.

NIRENBERG, M. W. (2004) Historical review: Deciphering the genetic code–a personal account. *Trends in Biochemical Sciences,* 29**,** 46-54.

NIRENBERG, M. W. & MATTHAEI, J. H. (1961) The dependence of cell free protein synthesis in *E. coli* upon naturally occurring or synthetic polysaccharides. *Proceedings of the National Academy of Sciences of the United States of America,* 47, 1588-1602.

NOVACEK, M. J. (1992) Mammalian phylogeny: Shaking the tree. *March,* 12**,** 121-125.

OHNO, S. (1970) Evolution by Gene Guplication. Eds: George Allen and Unwin, London.

O'HUIGIN, C., SATTA, Y., TAKAHATA, N. & KLEIN, J. (2002) Contribution of homoplasy and of ancestral polymorphism to the evolution of genes in anthropoid primates. *Molecular Biology and Evolution*, 19, 1501-1513.

OWEN, R. (1843) Lectures on the comparative anatomy and physiology of the invertebrate animals. Eds: Longman, Brown, Green & Longmans, 6**,** 34-43, London.

PAGE, R. D. M. (1994) Maps between trees and cladistic analysis of historical associations among genes, organisms, and areas. *Systematic Biology,* 43**,** 58.

PAGE, R. D. M & HOLMES, E. C.(1998) Molecular evolution: A phylogenetic approach. Blackwell Science Ltd, Oxford.

PAGÈS, M., CALVIGNAC, S., KLEIN, C., PARIS, M., HUGHES, S. & HÄNNI, C. (2008) Combined analysis of fourteen nuclear genes refines the Ursidae phylogeny. *Molecular Phylogenetics and Evolution,* 47, 73 – 83.

PANOPOULOU, G. & POUSTKA, A. J. (2005) Timing and mechanism of ancient vertebrate genome duplications–the adventure of a hypothesis. *Trends in Genetics,* 21, 559-567.

PAPP, B., PÁL, C. & HURST, L. D. (2003) Dosage sensitivity and the evolution of gene families in yeast. *Nature,* 424, 194-197.

PAZOS, F. & VALENCIA, A. (2001) Similarity of phylogenetic trees as indicator of protein–protein interaction. *Protein Engineering,* 14, 609-614.

PAZOS, F. & VALENCIA, A. (2008) Protein co-evolution, co-adaptation and interactions. *The EMBO Journal,* 27, 2648-2655.

PEDEN, J. (1997) CodonW. Trinity College, Dublin.

PEREIRA-LEAL, J. B., LEVY, E. D., KAMP, C. & TEICHMANN, S. A. (2007) Evolution of protein complexes by duplication of homomeric interactions. *Genome Biology,* 8, R51.

PERELMAN, P., JOHNSON, W. E., ROOS, C., SEUÁNEZ, H. N., HORVATH, J. E., MOREIRA, M. A. M., KESSING, B., PONTIUS, J., ROELKE, M. & RUMPLER, Y. (2011) A molecular phylogeny of living primates. *PLoS Genetics,* 7**,** e1001342.

PERRIÈRE, G. & THIOULOUSE, J. (2002) Use and misuse of correspondence analysis in codon usage studies. *Nucleic Acids Research,* 30**,** 4548-4555.

PERRY, G. H., DOMINY, N. J., CLAW, K. G., LEE, A. S., FIEGLER, H., REDON, R., WERNER, J., VILLANEA, F. A., MOUNTAIN, J. L. & MISRA, R. (2007) Diet and the evolution of human amylase gene copy number variation. *Nature Genetics,* 39**,** 1256-1260.

PHILIPPE, H. (2000) Opinion: Long branch attraction and protist phylogeny. *Protist,* 151**,** 307-316.

PHILIPPE, H., DELSUC, F., BRINKMANN, H. & LARTILLOT, N. (2005) Phylogenomics. *Annual Review of Ecology, Evolution, and Systematics,* 541-562.

PHILIPPE, H. & LOPEZ, P. (2001) On the conservation of protein sequences in evolution. *Trends in Biochemical Sciences,* 26**,** 414-416.

PHILLIPS, M. J., BENNETT, T. H. & LEE, M. (2009) Molecules, morphology and ecology indicate a recent amphibuous ancestry for echidnas. *Proceedings of the National Academy of Sciences of the United States of America*, 106, 17089-17094.

PHILLIPS, M., DELSUC, F. & PENNY, D. (2004) Genome-scale phylogeny and detection of systematic biases. *Molecular Biology and Evolution*, 21, 1455-1458.

PLOTKIN, J. B. & KUDLA, G. (2010) Synonymous but not the same: The causes and consequences of codon bias. *Nature Reviews Genetics,* 12**,** 32-42.

POE, S. (2003) Evaluation of the strategy of long-branch subdivision to improve the accuracy of phylogenetic methods. *Systematic Biology*, 52, 423-428.

POLLOCK, D. D., ZWICKL, D. J., MCGUIRE, J. & HILLIS, D. M. (2002) Increased taxon sampling is advantageous for phylogenetic inference. *Systematic Biology*, 51, 664-671.

POSADA, D. & BUCKLEY, T. R. (2004) Model selection and model averaging in phylogenetics: Advantages of Akaike Information Criterion and Bayesian approaches over likelihood ratio tests. *Systematic Biology,* 53**,** 793-808.

PRACHUMWAT, A. & LI, W. H. (2006) Protein function, connectivity, and duplicability in yeast. *Molecular Biology and Evolution,* 23**,** 30-39.

PRASAD, A. B., ALLARD, M. W. & GREEN, E. D. (2008) Confirming the phylogeny of mammals by use of large comparative sequence datasets. *Molecular Biology and Evolution*, 25, 1795-1808.

PRUITT, K. D., TATUSOVA, T. & MAGLOTT, D. R. (2007) NCBI reference sequences (refseq): A curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Research,* 35**,** D61-D65.

QIAN, W., YANG, J. R., PEARSON, N. M., MACLEAN, C. & ZHANG, J. (2012) Balanced codon usage optimizes eukaryotic translational efficiency. *PLoS Genetics,* 8**,** e1002603.

RAO, Y., WU, G., WANG, Z., CHAI, X., NIE, Q. & ZHANG, X. (2011) Mutation bias is the driving force of codon usage in the *Gallus gallus* genome. *DNA Research,* 18**,** 499-512.

RASMUSSEN, M. D. & KELLIS, M. (2011) A Bayesian approach for fast and accurate gene tree reconstruction. *Molecular Biology and Evolution,* 28**,** 273-290.

REN, F., TANAKA, H. & YANG, Z. (2009) A likelihood look at the supermatrix–supertree controversy. *Gene,* 441**,** 119-125.

RICE, P., LONGDEN, I. & BLEASBY, A. (2000) EMBOSS: The European Molecular Biology Open Software Suite. *Trends in Genetics,* 16**,** 276-277.

RIPPLINGER, J. & SULLIVAN, J. (2008) Does choice in model selection affect maximum likelihood analysis? *Systematic Biology*, 57, 76-85.

ROKAS, A., WILLIAMS, B. L., KING, N. & CARROLL, S. B. (2003) Genome-scale approaches to resolving incongruence in molecular phylogenies. *Nature,* 425**,** 798-804.

ROMERO, H., ZAVALA, A., MUSTO, H. & BERNARDI, G. (2003) The influence of translational selection on codon usage in fishes from the family Cyprinidae. *Gene,* 317**,** 141-147.

ROTA-STABELLI, O. , KAYAL, E., GLEESON, D., DAUB, J., BOORE, J., TELFORD, M., PISANI, D., BLAXTER, M. & LAVROV, D. (2010) Ecdysozoan mitogenomics: Evidence for a common origin of the legged invertebrates, the Panarthropoda. *Genome Biology and Evolution*, 2, 425-440.

RZHETSKY, A. & NEI, M. (1992) A simple method for estimating and testing minimum-evolution trees. *Molecular Biology and Evolution,* 9**,** 945-967.

RZHETSKY, A. & NEI, M. (1993) Theoretical foundation of the minimum-evolution method of phylogenetic inference. *Molecular Biology and Evolution,* 10**,** 1073-1095.

SAITOU, N. & NEI, M. (1987) The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution,* 4**,** 406-425.

SANDERSON, M. J. (1997) A nonparametric approach to estimating divergence times in the absence of rate constancy. *Molecular Biology and Evolution,* 14**,** 1218-1231.

SANDERSON, M. J., MCMAHON, M. M. & STEEL, M. (2010) Phylogenomics with incomplete taxon coverage: The limits to inference. *BMC Evolutionary Biology,* 10**,** 155.

SANGER, F., NICKLEN, S. & COULSON, R. (1977a) DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America*, 74, 5463-5467.

SANGER, F., AIR, G. M., BARRELL, B. G., BROWN, N. L., COULSON, A. R., FIDDES, J. C., HUTCHISON III, C. A., SLOCOMBE, P. M. & SMITH, M. (1977b) Nucleotide sequence of bacteriophage ϕX174 DNA. *Nature*, 24, 687-695.

SAUNA, Z. E. & KIMCHI-SARFATY, C. (2011) Understanding the contribution of synonymous mutations to human disease. *Nature Reviews Genetics,* 12**,** 683-691.

SCHUSTER, S. C. (2008) Next-generation sequencing transforms today's biology. *Nature,* 18, 16-18.

SCHWARZ, G. (1978) Estimating the dimension of a model. *The Annals of Statistics,* 6**,** 461-464.

SHARP, P. M., AVEROF, M., LLOYD, A. T., MATASSI, G. & PEDEN, J. F. (1995) DNA sequence evolution: The sounds of silence. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences,* 349**,** 241-247.

SHARP, P. M., COWE, E., HIGGINS, D. G., SHIELDS, D. C., WOLFE, K. H. & WRIGHT, F. (1988) Codon usage patterns in *Escherichia coli*, *Bacillus subtilis*, *Saccharomyces cerevisiae*, *Schizosaccharomyces pombe*, *Drosophila melanogaster* and *Homo sapiens*: a review of the considerable within-species diversity. *Nucleic Acids Research,* 16**,** 8207-8211.

SHARP, P. M., TUOHY, T. M. F. & MOSURSKI, K. R. (1986) Codon usage in yeast: Cluster analysis clearly differentiates highly and lowly expressed genes. *Nucleic Acids Research,* 14**,** 5125-5143.

SHENDURE, J. & JI, H. (2008) Next-generation DNA sequencing. *Nature Biotechnology,* 26**,** 1135-1145.

SHIELDS, D. C., SHARP, P. M., HIGGINS, D. G. & WRIGHT, F. (1988) "Silent" sites in drosophila genes are not neutral: Evidence of selection among synonymous codons. *Molecular Biology and Evolution,* 5**,** 704-716.

SHIMODAIRA, H. (2002) An approximately unbiased test of phylogenetic tree selection. *Systematic Biology,* 51**,** 492-508.

SHIMODAIRA, H. & HASEGAWA, M. (2001) Consel: For assessing the confidence of phylogenetic tree selection. *Bioinformatics,* 17**,** 1246-1247.

SOLTIS, P. S. & SOLTIS, D. E. (2003) Applying the bootstrap in phylogeny reconstruction. *Statistical Science*, 18, 256-267.

SPENCER, M., SUSKO, E. & ROGER, A. J. (2005) Likelihood, parsimony, and heterogeneous evolution. *Molecular Biology and Evolution,* 22**,** 1161-1164.

SPRINGER, M. S., BURK-HERRICK, A., MEREDITH, R., EIZIRIK, E., TEELING, E., O'BRIEN, S. J. & MURPHY, W. J. (2007) The adequacy of morphology for reconstructing the early history of placental mammals. *Systematic Biology*, 56, 673-684.

SPRINGER, M. S. & MURPHY, W. J. (2007) Mammalian evolution and biomedicine: new views from phylogeny. *Biology Reviews*, 82, 375-392.

STAMATAKIS, A. (2006) Raxml-vi-hpc: Maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics,* 22**,** 2688-2690.

STARK, C., BREITKREUTZ, B. J., CHATR-ARYAMONTRI, A., BOUCHER, L., OUGHTRED, R., LIVSTONE, M. S., NIXON, J., VAN AUKEN, K., WANG, X. & SHI, X. (2011) The BioGRID interaction database: 2011 update. *Nucleic Acids Research,* 39**,** D698-D704.

STEIPER, M. E. & SEIFFERT, E. R. (2012) Evidence for a convergent slowdown in primate molecular rates and its implications for the timing of early primate evolution. *Proceedings of the National Academy of Sciences of the United States of America,* 109**,** 6006-6011.

STENICO, M., LLOYD, A. T. & SHARP, P. M. (1994) Codon usage in *Caenorhabditis elegans*: Delineation of translational selection and mutational biases. *Nucleic Acids Research,* 22**,** 2437-2446.

STOLETZKI, N. & EYRE-WALKER, A. (2007) Synonymous codon usage in *E. coli*: selection for translational accuracy. *Molecular Biology and Evolution*, 24, 374-381.

STOLL, M., KWITEK-BLACK, A., COWLEY JR., A., HARRIS, E., HARRAP, S., KREIGER, J.,PRINTZ, M., PROVOOST, A., SASSARD, J. & JACOB, H. (2000) New target regions for human hypertension via comparative genomics. *Genome Research*, 10, 473-482.

SU, A. I., WILTSHIRE, T., BATALOV, S., LAPP, H., CHING, K. A., BLOCK, D., ZHANG, J., SODEN, R., HAYAKAWA, M. & KREIMAN, G. (2004) A gene atlas of the mouse and human protein-encoding transcriptomes. *Proceedings of the National Academy of Sciences of the United States of America,* 101**,** 6062-6067.

SUZUKI, Y., GOJOBORI, T. & NEI, M. (2001) ADAPTSITE: Detecting natural selection at single amino acid sites. *Bioinformatics,* 17**,** 660-661.

SWOFFORD, D. L., WADDELL, P. J., HUELSENBECK, J. P., FOSTER, P. G., LEWIS, P. O. & ROGERS, J. S. (2001) Bias in phylogenetic estimation and its relevance to the choice between parsimony and likelihood methods. *Systematic Biology,* 50**,** 525-539.

TALBOT, S. L. & SHIELDS, G. F. (1996) A phylogeny of the bears (Ursidae) inferred from complete sequences of three mitochondrial genes. *Molecular Phylogenetics and Evolution,* 5**,** 567-575.

TAMURA, K., NEI, M. & KUMAR, S. (2004) Prospects for inferring very large phylogenies by using the neighbor-joining method. *Proceedings of the National Academy of Sciences of the United States of America,* 101**,** 11030-11035.

TAVARÉ, S. (1986) Some probabilistic and statistical problems in the analysis of DNA sequences. *Lectures on Mathematics in the Life Sciences,* 17**,** 57-86.

TAYLOR, J. S. & RAES, J. (2004) Duplication and divergence: The evolution of new genes and old ideas. *Annual Review of Genetics*, 38, 615-643.

THENIUS, E. (1979) Zur systematischen und phylogenetischen stellung des bambusbaren: Ailuropoda melanoleuca david (Carnivora, Mammalia). *Z. Saugetuerk.*

THOMPSON, J. D., HIGGINS, D. G. & GIBSON, T. J. (1994) ClustalW: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research,* 22**,** 4673-4680.

THORNE, J. (2003) Multidistribute. Available from the author (http://statgen.ncsu. edu/thorne/multidivtime.html)

THORNE, J. L., KISHINO, H. & PAINTER, I. S. (1998) Estimating the rate of evolution of the rate of molecular evolution. *Molecular Biology and Evolution,* 15**,** 1647-1657.

TIERNEY, L. (1994) Markov chains for exploring posterior distributions. *The Annals of Statistics,* 1701-1728.

TRUETT, G. E., BAHARY, N., FRIEDMAN, J. M. & LEIBEL, R. L. (1991) Rat obesity gene fatty (*fa*) maps to chromosome 5: Evidence for homology with the mouse gene diabetes (*db*). *Proceedings of the National Academy of Sciences of the United States of America,* 88**,** 7806-7809.

URRUTIA, A. O. & HURST, L. D. (2001) Codon usage bias covaries with expression breadth and the rate of synonymous evolution in humans, but this is not evidence for selection. *Genetics,* 159**,** 1191-1199.

URRUTIA, A. O. & HURST, L. D. (2003) The signature of selection mediated by expression on human genes. *Genome Research,* 13**,** 2260-2264.

VAN DER PEER, Y., MAERE, S. & MEYER, A. (2009) The evolutionary significance of ancient genome duplications. *Nature Reviews Genetics*, 10, 725-732.

VAN RHEEDE, T., BASTIAANS, T., BOONE, D. N., HEDGES, S., DE JONG, W & MADSEN, O. (2006) The platypus is in its place: nuclear genes and indels confirm the sister group relation of Monotremes and Therians. *Molecular Biology and Evolution*, 23, 587-597.

VENTER, J. C., ADAMS, M. D., MYERS, E. W., LI, P. W., MURAL, R. J., SUTTON, G. G., SMITH, H. O., YANDELL, M., EVANS, C. A. & HOLT, R. A. (2001) The sequence of the human genome. *Science's STKE,* 291**,** 1304.

VÉZQUEZ, D. P. & GITTLEMAN, J. L. (1998) Biodiversity conservation: Does phylogeny matter? *Current Biology,* 8**,** R379-R381.

VON HAESELER, A. (2012) Do we still need supertrees? *BMC Biology,* 10**,** 13.

WAITS, L. P., SULLIVAN, J., O'BRIEN, S. J. & WARD, R. (1999) Rapid radiation events in the family Ursidae indicated by likelihood phylogenetic estimation from multiple fragments of mtdna. *Molecular Phylogenetics and Evolution,* 13**,** 82-92.

WANG, J., WANG, W., LI, R., LI, Y., TIAN, G., GOODMAN, L., FAN, W., ZHANG, J., LI, J., CHANG, J., GUO, Y., FENG, B., LI, H., FANG, X., LIANG, H., DU., Z., LI, D., ZHAO, Y., HU., Y., YANG., Z., ZHENG, H., HELLMANN, I., INOUYE, M., POOL, J., YI, X., ZHAO, J., DUAN, J., ZHOU, Y., QIAN, J., MA, L., GUOQING, L., YANG, Z., ZHANG, G., YANG, B., YU, C., LIANG, F., LI, W., LI, S., LI, D., NI, P., RUAN, J., LI, Q., ZHU, H., LIU, D., LU, Z., LI, N., GUO, G., ZHANG, J., YE, J., FANG, C., LI, L., ZHOU, K., ZHENG, H., REN, Y., YANG, L., GAO, Y., YANG, G., LI, Z., FENG., X., KRISTIANSEN, K., WONG, G., NIELSON, R., DURBIN, R., BOLUND, L., ZHANG, X., LI, S., YANG, H. & WANG, J. (2008) The diploid genome sequence of an Asian individual. *Nature*, 456, 60-65.

WARNOW, T. (2012) Standard maximum likelihood analyses of alignments with gaps can be statistically inconsistent. *PLoS currents,* 4, RRN1308.

WATSON, J. D. & CRICK, F. H. C. (1953) Molecular structure of nucleic acids. *Nature,* 171**,** 737-738.

WAYNE, R., VAN VALKENBURGH, B. & O'BRIEN, S. (1991) Molecular distance and divergence time in carnivores and primates. *Molecular Biology and Evolution,* 8**,** 297-319.

WHELAN, S., BLACKBURNE, B. P. & SPENCER, M. (2011) Phylogenetic substitution models for detecting heterotachy during plastid evolution. *Molecular Biology and Evolution,* 28**,** 449-458.

WHELAN, S. & GOLDMAN, N. (2001) A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Molecular Biology and Evolution,* 18**,** 691-699.

WHELAN, S., LIÒ, P. & GOLDMAN, N. (2001) Molecular phylogenetics: state-of-the-art methods for looking into the past. *Trends in Genetics*, 17, 262-271.

WIENS, J. J. (2003) Missing data, incomplete taxa, and phylogenetic accuracy. *Systematic Biology,* 52**,** 528-538.

WIENS, J. J. (2006) Missing data and the design of phylogenetic analyses. *Journal of Biomedical Informatics,* 39**,** 34-42.

WIIG, O., AARS, J. & BORN, E. W. (2008) Effects of climate change on polar bears. *Science Progress,* 91**,** 151-173.

WILDMAN, D. E., UDDIN, M., OPAZO, J. C., LIU, G., LEFORT, V., GUINDON, S., GASCUEL, O., GROSSMAN, L., ROMERO, R. & GOODMAN, M. (2007) Genomics, biogeography and the diversification of placental mammals. *Proceedings of the National Academy of Sciences of the United States of America*. 36, 14395-14400.

WILES, A. M., DODERER, M., RUAN, J., GU, T. T., RAVI, D., BLACKMAN, B. & BISHOP, A. J. R. (2010) Building and analyzing protein interactome networks by cross-species comparisons. *BMC Systems Biology,* 4**,** 36.

WILSON, D. E., REEDER, D.M. (2005) Mammal species of the world: A taxonomic and geographic reference. In: *Mammal species of the world: A taxonomic and geographic reference* City: John Hopkins University Press, Baltimore.

WOESE, C., ACHENBACH, L., ROUVIERE, P., & MANDELCO, L. (1991) Archael phylogeny: re-examination of the phylogenetic position of *Archaeaglobus fulgidas* in light of certain composition induced artifacts. *Systematic and Applied Microbiology*, 14, 364-371.

WOLFE, K. & Li, W. H. (2003) Molecular evolution meets the genomics revolution. *Nature*, 33, 255-265.

WOLFE, K. H., LI, W. H. & SHARP, P. M. (1987) Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast, and nuclear DNAs. *Proceedings of the National Academy of Sciences of the United States of America,* 84**,** 9054-9058.

WOROBEY, M., SANTIAGO, M. L., KEELE, B. F., NDJANGO, J. B. N., JOY, J. B., LABAMA, B. L., DHED'A, B. D., RAMBAUT, A., SHARP, P. M. & SHAW, G. M. (2004) Origin of AIDS: Contaminated polio vaccine theory refuted. *Nature,* 428**,** 820-820.

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

WRIGHT, S. (1931) Evolution in mendelian populations. *Genetics,* 16**,** 97-159.

WU, X., HART, H., CHENG, C., ROACH, P. & TATCHELL, K. (2001) Characterization of Gac1p, a regulatory subunit of protein phosphatase type 1 involved in glycogen accumulation in *S. cerevisiae*. *Molecular Genetics and Genomics*, 265, 622-635.

YANG, J. & LI, W. H. (2004) Developmental constraint on gene duplicability in fruit flies and nematodes. *Gene,* 340**,** 237-240.

YANG, Z. (1996a) Among-site rate variation and its impact on phylogenetic analyses. *Trends in Ecology & Evolution,* 11**,** 367-372.

YANG, Z. (1996b) Phylogenetic analysis using parsimony and likelihood methods. *Journal of Molecular Evolution,* 42**,** 294-307.

YANG, Z. & NIELSEN, R. (1998) Synonymous and nonsynonymous rate variation in nuclear genes of mammals. *Journal of Molecular Evolution,* 46**,** 409-418.

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

YANG, Z. & RANNALA, B. (2006) Bayesian estimation of species divergence times under a molecular clock using multiple fossil calibrations with soft bounds. *Molecular Biology and Evolution*, 23, 212-226.

YANG, Z. & RANNALA, B. (1997) Bayesian phylogenetic methods using DNA sequences: A Markov Chain Monte Carlo method. *Molecular Biology and Evolution*, 14, 717-724.

YANG, Z. & RANNALA, B. (2012) Molecular phylogenetics: Principles and practice. *Nature Reviews Genetics,* 13**,** 303-314.

YI, S., ELLSWORTH, D. L. & LI, W. H. (2002) Slow molecular clocks in old world monkeys, apes, and humans. *Molecular Biology and Evolution,* 19**,** 2191-2198.

YOKOYAMA, S. (1994) Gene duplications and evolution of the short wavelength-sensitive visual pigments in vertebrates. *Molecular Biology and Evolution,* 11**,** 32-39.

YU, L., LI, Q., RYDER, O. & ZHANG, Y. (2004) Phylogeny of the bears (Ursidae) based on nuclear and mitochondrial genes. *Molecular Phylogenetics and Evolution,* 32**,** 480-494.

YU, L., LI, Y. W., RYDER, O. & ZHANG, Y. P. (2007) Analysis of complete mitochondrial genome sequences increases phylogenetic resolution of bears (Ursidae), a mammalian family that experienced rapid speciation. *BMC Evolutionary Biology,* 7**,** 198.

ZAVALA, A., NAYA, H., ROMERO, H. & MUSTO, H. (2002) Trends in codon and amino acid usage in *Thermotoga maritima*. *Journal of Molecular Evolution,* 54**,** 563-568.

ZHANG, F., GU, W., HURLES, M. E. & LUPSKI, J. R. (2009) Copy number variation in human health, disease, and evolution. *Annual Review of Genomics and Human Genetics,* 10**,** 451-481.

ZHANG, J. (2003) Evolution by gene duplication: An update. *Trends in Ecology & Evolution,* 18**,** 292-298.

ZHANG, L, HUA, N. & SUN, S. (2008) Wildlife trade, consumption and conservation awareness in Southwest China. *Biodiversity and Conservation*, 17, 1493-1516.

ZHANG, Z., LI, J., CUI, P., DING, F., LI, A., TOWNSEND, J. P. & YU, J. (2012) Codon Deviation Coefficient: A novel measure for estimating codon usage bias and its statistical significance. *BMC Bioinformatics,* 13**,** 43.

ZHANG, Z. & YU, J. (2010) Modeling compositional dynamics based on GC and purine contents of protein-coding sequences. *Biology Direct,* 5**,** 1-15.

ZHAO, Z., JIN, L., FU, Y. X., RAMSAY, M., JENKINS, T., LESKINEN, E., PAMILO, P., TREXLER, M., PATTHY, L. & JORDE, L. B. (2000) Worldwide DNA sequence variation in a 10-kilobase noncoding region on human chromosome 22. *Proceedings of the National Academy of Sciences of the United States of America,* 97**,** 11354-11358.

ZHOU, T., WEEMS, M. & WILKE, C. O. (2009) Translationally optimal codons associate with structurally sensitive sites in proteins. *Molecular Biology and Evolution,* 26**,** 1571-1580.

ZUCKERKANDL, E. & PAULING, L. (1962) Molecular disease, evolution and genetic heterogeneity. In: *Horizons in Biochemistry*, 189-225. Kasha, M. Pullman, B. (eds). Academic Press, New York.

ZUCKERKANDL, E. & PAULING, L. (1965) Molecules as documents of evolutionary history. *Journal of Theoretical Biology,* 8**,** 357-366.

Appendix

Appendix A1: Bayes Factor Analysis

Each model versus the deconstrained model (that was generated from multidistribute):
***UGam* = 0.045054 [-0.393527: 0.097022]**
***Ln* = -0.54 [-0.587666: -0.29091]**
***CIR* = -0.05 [-0.08: 0.102636]**

Equation that was used to compare models:

$$\frac{Model\ 1}{Model\ 2} = model1 - model2\ [\ model1_{min} - model2_{max} : model1_{max} - model2_{min}\ ]$$

Comparison of each molecular clock model to every other model:

UGam versus CIR:

$$\frac{UGAM}{CIR} = 0.045054 + 0.05\ [-0.393527 - 0.102636: 0.097022 + 0.08]$$

$$\frac{UGAM}{CIR} = 0.095054\ [-0.496163: 0.177022]$$

UGam versus LN:

$$\frac{UGAM}{LN} = 0.045054 + 0.54\ [-0.393527 + 0.29091: 0.097022 + 0.587666]$$

$$\frac{UGAM}{LN} = 0.585054\ [-0.102617: 1.557886]$$

CIR versus LN:

$$\frac{CIR}{LN} = -0.05 + 0.54\ [-0.08 + 0.29091: 0.102636 + 0.587666]$$

$$\frac{CIR}{LN} = 0.49\ [0.21091: 0.690302]$$

Publication

Increased Genome Sampling Reveals a Dynamic Relationship between Gene Duplicability and the Structure of the Primate Protein–Protein Interaction Network

Aoife Doherty,† David Alvarez-Ponce,† and James O. McInerney*

Department of Biology, National University of Ireland Maynooth, Maynooth, County Kildare, Ireland

†These authors contributed equally to this work.

***Corresponding author:** E-mail: james.o.mcinerney@nuim.ie.

Associate Editor: Michael Purugganan

Abstract

Although gene duplications occur at a higher rate, only a small fraction of these are retained. The position of a gene's encoded product in the protein–protein interaction network has recently emerged as a determining factor of gene duplicability. However, the direction of the relationship between network centrality and duplicability is not universal: In *Escherichia coli*, yeast, fly, and worm, duplicated genes more often act at the periphery of the network, whereas in humans, such genes tend to occupy the most central positions. Herein, we have inferred duplication events that took place in the different branches of the primate phylogeny. In agreement with previous observations, we found that duplications generally affected the most central network genes, which is presumably the process that has most influenced the trend in humans. However, the opposite trend—that is, duplication being more common in genes whose encoded products are peripheral in the network—is observed for three recent branches, including, quite counterintuitively, the external branch leading to humans. This indicates a shift in the relationship between centrality and duplicability during primate evolution. Furthermore, we found that genes encoding interacting proteins exhibit phylogenetic tree topologies that are more similar than expected for random pairs and that genes duplicated in a given branch of the phylogeny tend to interact with those that duplicated in the same lineage. These results indicate that duplication of a gene increases the likelihood of duplication of its interacting partners. Our observations indicate that the structure of the primate protein–protein interaction network affects gene duplicability in previously unrecognized ways.

Key words: gene duplicability, protein–protein interaction network, network evolution, primate evolution.

Introduction

One of the key insights provided by fully sequenced genomes is the pervasiveness of gene duplication and loss in all organisms (Ohno 1970; Zhang 2003), which has resulted in modern-day genomes being replete with multigene families and a confusing pattern of orthologs and paralogs distributed throughout life on the planet. However, genes widely differ in their propensity to retain duplicates, whereas some gene families are represented by dozens or even hundreds of members in a given genome, others remain as singleton genes over time. This observation naturally leads to the key question concerning why genes duplicate and the even bigger question of what constraints exist that might prevent duplications from occurring or at least retard their rate of occurrence.

Gene duplication is often the key for understanding the origin and evolution of important advantageous traits. For example, the acquisition of color vision in vertebrates is the result of the duplication of retinal visual pigment genes (Yokoyama 2002), and salivary amylase gene copy number is positively correlated with dietary starch intake in human populations (Perry et al. 2007). On the other hand, gene duplication is a significant factor in the pathogenesis of various diseases such as cancer (Slamon et al. 1987; Lahortiga et al. 2007). A duplicated gene is very likely to be lost unless it offers a selective advantage to the organism in which it is found, and therefore, only a fraction of duplicated genes are retained after duplication (Ohno 1970; Lipinski et al. 2011). Over the past decade, the combination of genomic and functional data has allowed us to identify the factors correlating with gene duplicability, that is, the tendency to retain both gene copies after duplication. These factors include gene function (Marland et al. 2004) and complexity (Papp et al. 2003; Yang et al. 2003; He and Zhang 2005), subcellular location (Prachumwat and Li 2006), and timing of expression during development (Castillo-Davis and Hartl 2002; Yang and Li 2004). Yet, a large fraction of the variability of gene duplicability remains unexplained.

Genes and proteins rarely act in isolation, and over the past few years, in particular, we have been gaining a better understanding of the complex networks of interactions in which these molecules find themselves. The high throughput accumulation of interactomic data now allows us to investigate the relationship between the patterns of molecular evolution of genes and the position that their encoded products occupy in protein–protein interaction networks (PINs) (see Cork and

© The Author 2012. Published by Oxford University Press on behalf of the Society for Molecular Biology and Evolution. All rights reserved. For permissions, please e-mail: journals.permissions@oup.com

Research article

Downloaded from http://mbe.oxfordjournals.org/ at National University of Ireland, Maynooth on March 7, 2013

Purugganan 2004; Eanes 2011; Zera 2011; Alvarez-Ponce et al. 2012). The position of a protein in the network can be measured from its network centrality, which can be computed as its degree (number of proteins with which it interacts), betweenness (number of shortest paths between protein pairs to which it belongs), or closeness (the inverse of the average distance to all other proteins in the network) (Borgatti 2005; Mason and Verwoerd 2007). Some aspects of the evolution of genes have been shown to be affected by the centrality of their encoded products in the PIN (e.g., Luisi et al. 2012). For instance, genes occupying the most central positions tend to be more selectively constrained (Fraser et al. 2002; Hahn and Kern 2005; Lemos et al. 2005). Although gene duplicability is also affected by centrality, the direction of the relationship between centrality and duplicability is not universal. In *E. coli*, yeast, and fly, singleton genes tend to occupy more central positions in the network than duplicated genes (Hughes and Friedman 2005; Prachumwat and Li 2006; Makino et al. 2009). A possible explanation for this phenomenon is that duplication of a gene may disrupt the dosage balance of the interactions in which it is involved (Veitia 2002; Papp et al. 2003), and this may have more deleterious effects for the most highly connected genes. Conversely, duplicated genes tend to be more central than singleton genes in the human PIN (Liang and Li 2007), which is a derived character resulting from the high duplicability of metazoan-specific genes (D'Antonio and Ciccarelli 2011). However, it remains unclear why this different pattern is observed in humans. These contrasting observations indicate that, although network position has a clear effect on a gene's duplicability, the relationship between duplicability and PIN centrality has undergone modification in the vertebrate lineage. This dynamic behavior of the relationship between centrality and duplicability opens the question of whether more shifts have taken place during evolution and, if so, how often did they occur and when.

Further evidence for the dependence between the position of genes in an interaction network and their patterns of evolution comes from the observation that genes encoding interacting proteins tend to exhibit correlated evolutionary histories (for a review, see Lovell and Robertson 2010). For example, their rates of evolution are more similar than expected from random protein pairs (Fraser et al. 2002; Lemos et al. 2005; Alvarez-Ponce et al. 2011). This similarity is generally attributed to molecular co-evolution or to interacting proteins being subject to similar evolutionary forces and it can be potentially used to infer protein–protein interactions from sequence data (Codoñer and Fares 2008; Fares et al. 2011). For instance, several studies have shown that interacting genes manifest phylogenetic histories that are more similar than expected in a random network, as evidenced by the similarity in the lengths of the branches in the phylogeny. However, gene tree similarities have usually been assessed using the mirrortree approach, which relies on the underlying distance matrices (Goh et al. 2000; Pazos and Valencia 2001; Pazos et al. 2008). It is less clear whether the actual phylogenetic trees inferred from interacting proteins are more topologically similar than expected from random protein pairs. In fact, Kelly and Stumpf (2010) found only negligible evidence for such an increased level of similarity between pairs of trees inferred from interacting proteins in sets of yeast orthologous sequences. However, both the mirrortree approach and the approach used by Kelly and Stumpf rely on sets of 1:1 orthologs. Although computationally convenient, this approach does not address the potential gene tree similarity resulting from similar duplication histories. Almost 20 years ago, it was hypothesized that interacting genes may tend to exhibit topologically similar phylogenetic trees owing to co-duplication at similar evolutionary times (Fryxell 1996). Arguably, duplication of a gene with interacting partners may be deleterious unless the interacting genes co-duplicate soon after or before the event (Papp et al. 2003). Alternatively, the functional diversification of duplicated genes could be facilitated by a pre-existing heterogeneity in proteins that interact with their products (Fryxell 1996). Although a number of examples of correlated tree topologies for interacting genes have been reported (e.g., Fryxell 1996; Koretke et al. 2000; Alvarez-Ponce et al. 2009), an analysis at the level of the entire interactome has not been carried out to date.

Herein, we combine comparative genomics and protein–protein interaction data to explore the relationship between the structure of the primate PIN and the duplicability of genes encoding its components. For that purpose, we inferred the gene duplication events that took place in each of the branches of a phylogeny consisting of six primates and one rodent and evaluated the dependence between the duplicability of genes and the position of their encoded products in the PIN. The results revealed a complex relationship between network position and duplicability. We found that 1) in agreement with previous observations, duplicated genes act at the most central positions of the human PIN; however, when we examined the trend across different portions of the primate phylogeny, the opposite (i.e., gene duplication preferentially affecting genes whose encoded products are peripheral in the PIN) was observed for genes duplicated in the external branch leading to humans and the two internal branches subtending the human/chimpanzee and the human/chimpanzee/gorilla clades, indicating that the relationship between duplicability and centrality has undergone modification more than once during animal evolution; 2) genes encoding interacting proteins exhibit more similar tree topologies than expected in a random network; and 3) genes that duplicated in a given branch tend to interact with genes that duplicated in the same branch, indicating that the duplication of a gene increases the likelihood of duplication of its interacting partners in the network. Taken together, these results indicate that the structure of the primate network constrains the patterns of duplication of their components at multiple levels and in a dynamic manner.

Materials and Methods

Genomic Data

We retrieved all protein-coding sequences (CDSs) and family assignments for human, chimpanzee, gorilla, orangutan,

Downloaded from http://mbe.oxfordjournals.org/ at National University of Ireland, Maynooth on March 7, 2013

macaque, marmoset, and mouse from the Ensembl database (version 61; Flicek et al. 2011). We eliminated the following from our analyses: 1) coding sequences that were interrupted by a stop codon or whose length was not a multiple of three; 2) sequences that had not been assigned to any gene family; and 3) gene families consisting of less than four sequences. After this filtering, a dataset comprising 125,999 genes belonging to 12,158 gene families was retained.

Phylogenetic Tree Reconstruction and Duplication Inference

For each gene family, we aligned the protein sequences using MUSCLE (Edgar 2004). The resulting protein alignments were used to guide the alignment of the corresponding CDSs using TranslatorX (Abascal et al. 2010). The CDS alignments were subsequently used to reconstruct Bayesian phylogenetic trees using SPIMAP (Rasmussen and Kellis 2011). Gene duplications were inferred using the species/gene tree reconciliation approach implemented in the SPIMAP software and the species overlap method (Huerta-Cepas et al. 2007; Gabaldón 2008) implemented in the ETE package (Huerta-Cepas et al. 2010). For these analyses, we used the reference species tree provided in the study by Benton et al. (2009) (fig. 1).

We assigned each duplication event to a branch of the reference species tree and to one or more human genes. For each duplication node, we examined the species represented in the descendant leaves. We assigned the duplication event to the branch preceding the deepest node in the reference species tree whose descendants include all the species affected by the duplication. For instance, if sequences descending from a duplication node included sequences from all great apes, the duplication event was assigned to the branch subtending the radiation of the great apes. Subsequently, we assigned the duplication event to the set of human genes that are the result of this duplication event or are the closest human homologs to the genes involved in this duplication. If there was at least one human gene in the set of descendant leaves of a duplication node, the duplication event was assigned to this human gene or set of genes. Otherwise, we systematically examined the parental node of that node until the descendant leaves contained at least one human homolog.

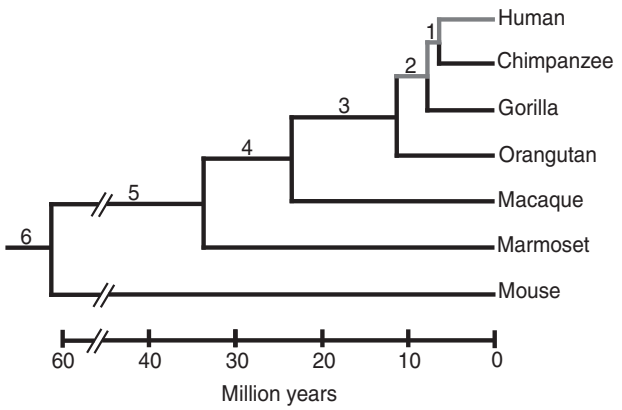


FIG. 1. Phylogeny of the species included in the analysis. Divergence dates were retrieved from the study by Benton et al. (2009). The number above each internal branch (1–6) is the name that we have assigned to that branch. Branches for which duplicated genes tend to be less connected than nonduplicated genes are represented in gray.

Network-Level Analysis

The human interactome was assembled from the interactions available from the BioGRID database version 3.1.81 (Stark et al. 2011). Only nonredundant physical interactions among pairs of human proteins with an Ensembl ID were considered. The network (termed PIN0) contains 9,087 proteins connected by 39,883 interactions. For each protein, degree was computed as the number of interacting partners, and betweenness and closeness centralities were computed using the NetworkX package (http://networkx.lanl.gov/). Proteins not represented in the PIN0 network were not used in network-level analyses.

We evaluated whether the phylogenetic trees of genes encoding interacting proteins were more similar than expected in a random network and whether genes that duplicated in a given branch of the species tree tend to interact with genes that duplicated in the same branch. For that purpose, a subnetwork containing only proteins with a nontrivial tree was used (PIN1; supplementary fig. S1, Supplementary Material online). We used as statistics the average tree topological similarity of interacting proteins (see below) and the number of interactions among proteins encoded by duplicated genes. The statistical significance of measured network parameters was evaluated from an ensemble of 250 or 10,000 randomized networks. Random networks were generated using a network rewiring approach. Each random network was generated from PIN1 by repeatedly choosing two edges at random (e.g., A–B and C–D) and swapping them (yielding A–D and C–B, or A–C and B–D). This operation was iterated $100 \times m$ times on each random network, where m is the number of edges. Therefore, each random network contains the same nodes, the same number of edges, and the same degree for each node as the original network. P values were computed as the proportion of random networks with a parameter value higher or equal to the observed one.

To discard the potential impact of confounding network features in our results, analyses were repeated on two subnetworks of PIN1. PIN2 is a subnetwork of PIN1 with no self-interactions or interactions among proteins encoded by paralogous genes; and PIN3 is a subnetwork of PIN2 without interactions among proteins encoded by genes locating in the same chromosome arm (supplementary fig. S1, Supplementary Material online). A separate network ensemble was generated for each of these networks. The same restrictions used to generate each subnetwork were imposed to the corresponding ensembles, only allowing edge swaps respecting these restrictions.

Comparison of Phylogenetic Trees

We used the "tree comparison" program from the treeKO package (Marcet-Houben and Gabaldón 2011) to compare

Downloaded from http://mbe.oxfordjournals.org/ at National University of Ireland, Maynooth on March 7, 2013

the tree topologies of pairs of interacting proteins. The "strict distance" was used. Trees were rooted in the branch that minimized the number of gene duplications in the tree.

Age of Human Genes

To establish the age of each human gene, we carried out a similarity search against the nr database (downloaded on 12 October 2010; Pruitt et al. 2007) using the BLASTP algorithm (Altschul et al. 1990). Only genes that aligned to more than 80% of the query sequence were retained. If at least 5% of the hits corresponded to nonmetazoan genomes, the human gene was considered to be of premetazoan ancestry (i.e., "ancient").

Results

Identifying Duplication Events in the Primate Phylogeny

We retrieved all CDSs for six primates (human, chimpanzee, gorilla, orangutan, macaque, and marmoset) and one rodent (mouse). After filtering the dataset (see Materials and Methods), we retained a total of 125,999 genes belonging to 12,158 gene families (supplementary table S1, Supplementary Material online). For each family, we reconstructed a phylogenetic tree using a Bayesian approach. Using these phylogenetic trees, we inferred the duplication and loss events that took place during the evolution of each family using the gene tree/species tree reconciliation approach (Goodman et al. 1979; Page 1994). This algorithm compares each gene tree with an established species tree topology (fig. 1), and discrepancies between the two are attributed to duplication or loss events. Because inference of gene losses is methodologically problematic (Hahn 2007), only gene duplications are considered in the current analysis. In addition to the gene tree/species tree reconciliation approach, we used the reconciliation-independent species overlap method (Huerta-Cepas et al. 2007; Gabaldón 2008), which is based on the pattern of species overlap in the descendant leaves of each duplication node. The gene tree/species tree reconciliation approach inferred a total of 22,969 duplications across the studied phylogeny, whereas the more conservative species overlap method inferred 15,814 duplications (table 1 and supplementary table S2, Supplementary Material online). Unless otherwise stated, the results reported throughout this article correspond to duplications inferred using the gene tree/species tree reconciliation method; however, we carried out all analyses in parallel using both approaches, with qualitatively equivalent results. These results are detailed in the relevant tables, and all analyses and data are available in the Supplementary Material online or on request from the authors.

We estimated an overall gene duplication rate of 0.00348 duplications/gene/My across the phylogeny of the studied species. However, we found that the duplication rate varied widely across the different branches of the tree, ranging from 0.0012 duplications/gene/My on the chimpanzee external branch to 0.0252 duplications/gene/My on the internal branch subtending the human, chimpanzee, and gorilla clade (labeled as branch 2; see table 1). This represents a greater than 20-fold difference in duplication rate between these branches. The remarkable acceleration in the rate of gene duplication in branch 2 has been described previously and has been suggested to be the result of changes in the effective population size or the generation time during the evolution of the great apes (Marques-Bonet et al. 2009). In agreement with previous reports (Hahn et al. 2007), we observed an increased rate of gene duplication in the primate lineage (0.00388 duplications/gene/My) compared with the mouse branch (0.0018 duplications/gene/My). Furthermore, we observed an increased rate of duplication in the great apes (0.0041 duplications/gene/My) compared with the average rate in primates, also consistent with previous observations (Fortna et al. 2004; Hahn et al. 2007).

Of particular interest in the assessment of gene duplication is the issue of what kinds of genes have duplicated and in which evolutionary time. For each branch in the species tree, we obtained a list of human genes that are either the result of

Table 1. Summary Statistics for Each Branch of the Studied Phylogeny (Species/Gene Tree Reconciliation Method).

Branch Name	Branch Length (My)	Number of Duplications	Rate of Duplication	Number of Human Homologs	Ancient Human Homologs (%)	Ancient Human Homologs in PIN0 (%)
Human	6.5	495	0.0037	790	9.49	18.95
Chimpanzee	6.5	157	0.0012	217	10.60	30.77
Gorilla	8.0	424	0.0025	426	22.06	38.79
Orangutan	11.2	292	0.0013	342	22.22	41.82
Macaque	23.5	902	0.0018	731	32.15	45.41
Marmoset	33.7	1,526	0.0021	1,043	30.97	39.99
Mouse	61.5	2,584	0.0018	796	12.81	28.38
Branch 1	1.5	90	0.0030	179	17.32	26.47
Branch 2	3.2	1,655	0.0252	1,805	23.16	29.13
Branch 3	12.3	1,770	0.0071	1,906	21.30	26.99
Branch 4	10.2	2,127	0.0099	2,274	23.04	31.25
Branch 5	27.8	3,220	0.0055	3,108	21.30	26.15
Branch 6	—	7,727	—	8,668	20.72	22.83

Downloaded from http://mbe.oxfordjournals.org/ at National University of Ireland, Maynooth on March 7, 2013

duplication events that occurred in that branch or the closest human homologs to the genes involved in the duplications that occurred at that branch (see Materials and Methods). We considered whether each of the resulting gene lists was enriched in certain Gene Ontology (GO; Ashburner et al. 2000) terms. For that purpose, we compared the frequency of each GO term in the list of duplicated genes with the rest of the human genome using the FatiGO software (Al-Shahrour et al. 2004), which specifically seeks to find significant associations between GO terms and lists of genes. A total of 67 unique biological processes are enriched among genes duplicated in any of the external branches of the phylogeny (supplementary table S3, Supplementary Material online). In general, we observed enrichment in the "reproduction," "transcription," "translation," and "environment perception" GO categories, in agreement with previous results (e.g., Demuth and Hahn 2009; Huerta-Cepas et al. 2007). Interestingly, we observed a clear enrichment in GO categories associated with olfactory transduction in mouse-specific duplications, as reported previously by Niimura and Nei (2005, 2007).

From these results, we can conclude that our dataset and treatments of the data are in line with previous work. In this study, we have conducted an interactome-wide analysis of gene duplication.

The Relationship between Centrality and Duplicability Underwent Modification during Primate Evolution

Having identified the genes that underwent duplication in each branch of the phylogeny of the studied species, we sought to investigate the relationship between the structure of the network and the duplicability of its components. For that purpose, we assembled a human interactome (termed PIN0) from all interactions available in the BioGRID database (Stark et al. 2011). For each gene in the network, we computed three centrality measures (degree, betweenness, and closeness) and compared their values for nonduplicated genes and genes that underwent duplication in any branch of the phylogeny. In agreement with previous observations in the human interactome (Liang and Li 2007; D'Antonio and Ciccarelli 2011), we found that duplicated genes occupy more central positions in the human PIN than nonduplicated genes using the Mann–Whitney U test ($P = 2.89 \times 10^{-13}$ for degree; $P = 3.01 \times 10^{-10}$ for betweenness; and $P = 2.11 \times 10^{-14}$ for closeness; supplementary table S4, Supplementary Material online, and fig. 2). Crucially, however, this analysis only takes into account whether genes underwent duplication in any branch of the phylogeny, and therefore, it does not consider the specific branches on the species tree in which those duplications occurred. A more interesting analysis is to consider duplications that happened at approximately the same time and whether different parts of the interactome were perturbed by duplication at different times.

We conducted an analysis that partitioned duplication events into the branches in the phylogeny in which they occurred. We observed that duplicated genes exhibit a higher average degree (i.e., number of interacting partners) than nonduplicated genes in 10 of the 13 branches of the species tree, with statistically significant differences in 5 of the branches (supplementary table S4, Supplementary Material online, and fig. 2). Unexpectedly, the opposite trend (i.e., a higher degree for nonduplicated genes) is observed in the three remaining branches (the external branch leading to the human lineage and internal branches 1 and 2), with statistically significant differences in two of these branches (the human branch and internal branch 1; supplementary table S4, Supplementary Material online, and fig. 2). We obtained similar results when we used betweenness and closeness as the measures of network centrality and when the more conservative species overlap method was used as the method for inferring duplication events (supplementary table S4, Supplementary Material online). Therefore, despite the general tendency of duplications to occur at the most central genes of the network, the relationship between centrality and duplicability has inverted during the primate radiation.

These observations presents us with a picture of the relationship between network position and gene duplicability

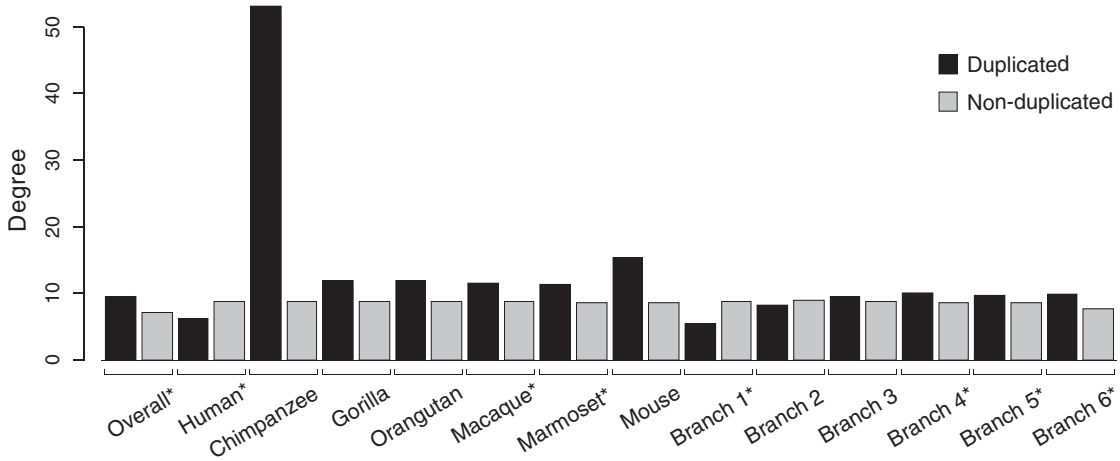


Fig. 2. Degree for proteins encoded by duplicated and nonduplicated genes in each branch of the phylogeny. Results for the corresponding statistical tests are reported in supplementary table S4, Supplementary Material online. The asterisk indicates statistically significant differences.

Downloaded from http://mbe.oxfordjournals.org/ at National University of Ireland, Maynooth on March 7, 2013

that is more complex than has been reported previously and that up to now was assumed to be the general rule for vertebrates. To gain a more complete understanding into the relationship between the structure of the network and the patterns of duplication of its components, we considered whether duplications of genes encoding interacting proteins were correlated.

Interacting Proteins in the Human PIN Tend to Exhibit Topologically Similar Phylogenetic Trees

For each pair of interacting proteins, we compared the topologies of the corresponding phylogenetic trees using the treeKO algorithm (Marcet-Houben and Gabaldón 2011). We used the "strict distance," which takes into account both the patterns of speciation and the duplication and loss patterns. For this analysis, we used a subnetwork of PIN0 (termed PIN1; see supplementary fig. S1, Supplementary Material online) that contained only proteins encoded by genes capable of reconstructing nontrivial phylogenetic trees (those belonging to gene families with four or more members). According to the treeKO algorithm, trees derived from interacting proteins exhibit an average distance of $D = 0.319$. To assess the significance of this value, we compared it with a null distribution obtained from a set of randomized networks with the same nodes, number of interactions, and degree for each node as the original network (PIN1; see Materials and Methods). Of 250 randomized versions of PIN1, none showed a D value lower than or equal to the observed one (average value for the simulations, $D = 0.339$; $P < 0.004$; fig. 3), indicating that interacting proteins in the human interactome manifest tree topologies that are more similar than expected from a random network.

This similarity might be the result of molecular coevolution of genes encoding interacting proteins. However, a number of features of PINs might also produce such a similarity, and their effects should be ameliorated as much as possible to eliminate potential sources of confounding bias. First, PINs are known to be enriched in self-interactions (i.e., interactions among identical proteins) and interactions among proteins encoded by paralogous genes (Ispolatov et al. 2005; Pereira-Leal et al. 2007; Alvarez-Ponce and McInerney 2011). Because genes involved in these interactions are represented in the same phylogenetic trees, this enrichment could potentially contribute to the low observed D value. To discard this possibility, analyses were repeated in a subnetwork of PIN1 in which all such interactions were removed (PIN2; see supplementary fig. S1, Supplementary Material online). We found that interacting proteins still exhibit a higher similarity than expected in a random network ($D = 0.331$; average value for the simulations, $D = 0.338$; $P < 0.004$), indicating that these features do not affect our observations. Second, duplication events sometimes affect large chromosomal regions, thereby involving simultaneous duplication of multiple adjacent genes, which would consequently have similar duplication histories. In addition, genes encoding interacting proteins tend to cluster together in the genome (Lee and Sonnhammer 2003; Makino and McLysaght 2008). Taken together, these tendencies may also contribute to the similarity in tree topologies observed among interacting proteins. However, the topological similarity of trees in the observed interactome is still significantly higher than expected at random when interactions involving proteins encoded by genes that localize to the same chromosome arm are also removed ($D = 0.331$; average value for the simulations, $D = 0.338$; $P < 0.004$ for PIN3; supplementary fig. S1, Supplementary Material online).

These results indicate that genes encoding interacting proteins manifest more similar tree topologies than expected from random pairs and that this pattern is independent of the enrichment of the network in self-interactions, interactions among paralogous genes, and interactions among genes that co-localize in the genome. This similarity can potentially be the result of genes that encode interacting proteins exhibiting similar duplication histories. To test this possibility, we investigated whether the duplications of interacting genes tend to occur in the same branches of the species tree.

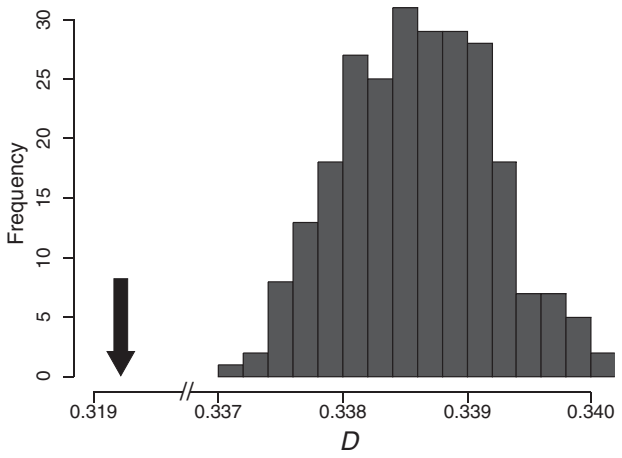


FIG. 3. Topological distance between the trees corresponding to pairs of interacting proteins in the human interactome. The observed value in the actual interactome (PIN1) is represented as an arrow, and the distribution inferred from 250 randomized networks is represented as a histogram.

Genes Encoding Interacting Proteins Tend to Co-duplicate in the Same Branches of the Phylogeny

We considered whether the human interactome was enriched in interactions among proteins encoded by duplicated genes. For that purpose, we computed the number of interactions involving genes that have undergone duplication in any branch of the phylogeny ($N = 22,988$ in PIN1) and compared this number to the null distribution obtained from a collection of 10,000 random networks. None of these random networks exhibits an N value higher than or equal to the observed one ($P < 0.0001$), indicating that duplicated genes tend to interact with each other in the real network. This result holds when self-interactions and interactions among paralogs ($N = 21,872$; $P < 0.0001$ for PIN2), and interactions between genes locating in the same

Downloaded from http://mbe.oxfordjournals.org/ at National University of Ireland, Maynooth on March 7, 2013

chromosome arm ($N = 21,152$; $P < 0.0001$ for PIN3), are removed from the analyses.

We carried out an equivalent analysis for genes duplicated in each of the 13 branches of the studied phylogeny; that is, we examined whether genes that duplicated in a given branch tend to interact with genes that duplicated in the same branch. For each branch *i*, we computed the number of interactions between genes that underwent duplication in that branch, N_i, and evaluated its statistical significance as above. When all interactions are considered (PIN1), the N_i values are significantly higher than expected from a random network in all 13 branches ($P < 0.05$; supplementary table S5, Supplementary Material online), indicating that genes that have undergone duplication in each of these branches tend to interact with each other. When self-interactions and interactions among paralogs are removed (PIN2), the N_i values are higher than the average values for the random networks for 10 of the 13 branches, with statistically significant differences in 4 of the branches (the external branches leading to gorilla, marmoset, and mouse, and internal branch 6; supplementary table S5 and supplementary fig. S2, Supplementary Material online). Qualitatively equivalent results were obtained when interactions among genes in the same chromosome arm were also removed from the analysis (PIN3; supplementary table S5, Supplementary Material online). Similar results are obtained using the species overlap method (supplementary table S5, Supplementary Material online). These results indicate that although the tendency of genes that duplicated in a given branch to interact with each other is in part the result of the enrichment of the network in self-interactions and interactions among paralogs (Ispolatov et al. 2005; Pereira-Leal et al. 2007; Alvarez-Ponce and McInerney 2011), these features cannot completely account for the observed trend.

Discussion

We used phylogenetic methods to accurately determine the branches of the primate phylogeny at which each gene family duplicated and investigated the relationship between a gene's pattern of duplication and the position of its encoded product in the primate PIN. We addressed this dependency from three perspectives. First, we evaluated the relationship between network centrality of a protein and the duplicability of the encoding gene in the different branches of the studied phylogeny. Second, we tested whether interacting proteins manifest topologically similar phylogenetic trees, in particular, when we look beyond the analysis of 1:1 orthologs. Finally, we considered whether genes encoding interacting proteins tend to duplicate at the same branches of the phylogeny. In all three cases, we found new significant results, with some patterns being more complex than previously thought.

The Dynamic Relationship between Centrality and Duplicability

In *E. coli*, yeast, and fly, genes occupying central positions tend to remain singleton, whereas those acting at the periphery of the network can more often retain duplicated copies (Hughes and Friedman 2005; Prachumwat and Li 2006; Makino et al. 2009). This has been attributed to the deleterious effects of altering the dosage balance of protein–protein interactions (Veitia 2002; Papp et al. 2003). In contrast with the pattern observed in the aforementioned organisms, duplicated genes tend to be more central in the human interactome (Liang and Li 2007; D'Antonio and Ciccarelli 2011), indicating that the relationship between duplicability and centrality has undergone modification during animal evolution. The pattern observed in the human interactome has been attributed to the possibility that the involvement of a gene in a higher number of interactions would facilitate the functional diversification of paralogs, for example, through tissue specialization or that highly connected proteins would be required in higher dosages (Liang and Li 2007).

Consistent with previous observations in the human genome (Liang and Li 2007; D'Antonio and Ciccarelli 2011), we found that primate genes that duplicated in any branch of the species tree tend to be more central than singleton genes (supplementary table S4, Supplementary Material online, and fig. 2). According to the dosage balance hypothesis, duplication of a gene would be deleterious unless its interacting partners underwent co-duplication soon after or before (Papp et al. 2003). An extreme example of co-duplication is whole genome duplication (WGD), which maintains the relative dosage of all balanced sets (Veitia 2004, 2005). Therefore, the high content of ohnologs (i.e., genes resulting from the two WGD events that occurred in early vertebrate evolution; Wolfe 2000) in mammalian genomes (Nakatani et al. 2007; Makino and McLysaght 2010) might potentially provide an explanation for the lack of a negative association between duplicability and centrality in mammals. Indeed, when the relationship between duplicability and centrality was analyzed separately for genes duplicated in each branch of the phylogeny, we found that genes that duplicated in the ancestral branch to all studied species (branch 6; fig. 1), which include ohnologs, tend to be more central than genes that did not duplicate at that branch (supplementary table S4, Supplementary Material online, and fig. 2). However, we also observed the same pattern in most of the other branches of the phylogeny (all of them post-WGD): Duplicated genes encode more central proteins than nonduplicated genes (supplementary table S4, Supplementary Material online, and fig. 2). This indicates that the preferential duplication of central genes is an ongoing process that can be observed in relatively recent branches (e.g., the macaque branch, which encompasses the last ~23.5 My; Benton et al. 2009; figs. 1 and 2) and not solely the result of WGD.

Unexpectedly, the opposite relationship between duplicability and centrality is observed in the external branch leading to humans and in the internal branches subtending the human/chimpanzee (branch 1) and the human/chimpanzee/gorilla (branch 2) clades, with statistically significant differences for the human branch and branch 1 (fig. 1). That is to say, in contrast to the overall trend, genes that duplicated in these lineages tend to occupy more peripheral positions in the network than nonduplicated genes (supplementary table S4, Supplementary Material online, and fig. 2), resembling the

Downloaded from http://mbe.oxfordjournals.org/ at National University of Ireland, Maynooth on March 7, 2013

pattern observed in *E. coli*, yeast, and fly (Hughes and Friedman 2005; Prachumwat and Li 2006; Makino et al. 2009). Therefore, the relationship between duplicability and centrality seems to have undergone a reversal during the evolution of great apes, revealing that this relationship is highly dynamic.

D'Antonio and Ciccarelli (2011) recently showed that the particular relationship between duplicability and centrality observed in humans is the result of the high content of the human genome in genes that arose late in evolution. Human genes of ancient (premetazoan) origin exhibit the same pattern as observed in *E. coli*, yeast, and fly (duplicated genes are less central), whereas human genes of more recent origin (those that originated within the metazoans) exhibit the opposite trend (duplicated genes tend to be more central). This contrasting pattern observed among ancient and new human genes could potentially provide an explanation for the different relationship between centrality and duplicability that we observe in the different branches of the phylogeny if duplications in the human branch and internal branch 1 involved preferentially ancient genes. However, we found that the proportion of ancient genes among genes that duplicated in these branches (9.49–17.32%) is generally lower than for genes that duplicated in the other branches of the phylogeny (table 1) (qualitatively similar results are obtained when the analysis is restricted to genes represented in the human interactome; table 1), indicating that the different age of genes that duplicated in the different branches of the phylogeny is not the factor responsible for the heterogeneity in the relationship between duplicability and centrality observed here.

Genes Encoding Interacting Proteins Exhibit Correlated Tree Topologies and Duplication Histories

To gain further insight into the relationship between the structure of the primate PIN and the duplicability of its components, we then considered whether genes encoding interacting proteins exhibit tree topologies that are more similar than expected from a random pair of proteins. We found that interacting genes exhibit phylogenetic trees with a higher similarity than expected from a random PIN (fig. 3). This tendency is not the result of the enrichment of the human interactome in self-interactions and interactions among paralogs (Ispolatov et al. 2005; Pereira-Leal et al. 2007; Alvarez-Ponce and McInerney 2011) or the clustering in the genome of genes encoding interacting proteins (Lee and Sonnhammer 2003; Makino and McLysaght 2008). Our observations contrast with those by Kelly and Stumpf (2010). They found only negligible evidence for pairs of yeast interacting proteins presenting phylogenetic trees topologically more similar than random pairs of proteins. At least three possible reasons might account for the different results obtained in both studies. First, they analyzed the yeast interactome, whereas we focused on primates; therefore, it might be possible that both interactomes would exhibit a different trend. Second, the datasets used by Kelly and Stumpf (2,528–5,109 proteins and 5,728–21,283 interactions) were remarkably smaller than the one used here, which could have limited statistical power in their analyses. Finally, they inferred phylogenetic trees from 1:1 orthologous sets, whereas we used entire gene families. Although computationally convenient, using 1:1 orthologous sets removes the effect of duplication and loss events in the tree topologies. Therefore, the different results obtained in the analysis by Kelly and Stumpf (2010) and our analysis may also potentially be the result, at least partially, of interacting genes exhibiting similar patterns of duplication and/or loss.

We found that the number of interactions between genes that underwent duplication at any branch of the phylogeny is higher than expected from a random network. This observation indicates that duplicated human genes tend to interact with each other in the PIN (supplementary fig. S2, Supplementary Material online), lending support to the hypothesis that duplication of a gene may increase the likelihood of duplication of its interacting partners (Fryxell 1996). This trend holds true when genes that duplicated in each particular branch of the phylogeny are analyzed separately (supplementary table S5 and supplementary fig. S2, Supplementary Material online). Although the significance vanishes for most of the branches when self-interactions and interactions among paralogs are removed, the trend remains significant for four of the branches (the external branches leading to gorilla, marmoset, and mouse and internal branch 6; supplementary table S5 and supplementary fig. S2, Supplementary Material online). Interestingly, these branches include the three longest branches in the phylogeny (the external branches leading to mouse and marmoset and internal branch 6; see fig. 1), suggesting that perhaps the lack of significance in the remaining branches may be the result of reduced statistical power in the shorter branches. Alternatively, the absence of significance in these branches might be a consequence of the reduced efficiency of selective mechanisms favoring the co-duplication of interacting genes in the same branches of the phylogeny. Indeed, we might expect that the selective advantage of duplicating the interacting partner of a protein would be often small. Furthermore, changes in gene dosage can be compensated by mechanisms different from complementary gene duplication, such as changes in expression levels, or may even be accommodated by stochastic variation in levels of protein expression. Therefore, the tendency of interacting genes to co-duplicate in the same branches of the species tree may be observed in organisms only in which natural selection is highly efficient. Primates have a lower effective population size than rodents, which is thought to involve a reduced efficiency of natural selection (Ohta 1973; Lynch 2007); therefore, evolutionary forces promoting the co-duplication of genes encoding interacting proteins may be less efficient in primates.

Conclusion

Taken together, our analyses indicate that the position of proteins in the primate PIN has an effect on the patterns of duplication of their encoding genes, indicating that the network imposes constraints on the fate of genes encoding its components. First, gene duplicability depends on the centrality of the encoded products in the network, although,

Downloaded from http://mbe.oxfordjournals.org/ at National University of Ireland, Maynooth on March 7, 2013

interestingly, the relationship between centrality and duplicability has varied during primate evolution. Second, interacting proteins exhibit similar duplication histories, tending to co-duplicate in the same branches of the phylogeny. Furthermore, we observed that interacting genes exhibit topologically similar phylogenetic trees, possibly owing to these correlated duplication histories.

Although separate analysis of individual genomes represents a valuable tool to provide a first glance at the patterns of gene duplication, this approach only allows a binary classification of genes as singleton or duplicated, thus providing only an aggregate overview. A more comprehensive characterization of duplication events can be gained by including multiple genomes in the analysis. This comparative approach allows, for instance, assigning duplication events to particular branches in the species tree. When applied to a relatively small selection of mammals, this approach allowed us to observe a dynamical relationship between the structure of the PIN and the patterns of duplication of genes encoding its components. Future work is warranted to understand how the structure of the PIN has influenced gene duplicability in other lineages. In particular, it will be interesting to see the relationship between duplicability and network position in organisms with effective population sizes that are larger than those for mammals. In addition, we note that a limitation of our analysis is that currently available interactomic data are highly incomplete and subject to a high rate of false negatives (Bader et al. 2004; Deeds et al. 2006). The future availability of more complete and accurate interactomes will allow a deeper understanding of the relationship between duplicability and centrality.

Supplementary Material

Supplementary tables S1–S5 and figs. S1 and S2 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

The authors thank Toni Gabaldón for helpful comments on the manuscript. The authors would like to thank the Irish Centre for High End Computing and the NUI Maynooth HPC facility. This work was supported in part by a Science Foundation Ireland grant 09/RFP/EOB2510 to J.O.M. and an Irish Research Council for Science Engineering and Technology grant to A.D.

References

Abascal F, Zardoya R, Telford MJ. 2010. TranslatorX: multiple alignment of nucleotide sequences guided by amino acid translations. *Nucleic Acids Res.* 38:W7.

Al-Shahrour F, Díaz-Uriarte R, Dopazo J. 2004. FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes. *Bioinformatics* 20:578.

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

Alvarez-Ponce D, Aguadé M, Rozas J. 2009. Network-level molecular evolutionary analysis of the insulin/TOR signal transduction pathway across 12 *Drosophila* genomes. *Genome Res.* 19:234–242.

Alvarez-Ponce D, Aguadé M, Rozas J. 2011. Comparative genomics of the vertebrate insulin/TOR signal transduction pathway: a network-level analysis of selective pressures. *Genome Biol Evol.* 3:87–101.

Alvarez-Ponce D, Guirao-Rico S, Orengo DJ, Segarra C, Rozas J, Aguade M. 2012. Molecular population genetics of the insulin/TOR signal transduction pathway: a network-level analysis in *Drosophila melanogaster*. *Mol Biol Evol.* 29:123–132.

Alvarez-Ponce D, McInerney JO. 2011. The human genome retains relics of its prokaryotic ancestry: human genes of archaebacterial and eubacterial origin exhibit remarkable differences. *Genome Biol Evol.* 3:782–790.

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT. 2000. Gene ontology: tool for the unification of biology. *Nat Genet.* 25:25.

Bader JS, Chaudhuri A, Rothberg JM, Chant J. 2004. Gaining confidence in high-throughput protein interaction networks. *Nat Biotechnol.* 22:78–85.

Benton MJ, Donoghue PCJ, Asher RJ. 2009. Calibrating and constraining molecular clocks. In: Hedges SB, Kumar S, editors. The timetree of life. Oxford: Oxford University Press. p. 35–86.

Borgatti SP. 2005. Centrality and network flow. *Soc Network.* 27:55–71.

Castillo-Davis CI, Hartl DL. 2002. Genome evolution and developmental constraint in *Caenorhabditis elegans*. *Mol Biol Evol.* 19:728–735.

Codoñer FM, Fares MA. 2008. Why should we care about molecular coevolution? *Evol Bioinform Online.* 4:29–38.

Cork JM, Purugganan MD. 2004. The evolution of molecular genetic pathways and networks. *Bioessays* 26:479–484.

D'Antonio M, Ciccarelli FD. 2011. Modification of gene duplicability during the evolution of protein interaction network. *PLoS Comput Biol.* 7:e1002029.

Deeds EJ, Ashenberg O, Shakhnovich EI. 2006. A simple physical model for scaling in protein-protein interaction networks. *Proc Natl Acad Sci U S A.* 103:311–316.

Demuth DP, Hahn MW. 2009. The life and death of gene families. *Bioessays* 31:29–39.

Eanes WF. 2011. Molecular population genetics and selection in the glycolytic pathway. *J Exp Biol.* 214:165–171.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 32:1792.

Fares MA, Ruiz-Gonzalez MX, Labrador JP. 2011. Protein coadaptation and the design of novel approaches to identify protein-protein interactions. *IUBMB Life.* 63:264–271.

Flicek P, Amode MR, Barrell D, et al. (52 co-authors). 2011. Ensembl 2011. *Nucleic Acids Res.* 39:D800–D806.

Fortna A, Kim Y, MacLaren E, Marshall K, Hahn G, Meltesen L, Brenton M, Hink R, Burgers S, Hernandez-Boussard T. 2004. Lineage-specific gene duplication and loss in human and great ape evolution. *PLoS Biol.* 2:e207.

Fraser HB, Hirsh AE, Steinmetz LM, Scharfe C, Feldman MW. 2002. Evolutionary rate in the protein interaction network. *Science* 296:750–752.

Fryxell KJ. 1996. The coevolution of gene family trees. *Trends Genet.* 12:364–369.

Gabaldón T. 2008. Large-scale assignment of orthology: back to phylogenetics? *Genome Biol.* 9:235.

Goh CS, Bogan AA, Joachimiak M, Walther D, Cohen FE. 2000. Co-evolution of proteins with their interaction partners. *J Mol Biol.* 299:283–293.

Downloaded from http://mbe.oxfordjournals.org/ at National University of Ireland, Maynooth on March 7, 2013

Goodman M, Czelusniak J, Moore GW, Romero-Herrera A, Matsuda G. 1979. Fitting the gene lineage into its species lineage, a parsimony strategy illustrated by cladograms constructed from globin sequences. *Syst Biol.* 28:132–163.

Hahn MW. 2007. Bias in phylogenetic tree reconciliation methods: implications for vertebrate genome evolution. *Genome Biol.* 8:R141.

Hahn MW, Demuth JP, Han SG. 2007. Accelerated rate of gene gain and loss in primates. *Genetics* 177:1941.

Hahn MW, Kern AD. 2005. Comparative genomics of centrality and essentiality in three eukaryotic protein-interaction networks. *Mol Biol Evol.* 22:803–806.

He X, Zhang J. 2005. Gene complexity and gene duplicability. *Current Biol.* 15:1016–1021.

Huerta-Cepas J, Dopazo H, Dopazo J, Gabaldón T. 2007. The human phylome. *Genome Biol.* 8:R109.

Huerta-Cepas J, Dopazo J, Gabaldón T. 2010. ETE: a python environment for tree exploration. *BMC Bioinformatics* 11:24.

Hughes AL, Friedman R. 2005. Gene duplication and the properties of biological networks. *J Mol Evol.* 61:758–764.

Ispolatov I, Yuryev A, Mazo I, Maslov S. 2005. Binding properties and evolution of homodimers in protein-protein interaction networks. *Nucleic Acids Res.* 33:3629–3635.

Kelly W, Stumpf M. 2010. Trees on networks: resolving statistical patterns of phylogenetic similarities among interacting proteins. *BMC Bioinformatics.* 11:470.

Koretke KK, Lupas AN, Warren PV, Rosenberg M, Brown JR. 2000. Evolution of two-component signal transduction. *Mol Biol Evol.* 17:1956–1970.

Lahortiga I, De Keersmaecker K, Van Vlierberghe P, Graux C, Cauwelier B, Lambert F, Mentens N, Beverloo HB, Pieters R, Speleman F. 2007. Duplication of the MYB oncogene in T cell acute lymphoblastic leukemia. *Nat Genet.* 39:593–595.

Lee JM, Sonnhammer EL. 2003. Genomic gene clustering analysis of pathways in eukaryotes. *Genome Res.* 13:875–882.

Lemos B, Bettencourt BR, Meiklejohn CD, Hartl DL. 2005. Evolution of proteins and gene expression levels are coupled in *Drosophila* and are independently associated with mRNA abundance, protein length, and number of protein-protein interactions. *Mol Biol Evol.* 22:1345–1354.

Liang H, Li WH. 2007. Gene essentiality, gene duplicability and protein connectivity in human and mouse. *Trends Genet.* 23:375–378.

Lipinski KJ, Farslow JC, Fitzpatrick KA, Lynch M, Katju V, Bergthorsson U. 2011. High spontaneous rate of gene duplication in *Caenorhabditis elegans*. *Current Biol.* 21:306–310.

Lovell SC, Robertson DL. 2010. An integrated view of molecular coevolution in protein-protein interactions. *Mol Biol Evol.* 27:2567–2575.

Luisi P, Alvarez-Ponce D, Dall'olio GM, Sikora M, Bertranpetit J, Laayouni H. 2012. Network-level and population genetics analysis of the insulin/TOR signal transduction pathway across human populations. *Mol Biol Evol.* 29:1379–1392.

Lynch M. 2007. The origins of genome architecture. Sunderland (MA): Sinauer Associates.

Makino T, Hokamp K, McLysaght A. 2009. The complex relationship of gene duplication and essentiality. *Trends Genet.* 25:152–155.

Makino T, McLysaght A. 2008. Interacting gene clusters and the evolution of the vertebrate immune system. *Mol Biol Evol.* 25:1855–1862.

Makino T, McLysaght A. 2010. Ohnologs in the human genome are dosage balanced and frequently associated with disease. *Proc Natl Acad Sci U S A.* 107:9270–9274.

Marcet-Houben M, Gabaldón T. 2011. TreeKO: a duplication-aware algorithm for the comparison of phylogenetic trees. *Nucleic Acids Res.* 39:e66.

Marland E, Prachumwat A, Maltsev N, Gu Z, Li WH. 2004. Higher gene duplicabilities for metabolic proteins than for nonmetabolic proteins in yeast and *E. coli*. *J Mol Evol.* 59:806–814.

Marques-Bonet T, Kidd JM, Ventura M, Graves TA, Cheng Z, Hillier LDW, Jiang Z, Baker C, Malfavon-Borja R, Fulton LA. 2009. A burst of segmental duplications in the african great ape ancestor. *Nature* 457:877.

Mason O, Verwoerd M. 2007. Graph theory and networks in biology. *IET Syst Biol.* 1:89–119.

Nakatani Y, Takeda H, Kohara Y, Morishita S. 2007. Reconstruction of the vertebrate ancestral genome reveals dynamic genome reorganization in early vertebrates. *Genome Res.* 17:1254–1265.

Niimura Y, Nei M. 2005. Comparative evolutionary analysis of olfactory receptor gene clusters between humans and mice. *Gene* 346:13–21.

Niimura Y, Nei M. 2007. Extensive gains and losses of olfactory receptor genes in mammalian evolution. *PLoS One.* 2:e708.

Ohno S. 1970. Evolution by gene duplication. Berlin: Springer-Verlag.

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

Page RDM. 1994. Maps between trees and cladistic analysis of historical associations among genes, organisms, and areas. *Syst Biol.* 43:58.

Papp B, Pal C, Hurst LD. 2003. Dosage sensitivity and the evolution of gene families in yeast. *Nature* 424:194–197.

Pazos F, Juan D, Izarzugaza JM, Leon E, Valencia A. 2008. Prediction of protein interaction based on similarity of phylogenetic trees. *Methods Mol Biol.* 484:523–535.

Pazos F, Valencia A. 2001. Similarity of phylogenetic trees as indicator of protein-protein interaction. *Protein Eng.* 14:609–614.

Pereira-Leal JB, Levy ED, Kamp C, Teichmann SA. 2007. Evolution of protein complexes by duplication of homomeric interactions. *Genome Biol.* 8:R51.

Perry GH, Dominy NJ, Claw KG, Lee AS, Fiegler H, Redon R, Werner J, Villanea FA, Mountain JL, Misra R. 2007. Diet and the evolution of human amylase gene copy number variation. *Nat Genet.* 39: 1256–1260.

Prachumwat A, Li WH. 2006. Protein function, connectivity, and duplicability in yeast. *Mol Biol Evol.* 23:30–39.

Pruitt KD, Tatusova T, Maglott DR. 2007. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.* 35:D61–D65.

Rasmussen MD, Kellis M. 2011. A bayesian approach for fast and accurate gene tree reconstruction. *Mol Biol Evol.* 28:273.

Slamon DJ, Clark GM, Wong SG, Levin WJ, Ullrich A, McGuire WL. 1987. Human breast cancer: correlation of relapse and survival with amplification of the HER-2/neu oncogene. *Science* 235:177.

Stark C, Breitkreutz BJ, Chatr-Aryamontri A, et al. (15 co-authors). 2011. The BioGRID interaction database: 2011 update. *Nucleic Acids Res.* 39:D698–D704.

Veitia RA. 2002. Exploring the etiology of haploinsufficiency. *Bioessays* 24:175–184.

Veitia RA. 2004. Gene dosage balance in cellular pathways: implications for dominance and gene duplicability. *Genetics* 168: 569–574.

Veitia RA. 2005. Paralogs in polyploids: one for all and all for one? *Plant Cell.* 17:4–11.

Downloaded from http://mbe.oxfordjournals.org/ at National University of Ireland, Maynooth on March 7, 2013

Wolfe K. 2000. Robustness—it's not where you think it is. *Nat Genet.* 25: 3–4.

Yang J, Li WH. 2004. Developmental constraint on gene duplicability in fruit flies and nematodes. *Gene* 340:237–240.

Yang J, Lusk R, Li WH. 2003. Organismal complexity, protein complexity, and gene duplicability. *Proc Natl Acad Sci U S A.* 100:15661.

Yokoyama S. 2002. Molecular evolution of color vision in vertebrates. *Gene.* 300:69–78.

Zera AJ. 2011. Microevolution of intermediary metabolism: evolutionary genetics meets metabolic biochemistry. *J Exp Biol.* 214:179–190.

Zhang J. 2003. Evolution by gene duplication: an update. *Tr Ecol Evol.* 18: 292–298.

Downloaded from http://mbe.oxfordjournals.org/ at National University of Ireland, Maynooth on March 7, 2013