The Molecular Phylogeny of Placental Mammals and its application to uncovering signatures of Molecular Adaptation

Claire C. Morgan B. Sc. (Hons) Biotechnology

A thesis presented to Dublin City University for the Degree of

Doctor of Philosophy

Supervisor: Dr. Mary J. O'Connell

School of Biotechnology

Dublin City University

External Supervisor: Dr. Simon Whelan

University of Manchester

January 2013

Declaration

‘I hereby certify that this material, which I now submit for assessment on the programme of study leading to the award of Doctor of Philosophy is entirely my own work, that I have exercised reasonable care to ensure that the work is original, and does not to the best of my knowledge breach any law of copyright, and has not been taken from the work of others save and to the extent that such work has been cited and acknowledged within the text of my work.’

Signed: _____

ID Number: 54308646

Date: _____

Abstract

Considerable conflict remains in the literature as to the position of the root of placental mammals, and the placement of several intra-ordinal groups. Debate continues over the use of DNA or amino acids datasets and over the use of Supertree or Supermatrix approaches. Known phenomena exist within mammal data that complicate the reconstruction of phylogeny. These include (but are not limited to), variation in longevity, body size, metabolic rates, and germ-line generation time that result in variation in mutation rates and composition biases. Previous attempts to resolve the placental mammal phylogeny have used homogeneous evolutionary models that cannot capture and adequately describe these features across the species sampled. In this thesis I explore the properties of different datasets and data types and their suitability to the resolution of the mammal phylogeny at different depths: (i) the position of the root of the placental mammals, and (ii), the intraordinal placements within the Laurasiatheria. The datasets tested were (i) mitochondrial and nuclear data types, (ii) previously published datasets for mammals, and (iii), datasets I assembled specifically for analyses at different phylogenetic depths. I propose and apply the use of heterogeneous models to resolve the position of the root of the placental mammal phylogeny to these datasets.

Reconstruction of a robust mammal phylogeny provides us with an essential framework for understanding the molecular underpinnings of adaptation to environment. The placental mammals display a huge variations in life traits such longevity, body size and DNA repair efficiency, since they emerged ~100 million years ago. With this robust phylogeny, I set out to determine the level of adaptive and non-adaptive processes acting on a set of mammal genes that are linked with longevity and cancer.

The results of these analyses yield important insights into data and model suitability, and provide strong evidence for a single hypothesis for the rooting of placental mammals. These results also show that Laurasiatheria intra-ordinal placements are not fully resolved and additional sampling from this diverse clade is required. Using this resolved phylogeny, specific molecular adaptations and non-adaptive mechanisms were identified in the mammalia for a set of telomere-associated genes.

Success

"Success is not final, failure is not fatal: it is the courage to continue that counts."

Winston S. Churchill

For My Family and Friends

Acknowledgements

It is difficult to know where to start as I have been so blessed with the support from all my friends and family over the past number of years.

To my husband John, who has been a rock of emotional support, for moving countries so that I could carry out my Ph.D., for maintaining calm during my bouts of irrational behaviour, for thinking of clever ways to help me work around my many computational problems and for being there every step of the way, Thank you.

I honestly don't think I can adequately word how thankful I am to my supervisor Dr. Mary O'Connell. She has been a keystone in the development of my scientific and critical thinking skills, providing me with every opportunity and pushing me to my full potential. Mary has been such an incredible influence in my life and I am so thankful to have had the opportunity to work with her here in Dublin City University (DCU).

To my fellow BME lab folk; Dr. Noeleen Loughran, Thomas Walsh, Mark Lynch, Andrew Webb, Kate Lee and Ann Mc Cartney, thank you for all the chats, the support, the laughs, shared frustration over queuing systems and for team effort in working through problems and supervising labs.

A sincere thanks to the following collaborators; Prof. James McInerney (NUIM), Dr. Davide Pisani (Brist.), Dr. Peter Foster (NHM), Dr. Chris Creevey (Teagasc), Prof. Heather Ruskin (DCU), Dr. Kieran Meade (Teagasc), Dr. Simon Berrow (IWDG), Conor Ryan (IWDG), Dr. Sinead Loughran (Dundalk IT), Dr. Padraig Doolan and Dr. Sinead Aherne.

A special thanks to Eoin McLoughlin for all his assistance with python programming.

Finally, a big thanks to all of you wonderful friends who have helped me maintain "a life" outside of the Ph.D.

Abbreviations

A	Adenine
aa	Amino Acids
ABL1	Abelson murine leukemia viral oncogene homolog 1
ABL2	Abelson murine leukemia viral oncogene homolog 2
Alp	Alpaca
ANKHD1	Ankyrin repeat and KH domain-containing protein 1
ANKHD1-EIF4EBP3	Ankyrin repeat and KH domain-containing protein 1 eukaryotic translation initiation factor 4E-binding protein 3
ANKRD17	Ankyrin repeat domain-containing protein 17
Arg	Arginine
Arm	Armadillo
ASRV	Among site rate variation
ATM	Ataxia telangiectasia mutated
ATP6	ATP synthase subunit a
ATP8	ATP synthase protein 8
ATRX	Alpha thalassemia/mental retardation syndrome X-linked
BEB	Bayes Empirical Bayes
BF	Bayes Factors
BGI	Beijing Genomics Institute
BI	Bayesian Inference
BIC	Bayesian information criterion
BLAST	Basic Local Alignment Search Tool
Blossum	Blocks substitution matrix
BLM	Bloom syndrome protein
bp	Base pairs
BRCA1	Breast cancer type 1 susceptibility protein
BRCA2	Breast cancer type 2 susceptibility protein
BRIP1	BRCA1 interacting protein C-terminal helicase 1
Bus	Bushbaby
BYA	Billions of years ago

C	Cytosine
CAT	Empirical profile mixture model
CAT-BP	Empirical profile mixture model with break points
CBX1	Chromobox homolog 1
CBX3	Chromobox homolog 3
CBX5	Chromobox homolog 5
CED	Certain Evolutionary Distance
Cfa	Dog
Chi	Chimpanzee
Chk	Chicken
Cho	Sloth
CO1	Cytochrome c oxidase subunit 1
CO2	Cytochrome c oxidase subunit 2
CO3	Cytochrome c oxidase subunit 3
CPREV	Chloroplast mitochondrial model
Cts	Guineapig
CYTB	Cytochrome b
Dayhoff	Dayhoff model
df	Degrees of freedom
DKC1	Dyskeratosis congenita 1, dyskerin
Dn	Non-synonymous substitution per non-synonymous site
Dn/Ds	Rate of non-synonymous substitution per non-synonymous site to Synonymous substitution per synonymous site.
DNA	Deoxyribonucleic acid
Dog	Dog
Dol	Dolphin
Ds	Synonymous substitution per synonymous site
Dvi	Opossum
Eca	Horse
Ele	Elephant
ERCC1	Excision repair cross-complementing rodent repair deficiency, complementation group 1 (includes overlapping antisense

	sequence)
ERCC4	Excision repair cross-complementing rodent repair deficiency, complementation group 4
Ete	Tenrec
EXO1	Exonuclease 1
Exp	Expected
+F	Amino acid frequencies
f	Fixed parameters
F81	Felsenstein (F81) model
FANCA	Fanconi anemia, complementation group A
FANCB	Fanconi anemia, complementation group B
FANCC	Fanconi anemia, complementation group C
FANCD2	Fanconi anemia, complementation group D2
FANCE	Fanconi anemia, complementation group E
FANCF	Fanconi anemia, complementation group F
FANCI	Fanconi anemia, complementation group I
FANCL	Fanconi anemia, complementation group L
Fca	Cat
FDR	False discovery rate
G	Guanine
gBGC	GC-biased gene conversion
GC	Content of guanine and cytosine
GC3	Content of 3 codon position of guanine and cytosine
Gip	Giant Panda
Gly	Glycine
GO	Gene Ontology
Gor	Gorilla
GTR	General Time Reversible
Gui	Guinea pig
H	Hypothesis
Hed	Hedgehog

HGT	Horizontol Gene Transfer
HIV	Human immunodeficiency virus
HKY	Hasegawa-Kishino-Yano
HMTases	Histone methytransferases
Hor	Horse
Hsa	Human
Hum	Human
ILS	Incomplete Lineage Sorting
+I	Proportion of invariable sites
JC	Jukes and Cantor
JTT	Jones and Taylor model
K2P	Kimura 2 parameter
K80	Kimura 2-parameter
K81	Kimura 3-parameter
K81uf	Kimura 3-parameter with unequal base frequencies
Kag	Kangaroo_rat
Kan	Kangaroo_rat
KPg	Cretaceous-Paleogene
KTR	Cretaceous Terrestrial Revolution
L	Likelihood
Laf	Elephant
LBA	Long Branch Attraction
LG	Le and Gascuel model
LM	Likelihood Mapping
ln	Natural logarithm
LRT	Likelihood Ratio Tests
Lys	Lysine
M	Model
Mac	Macaque

Mar	Marmoset
MC	Metropolos coupling
MCL	Markov Cluster Algorithm
MCMC	Markov Chain Monte Carlo
ME	Minimum Evolution
Mean % ID	Mean pairwise identity score
Meg	Megabat
Met	Methionine
Meu	Wallaby
Mic	Microbat
microRNA	Micro ribonucleic acid
ML	Maximum Likelihood
MLH1	MutL homolog 1, colon cancer, nonpolyposis type 2 (E. coli)
MLH3	MutL homolog 3 (E. coli)
Mma	Macaque
Mmu	Mouse
Mol	Mouse Lemur
Mou	Mouse
MR	Majority rule
MRCA	Most Recent Common Ancestor
MRE11A	MRE11 meiotic recombination 11 homolog A (S. cerevisiae)
mRNA	Messenger RNA
MSA	Multiple Sequence Alignment
MSH3	MutS homolog 3 (E. coli)
mtDNA	Mitochondrial DNA
mtGene	Mitochondrial Gene
mtMam	Mitochondrial mammal model
mtRev24	Mitochondrial vertebrate model
MUTYH	mutY homolog (E. coli)
MY	Million Years
MYA	Millions of years ago
n	Number
NBN	Nibrin

ND1	NADH-ubiquinone oxidoreductase chain 1
ND2	NADH-ubiquinone oxidoreductase chain 2
ND3	NADH-ubiquinone oxidoreductase chain 3
ND4	NADH-ubiquinone oxidoreductase chain 4
ND4L	NADH-ubiquinone oxidoreductase chain 4L
ND5	NADH-ubiquinone oxidoreductase chain 5
ND6	NADH-ubiquinone oxidoreductase chain 6
NDCH	Node-discrete composition heterogeneity
NDRH	Node-discrete rate matrix heterogeneity
N_e	Effective Population Size
NEB	Naïve Empirical Bayes
NJ	Neighbour Joining
Nmr	Naked Mole Rat
nucGene	Nuclear DNA
Obs	Observed
Ohy	Rabbit
Opo	Opossum
Ora	Orangutan
Orb	Orangutan
P_x	Population
P	Level of Significance (p-value)
P4	A python package for phylogenetics (software)
PALB2	Partner and localizer of BRCA2
PAML	Phylogenetic Analysis by Maximum Likelihood
Pan	Panda
PARP2	Poly (ADP-ribose) polymerase 2
PAUP	Phylogenetic Analysis Using Parsimony
PAUP	Phylogenetic analysi
Pca	Hyrax
PCNA	Proliferating cell nuclear antigen
Pig	Pig
Pik	Pika

Pla	Platypus
Ply	Platypus
PMS1	PMS1 postmeiotic segregation increased 1 (S. cerevisiae)
PMS2	PMS2 postmeiotic segregation increased 2 (S. cerevisiae)
Pob	Polar bear
Poisson	Poisson model
PP	Posterior Probability
Pr	Probability
PROT	Protein
Rab	Rabbit
RAD50	RAD50 homolog (S. cerevisiae)
RAD51D	RAD51 homolog D (S. cerevisiae)
Rat	Rat
RB1	Retinoblastoma 1
RBL1	Retinoblastoma-like 1 (p107)
RBL2	Retinoblastoma-like 2 (p130)
RF	Robinson-Fould
RGE	Rare Genomic Event
Rno	Rat
rtREV	Reverse Transcriptase empirical amino acid substitution model
Sar	Shrew
Ser	Serine
SGO	Single gene orthologs
Shr	Shrew
SLR	Sitewise likelihood-ratio test
SLX4	SLX4 structure-specific endonuclease subunit homolog (S. cerevisiae)
SM	Supermatrix
Squ	Squirrel
Ssr	Pig
SUV39H1	Suppressor of variegation 3-9 homolog 1 (Drosophila)
SUV39H2	Suppressor of variegation 3-9 homolog 2 (Drosophila)

SYM	Symmetrical model
T	Thymine
TDG	Thymine-DNA glycosylase
TERF1	Telomeric repeat binding factor (NIMA-interacting) 1
TERF2	Telomeric repeat binding factor 2
TERT	Telomerase reverse transcriptase
TIM	Transition model
Timef	Equal-frequency Transition Model
TINF2	TERF1 (TRF1)-interacting nuclear factor 2
TNKS	tankyrase, TRF1-interacting ankyrin-related ADP-ribose polymerase
TNKS2	tankyrase, TRF1-interacting ankyrin-related ADP-ribose polymerase 2
Tre	Treeshrew
TREX1	Three prime repair exonuclease 1
TrN	Tamura-Nei
TrNef	Equal-frequency Tamura-Nei
Trp	Tryptophan
Tsp	Tarsier
Ttr	Dolphin
TVM	Transversion model
TVMef	Equal-frequency Transversion
U_f	Probability of fixation of deleterious alleles
UPGMA	Unweighted Pair Group Method with Arithmetic Mean
v	Free parameters
Val	Valine
VT	Variable time matrix
WAG	Whelan and Goldman
WRN	Werner syndrome, RecQ helicase-like

X	Aligned sequence data
$\chi 2$	Chi-squared
XRCC3	X-ray repair complementing defective repair in Chinese hamster cells 3
XRCC5	X-ray repair complementing defective repair in Chinese hamster cells 5 (double-strand-break rejoining)
XRCC6	X-ray repair complementing defective repair in Chinese hamster cells 6
Zeb	Zebra finch
Δl	Difference between lnL scores of two models
$+\Gamma$	Gamma Distribution
τ	Topology
υ	Branch lengths
θ	Model parameters
ω	Rate of non-synonymous substitution per non-synonymous site to Synonymous substitution per synonymous site.

List of Figures

List of Tables

Chapter 1

1 Introduction

1.1 Natural Selection and Molecular Evolution

1.1.1 Evolutionary Theory

Charles Darwin laid the foundation for the study of evolutionary biology in his book "On the origin of species by means of natural selection, or the preservation of favoured races in the struggle for life" (Darwin 1859). He proposed a theory that could explain evolutionary change based on two major observations: (i) species change over time, and (ii), the process of natural selection dictates which variations are retained (Darwin 1859). This theory, coupled with the understanding of genetic inheritance that followed, revolutionized a modern way of thinking. The modern synthesis for evolutionary theory was based on the interaction between Natural Selection and Genetic Drift. The Hardy-Weinberg equation (Eqn. 1) describes an idealized state where allele frequencies do not change across generations. The frequencies of alleles are denoted by p and q, where p^2 or q^2 refer to the homozygous alleles and 2pq are the heterozygous alleles. If the population is in equilibrium then the sum of the frequencies of alleles should equal 1.

Eqn. 1 The Hardy-Weinberg equation

$$p^2 + 2pq + q^2 = 1$$

For a population to be at Hardy-Weinberg equilibrium the population must have: (i) random mating, (ii) an infinite population size, (iii) no gene flow between populations, (iv) no mutations, and (v), natural selection should not confer a survival advantage. Deviations from this equation provide evidence for natural selection or genetic drift. Natural populations do not adhere to the Hardy-Weinberg equation, as random genetic drift and directional natural selection alter allele frequencies.

1.1.2 Random Genetic Drift

According to the neutral theory of evolution, the majority of molecular evolutionary changes are caused by random genetic drift (Kimura 1968). This theory was expanded to "the nearly neutral theory of molecular evolution" which accounts for slightly advantageous or deleterious mutations that can become fixed in a population through random genetic drift (Ohta 1973, Ohta and Gillespie 1996). In a haploid population

with an effective population of size *Ne*, the probability that a neutral mutation will become fixed is P_x. If a new neutral mutation arises within a diploid population then the initial probability of the neutral mutation fixing within a population becomes halved (Equ. 2).

Eqn. 2 The probability of fixation of a neutral mutation within a population

$$P_x = \frac{1}{2N_e}$$

The number of new mutations per generation is $2N_e\mu$, therefore the fixation rate of a novel neutral mutation is multiplied by the probability of fixation giving an overall fixation rate of:

Eqn. 3 The overall fixation rate of novel neutral mutations

$$2N_e\mu\left(\frac{1}{2N_e}\right) = \mu$$

Therefore, the probability of a neutral mutation fixing in a population is higher if N is small (Kimura 1968). The retention time for a neutral allele within a population is related to the mutation rate (μ), where high μ decreases the probability of retention. If a new mutation occurs and it is destined to be fixed within a population, then Kimura calculated that the average time for fixation to occur is $4N_e$ generations (Kimura 1980a). If a new mutation is either beneficial or deleterious to the population then both N_e and the strength of selection (s) have been shown to impact the probability of fixation of that allele (Kimura 1957), see Figure 1.1. N_e is an important factor impacting allele frequency. For example, in the case of population bottlenecks, where N_e has gone through a period of substantial reduction, genetic variance is decreased (Lynch 1986) and the adaptive potential of the species is limited (Willi 2006). In a scenario where N_e is finite, recombination is absent, slightly deleterious mutation rates are high, and purifying selection is too weak to remove all new deleterious mutations, extinction can result through "Muller's ratchet effect" (Muller 1963). When a new population arises from an initial population with low N_e, it is called the founder effect (Mayr 1942), and this also results in the fixation of deleterious mutations, such situations are exemplified

by the frequency occurrence of Tay–Sachs and Gaucher disease in the Ashkenazic Jewish populations (Slatkin 2004).

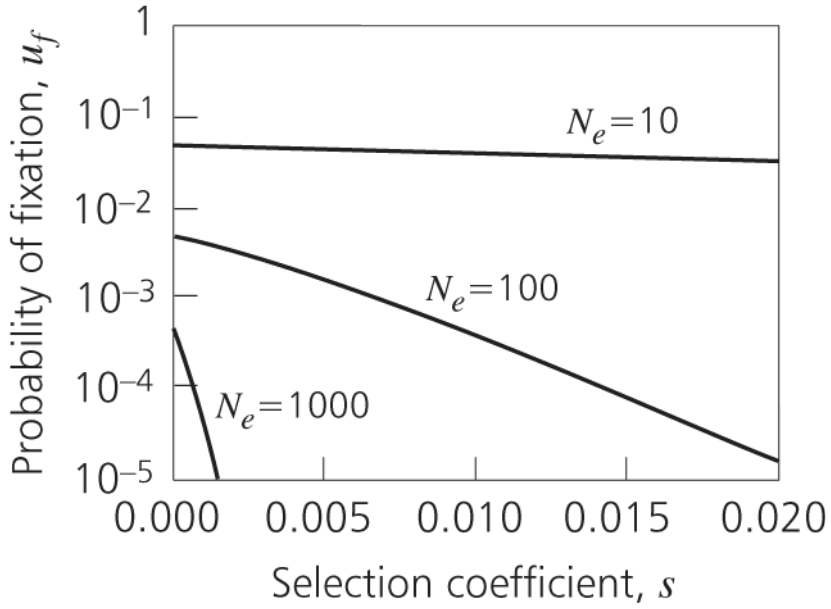


Figure 1.1 Fixation of deleterious mutations based on population size

The probability of fixation of deleterious alleles (U_f) is affected by the strength of selection against U_f (s) and the effective population size (N_e). The allele is initially present at frequency $1/(2N_e)$. Deleterious mutations will fix more readily in smaller populations (N_e = 10) compared to larger populations (N_e = 1000). Taken with permission from (Whitlock and Bürger 2004)

1.1.3 Natural Selection

Natural selection acts on phenotypes (observable traits) and the underlying genetic mutations. If the phenotypic trait is beneficial, and gives reproductive advantage to a species, then over time the mutation causing this phenotype will increase in frequency and improve the "fitness" of a population. In a similar fashion, natural selection works by ridding a population of deleterious mutations, this is also known as purifying selection. Natural selection has several levels at which it acts: individual, sexual and kin selection (Stearns 2008). Reproductive success is determined by the individual's ability to survive and produce fertile offspring, and it is dependent on success in finding a mate. Sexual selection acts on individuals to increase the likelihood of finding a mate even if it decreases the overall fitness of a species (Darwin 1871). For example, the ornate tail of the male peacock is attractive to the peahen, and while it increases the reproductive success of the bearer, it is also costly to generate and burdensome to carry leaving it more vulnerable to predators (Darwin 1871). Finally, kin selection a concept popularised by Hamilton (Hamilton 1964) has been used to describe the altruism and co-operative behaviour observed in communities of relatives. Kin selection is hypothesised to have been a major contributing factor in increased longevity in humans (Hawkes 2003). The "grand mother hypothesis" posses that increased longevity evolved in humans to assist in rearing the second-generation offspring ensuring that the grandparent genes will be successfully passed on through future generations (O'Connell et al. 1999). The concept of longevity will be revisited in section 1.1.6.

In a controversial publication entitled "Cancer and Evolution: Synthesis" Graham poses that cancer was a major selective pressure in the evolution of complex animals (Graham 1983). Graham describes this process as a cascade, whereby a mutagen initiates a mutation that triggers an oncogene resulting in transformation to a cancerous state. If this process occurs in a lineage before reproduction takes place the genes from that lineage are not passed on. Cancer selection is the process of natural selection selecting for individuals that have a defence system against cancer or a mechanism of postponing cancer until post-reproductive age (Graham 1992). He suggests that the constant pressure of cancer selection has lead to more precise mitosis that in turn lead to more accurate development (Graham 1992). In addition to the selective pressures mentioned in this section, there are many selective pressures acting upon a particular lineage at any

given time that ultimately are working on the underlying genetic material, e.g infection with new pathogen and diet regime (Fumagalli et al. 2011, Yang 1998).

1.1.4 The relationship between orthologs, paralogs and function

Homologs are defined as genes inherited from a common ancestor and are generally classified as (i) the product of a speciation event, i.e., ortholog, or (ii) the result of a gene duplication event, i.e., paralog (Fitch 1970), see Figure 1.2. More recently, paralogs have been further classified into "in-paralogs" based on a recent gene duplication within a species and "out-paralogs" that duplicated before a speciation event and can be in multiple species (Sonnhammer and Koonin 2002). Identification of orthologs and paralogs between the annotated genome from one species and an un-annotated genome of another can allow for inference of function as explored in Chapter 4 (Chen and Jeong 2000). Identification of homologous sequences is central to comparative genomic analyses and has been carried out using both sequence similarity and synteny analyses (Gabaldon et al. 2009, Wolf et al. 2001). Correct ortholog identification allows for reconstruction of a species phylogeny (Delsuc et al. 2005) and this will be explored in detail in Chapter 2.

The interpretation of the terms "ortholog" and "paralog" is continually refined and debated in the literature (Petsko 2001, Jensen 2001, Koonin 2001, Theissen 2002). The assumption that orthologs are more conserved in sequence and function than paralogs is a recent dispute (Dessimoz et al. 2012). The most recent debate that has ensued has hinged on the precise definition of an ortholog (Gabaldon et al. 2009) and whether orthologs should be defined based on: (i) sequence similarity (Gabaldon et al. 2009), (ii) conserved domain architecture (Forslund et al. 2011), (iii) position of introns (Henricson et al. 2010), (iv) function (Samsonova et al. 2002), or (v) protein structure (Peterson et al. 2009). An initial study by Nehrt et al. (2011) appeared to show how functional annotations are less similar among orthologs than they are among paralogs, thereby calling to question the relationship between orthology and function that is often assumed in molecular evolutionary studies (Nehrt et al. 2011). Subsequent studies showed how this analysis used computational and experimentally annotated Gene Ontology (GO) terms (du Plessis et al. 2011), and so contained biases resulting from algorithmic functional annotation that had impacted on the results and conclusions of the initial study (Altenhoff et al. 2012). The Nehrt et al. (2011) study also did not take

into account that the frequency of GO terms varies across species. On correcting for the frequencies of the GO terms by estimating them separately for each species, and on removal of computationally annotated GO terms, it was found that orthologs were more similar in function than paralogs (Altenhoff et al. 2012). The quest for harmonised definitions of orthologs is ongoing and a consortia of experts are presently working to standardise conventions (Dessimoz et al. 2012).

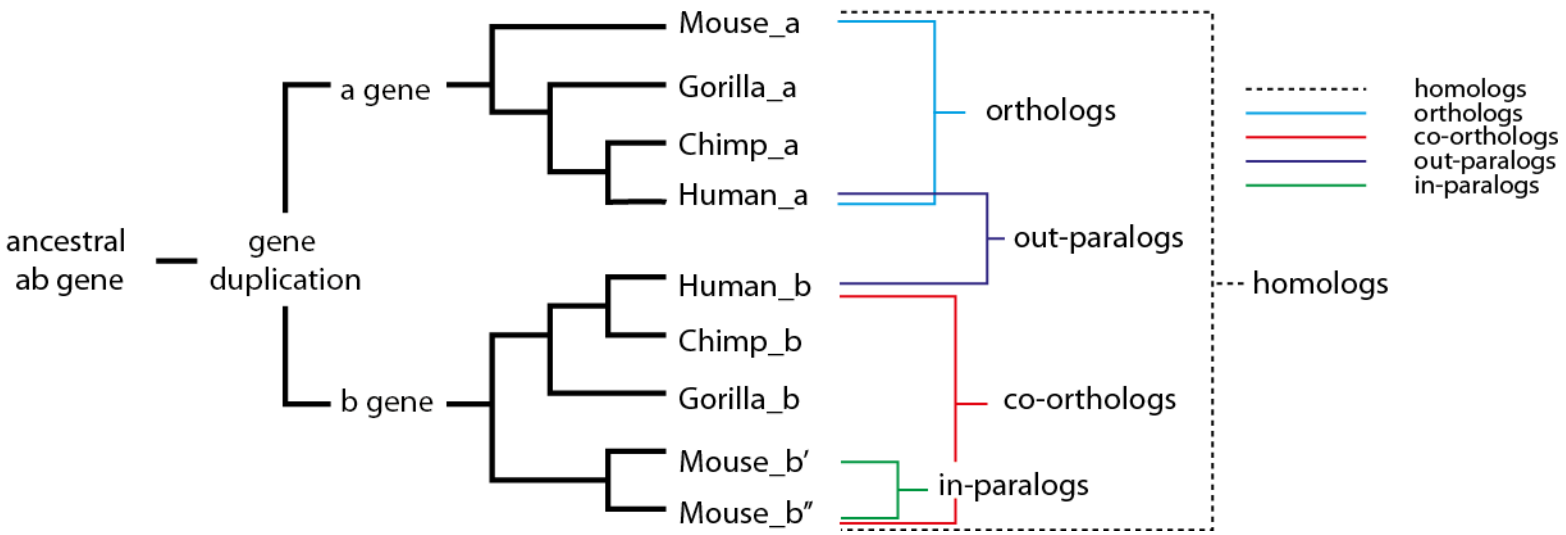


Figure 1.2 Gene phylogeny depicting the relationships amongst homologs.

Genes resulting from a speciation event are orthologs (blue), while genes resulting from a gene duplication event are paralogs. In-paralogs (green) are where species-specific gene duplication occurs and resulting orthologs cluster as co-orthologous (red). Out paralogs (purple) are not co-orthologs and are the results of a gene duplication event before speciation and are not necessarily in the same species.

Gene duplication is the mechanism by which paralogs arise and is believed to be the main engine for deriving new gene function. The mechanism of neofunctionalisation (duplication followed by the evolution of new function) is simply outlined as the duplication of a gene, the subsequent accumulation of mutations in the duplicate (while the other copy retains the original function) and the eventual evolution of a new function (Ohno 1970). Empirical studies have shown that most gene duplicates do not obtain novel functions and tend to either become subfunctionalised or pseudogenised (see review by Lynch and Force (2000)). Subfunctionalisation can result from conflicting selective pressure acting on the duplicates, with each duplicate performing part of the original function, therefore both duplicates are preserved (Force et al. 1999, Lynch and Force 2000). It has been shown that long term preservation of gene duplicates is only possible if sufficient levels of purifying selection are acting on these duplicated genes (Lynch and Katju 2004). Dosage effects can impact the fate of duplicates in two opposing ways: (i) dosage selection, when higher expression levels of a pair of duplicate genes are beneficial to the organism and result in the selection and fixation of both copies (Kondrashov and Koonin 2004), or (ii), dosage compensation, two copies of the gene are deleterious to the organism and one is subsequently pseudogenised (Heard and Disteche 2006). Finally, gene duplicates can undergo pseudogenisation and loose function (non-functionalisation) (Conant and Wolfe 2008). Estimates show that gene duplication and subsequent retention occurs in eukaryotic populations at rate of 1 gene per 100 million years (Lynch and Conery 2000), therefore birth and death of gene duplicates are common (Hughes and Nei 1989, Nei et al. 2000).

Gene duplication is not the only mechanism by which new functional genes emerge. For instance, recombination has been shown to be important in the evolution of new β-galactosidase (ebg) enzymatic function in *E. coli* K-12 genes (Hall and Zuzel 1980). Gene fusion events have also resulted in the evolution of novel chimeric genes such as human Ubiquitin-conjugating enzyme E2 variant 1 (KUA-UEV) (Thomson et al. 2000). In prokaryotic species such as bacteria it has been shown that pathogenic strains can emerge through the process of horizontal gene transfer (HGT) (Ochman 2001). Although HGT has been observed in plants (Bergthorsson et al. 2003) its prevalence in eukaryotes in general is not known. Other forms of introgression have been shown to be important in the evolution of new proteins (Bapteste et al. 2012) such as tissue-specific glue proteins in Drosophila (Aruna and Ranganath 2006). The process of rearranging portions of protein coding DNA by exon shuffling, as in the evolution of the

fucosyltransferase gene family (Javaud et al. 2003), has proven to be an important mechanism of new protein evolution (Patthy 1996, Long and Langley 1993). In human protein-coding genes, transposable elements are estimated to have generated 4% of new exons through transposition of sequences within the host genes (Nekrutenko and Li 2001). Retrotransposition produces new gene function through generation of gene duplicates in new genomic positions by reverse transcription, as evidenced by the sphinx gene in Drosophila (Wang et al. 2002) and the phosphoglycerate mutase gene in human and chimpanzee (Betran et al. 2002). Finally, *de novo* gene genesis from non-coding DNA sequences has been found in *Drosophila melanogaster* testis expressed genes (Levine et al. 2006), along with examples within the human lineage (Guerzoni and McLysaght 2011). There are other mechanisms by which genes acquire new functions and new genes originate and this is not an absolute list.

The mechanism of emergence of new gene function focused on specifically in Chapter 4 of this thesis is positive selection, described in more detail in section 1.1.5. Positive selection is the process by which advantageous mutations are retained and spread throughout a population - this has become synonymous with protein functional shift (Beall et al. 2010, Huang et al. 2012, Loughran et al. 2012).

1.1.5 Positive Selection and Functional Shift

There are three main types of Natural selection: (i) purifying selection, where a mutation is deemed deleterious and natural selection acts to decrease its frequency, (ii) positive selection (also referred to as adaptive evolution), where a mutation arises that results in a beneficial trait and natural selection acts to increase its frequency within the population, and finally (iii) neutral evolution where the mutation is neither deleterious nor beneficial and is carried through a population by random genetic drift at a frequency proportional to N_e. It is possible to detect the selective pressures acting on a species or a gene by calculating the frequency of alleles across the population; alternatively, the selective pressure on a particular protein coding sequence can be estimated by calculating the ratio of non-synonymous substitutions per non-synonymous site (Dn) to synonymous substitutions per synonymous site (D_S). This ratio (Dn/Ds) is denoted as ω throughout this thesis. The ω ratio can be interpreted as follows: (i) if $\omega > 1$, this is indicative of positive selection, (ii) if $\omega < 1$, this is indicative of purifying selection, and (iii), if $\omega = 1$, the protein is evolving under a neutral process, see Figure 1.3.

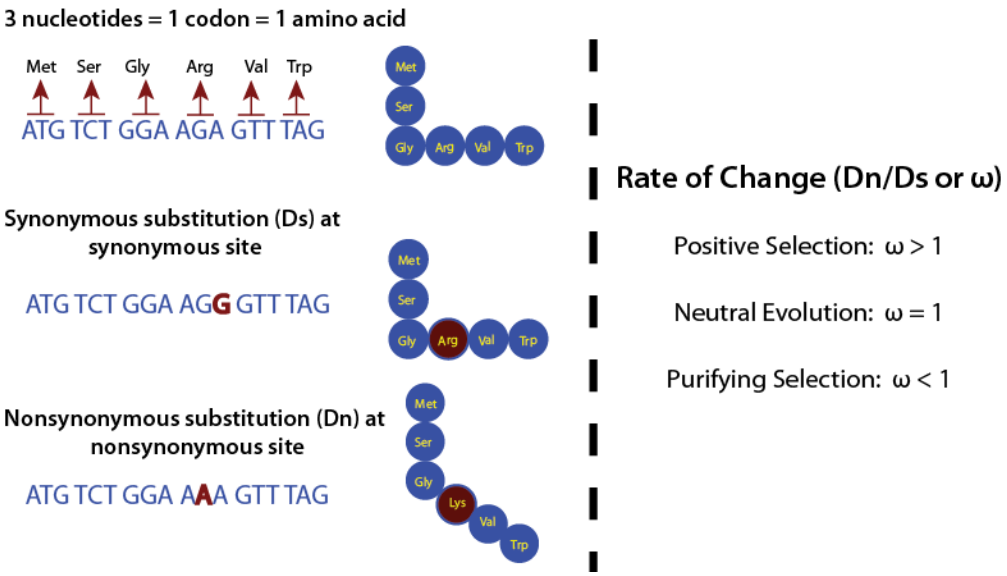


Figure 1.3 Diagramatic representation of the difference between ω ratios and protein functional shift on coding sequences

The nucleotides AGA code for arginine. Ds: a synonymous substitution at a synonymous site (AGA to AGG) does not alter protein, this is represented in the protein cartoon as no "kink" produced. Dn: a nonsynonymous substitution at a nonsynonymous site (AGA to AAA) results in a change from the amino acid serine to lysine, this is represented in the cartoon as a "kink" produced that may subsequently alter the proteins function.

At the nucleotide level, a non-synonymous substitutions (Dn) at a non-synonymous sites causes an amino acid substitution, and this biochemical change can alter the way in which the protein sequence is folded causing a structural change and subsequently a shift in the protein function. Synonymous substitutions at synonymous sites (Ds) in nucleotide sequences do not cause biochemical change at the amino acid level and therefore do not have a downstream effect on protein structure. Therefore when $\omega > 1$ is estimated from an aligned set of homologous sequences, it is indicative of positive selective pressure.

Despite high levels of positive selection reported in the literature, e.g. 45% of coding genes in Drosophila (Smith and Eyre-Walker 2002) and between 38% and 62% in mammals (Kosiol et al. 2008), it is generally held that the majority of codon positions in protein coding genes are evolving under purifying selection (Hughes 1999). The evolutionary constraint placed on proteins by their structure-function relationship means that much of sequence space is not viable for exploration (Peterson et al. 2009). Positive selection occurs at sites that confer an advantage to the population, and therefore only happen at specific points along a gene and not ubiquitously (Nielsen et al. 2005).

Although the link between positive selection and protein functional shift had been argued from a theoretical standpoint (Hughes 2007, Yang 1998), the empirical evidence was much slower to emerge. Initial studies of vision proteins showed that there was no clear connection between functional shift and positive selection (Yokoyama et al. 2008, Hughes and Piontkivska 2008). Austin Hughes has argued that elevated ω indicating positive selection does not necessarily correlate to protein functional shift and adaptive phenotypes are more likely to be caused by single amino-acid changes as opposed to repeated amino acid changes, deletion or silencing of genes or changes in gene expression (Hughes 2007). Empirical evidence by Yokoyama *et al.* (2008) demonstrated that upon reconstruction of ancestral vertebrate rhodopsin genes, repeated site mutagenesis on sites identified as positively selected did not correlate with protein functional shift (Yokoyama et al. 2008). More recent publications have shown a robust and biochemically validated link through the rational mutagenesis of various fungal, plant and mammal enzymes (Levasseur et al. 2006, Loughran et al. 2012, Moury and Simon 2011, Tennessen 2008). Following *in vitro* site-directed mutagenesis of sites identified as positively selected in the fungal lipase/feruloyl esterase A family, the results clearly demonstrated that certain amino acids under positive selection were

involved in a functional change in the enzyme (Levasseur et al. 2006). Site-directed mutagenesis experiments generating all possible combinations of positively selected sites in human myeloperoxidase, followed by detailed biochemical and biosynthetic pathway analyses confirmed the link between positive selection and functional shift to novel chlorination activity (Loughran et al. 2012). The empirical studies described here all employed codon-based maximum likelihood models of evolution, that are reliant on accurate phylogenetic reconstruction, to detect positive selection (this codon based approach is detailed in section 1.1.6.2). Yokoyama employed a neighbour joining method to reconstruct to phylogeny of the vertebrate rhodopsin genes (Yokoyama et al. 2008), while the two studies demonstrating links between positive selection and protein functional shift have employed more accurate maximum likelihood methods of phylogenetic reconstruction (Levasseur et al. 2006, Loughran et al. 2012). The success of maximum likelihood phylogenetic reconstruction methods compared to neighbour joining reconstruction methods have been detailed in section 1.2.1.

1.1.6 Methods for Detecting Positive Selection

The first methods employed to detect positive selection were distance based and assessed the rate of change among protein coding genes through a pairwise comparison of ω (Nei and Gojobori 1986, Ina 1995, Li et al. 1985). Improvements were made to this method by incorporating a phylogenetic tree - this allowed ancestral reconstruction at internal nodes, and the application of the distance-based method to calculate the ω between all sequences (Messier and Stewart 1997, Creevey and McInerney 2002, Yang 1998, Yang et al. 2000).

1.1.6.1 Distance Based Methods for Detecting Positive Selection

One of the first methods for estimating ω in protein coding sequences involved the classification of codons based on how often a nucleotide substitution would result in an amino acid replacement (Li et al. 1985). There are four separate classifications; non-degenerate sites are classed as non-synonymous sites, fourfold degenerate sites are classed as synonymous sites, twofold degenerate sites are classed as synonymous sites in the case of transitions, and non-synonymous sites in the case of transversions. The classification of sites is averaged across two sequences in a pairwise manner and ω is calculated. This earlier method is biased as it counts each twofold degenerate site as 1/3

synonymous and 2/3 non-synonymous, and this overestimates the synonymous counts due to transition mutations occurring more often than transversions and in addition most transitional mutations at twofold degenerate sites are synonymous (Li 1993). Therefore the Kimura two-parameter (Kimura 1980b) method was applied to correct for this bias (Li 1993). Even with the addition of this correction to the Li method (Li 1993) it was observed that ω estimations were less accurate when there was a high mutation rate, this produces underestimations of the non-synonymous substitution rate (Nei and Gojobori 1986). There are two major problems with pairwise detection methods, while it is possible to detect adaptive evolution, it was difficult to pinpoint the exact position within a sequence and it was not possible to assess lineage-specific positive selection, as there is no phylogeny implemented. In summary, pairwise methods of estimating adaptive evolutionary events lack precision and are not as effective as phylogeny-based approaches detailed in the next section. Sliding windows methods are not described as they have been shown to generate artifactual trends of synonymous and nonsynonymous rate variation, even if these values are completely constant and codons are evolving independently (Schmid and Yang 2008). A comparison of sliding windows based methods and the likelihood ratio test (LRT), showed that LRT's are far more rigorous in detecting positive selection and LRT methods have been detailed in the following section (Schmid and Yang 2008).

1.1.6.2 Phylogeny-Based Methods for Measuring Selective Pressure Variation

The first of the phylogeny based approaches for measuring selective pressure described here is the McDonald and Kreitman test which divides a phylogeny into two parts, "between-species" (inter-specific) and "within-species" (intra-specific). If a mutation occurs on a "between-species" branch, and is present in species that diverge from that branch, the mutation is considered fixed between those species. A mutation on a "within-species" branch is considered polymorphic within species (McDonald and Kreitman 1991). If positive selection is acting on polymorphic sites, then a mutation will fix quicker than if genetic drift is the sole driving force on the mutated site (McDonald and Kreitman 1991, Gillespie 1998). Therefore if adaptive evolution has occurred there will be more fixed mutations than polymorphic mutations observed between species (McDonald and Kreitman 1991). In this way it is possible to test if

signatures of adaptive evolution are present through rejection of the neutral mutation hypothesis (Kimura 1979).

The Creevey-McInerney method (Creevey and McInerney 2002) is an extension of the McDonald and Kreitman test and assumes sequences are neutrally evolving when the ratio of replacement polymorphic/variable to replacement fixed/invariable mutations and silent polymorphic/variable to silent fixed/invariable are the same. In the Creevey-McInerney method (Creevey and McInerney 2002) a rooted phylogenetic tree is employed and the hypothetical ancestral sequences are reconstructed at each internal node using maximum parsimony (Hennig 1966) (see section 1.2.2 for description). All substitutions that occur across the phylogeny are then identified. Deviation from equality between these two ratios (replacement polymorphic/variable to replacement fixed/invariable) means deviation from the neutral expectation. If high rates of replacement invariable sites are observed, this indicates directional selection, or if high rates of replacement variable sites are observed it indicates non-directional selection (Creevey and McInerney 2002). While the Creevey-McInerney method has been shown to be effective in detecting adaptive evolutionary events within lineages it is not capable of identifying lineage-site specific evolution (Creevey and McInerney 2002), and therefore these methods have not been employed in this thesis.

A more explicit approach to identifying adaptive evolution involves the incorporation of evolutionary models in a maximum likelihood (ML) and Bayesian framework (Goldman and Yang 1994, Nielsen and Yang 1998, Yang and Nielsen 2002). These methods employ a model of evolution that can assess the probability of observing the data given the model of evolution. One approach to detecting positive selection in a ML framework is through the sitewise likelihood-ratio (SLR) method (Massingham and Goldman 2005). The SLR method was developed to detect non-neutral evolution and while it can give the estimates for the strength of evidence for each site under selection, it is not effective as the following method in estimating the strength of selection acting on each site (Massingham and Goldman 2005).

Codon substitution models have been developed to account for physiochemical property differences or distances between amino acids. These models also allow for variation in selective pressure across different sites in an alignment and across different branches in a phylogeny (Goldman and Yang 1994, Nielsen and Yang 1998, Yang and Nielsen 2002). These models were further developed to allow for calculation of ω values across all species for each amino acid in the MSA and calculation of ω values for one or a subset of species, termed "foreground lineages", as compared to all other taxa in the dataset and are described in detail below (Yang et al. 2000).

The codon models used to calculate ω values are nested likelihood models, with each model differing in complexity from the previous by the addition of parameters, the models have been described graphically in Figure 1.4. The standard nomenclature from the literature for these models has been retained in this thesis (Yang et al. 2000, Z. Yang et al. 2005, Wong et al. 2004). The simplest model, M0, calculates a value for ω over the entire alignment. The M0 model assumes all sites and all lineages are evolving at the same rate. Model M3 is an extension of M0 and allows the ω values to vary freely. There are two variations of the M3 model, m3(k = 2) discrete which allows two variable site classes, and m3(k = 3) which allows three variable site classes. M1a is a neutral model that allows two site classes for proportion of sites where $\omega_0 = 0$ and $\omega_1 = 1$. M2a is a selection model, and allows three ω site classes where $\omega_0 = 0$, $\omega_1 = 1$ and ω_2 is estimated and may be greater than 1. M7 is the beta model, which allows for a beta distribution β(p,q), and the beta distribution can take many different shapes depending on the values of p and q, but the ω estimation is bound between 0 and 1. The number of categories used to approximate the beta distribution is user defined and has been set to 10, for the work described in Chapter 4. Therefore the M7 beta model allows for 10 different ω site classes and is compared against the more parameter rich M8 (beta & omega > 1). M8 allows 10 different ω site classes and contains an additional 11^{th} parameter for ω that is free to vary between 0 and > 1. M8a (beta & omega = 1) is the null hypothesis of M8 where the 11^{th} category must be neutral (ω = 1). All site-specific models are represented graphically in Figure 1.4(A).

Codon-substitution models were developed for detecting positive selection at individual sites along specified lineages (Yang and Nielsen 2002). Two lineage-site models were developed, model A and model B and each work on a phylogeny that has been separated into "foreground" and "background" lineages. Model A assumes two site

classes are the same in both foreground and background lineages with $\omega_0 = 0$ (class 0) and $\omega_1 = 1$ (class 1), while an additional site class ω_2 (class 2) allows for calculation of $\omega_2 > 1$. Model B is similar to model A except that it allows estimation of site classes ω_0 and ω_1 from the data instead of being set *a priori*. In simulation studies, model B had a false detection rate for estimating positive selection of 33-66% when there was relaxation of purifying selection in foreground branches, while model A did marginally better with 19-54% false positives under the same conditions (Zhang 2004). Modifications were later made to model A allowing for additional site class estimation, there were two site classes belonging to the null model M1a distribution, i.e. $0 < \omega_0 < 1$ and $\omega_1 = 1$ for both foreground and background, and two additional classes that allowed for positive selection, i.e., $\omega_{2a} > 1$ and $\omega_{2b} > 1$, as shown in Figure 1.4(B) (Zhang et al. 2005). The null hypothesis of model A is referred to as "modelA-null" has fixed $\omega_2 = 1$ and allows for sites evolving under purifying selection, or neutrally evolving, in the background lineages. These improved models, i.e. modelA and modelA-null, were shown to perform well when used in conjunction with one another and out-performed the previous versions of modelA and modelB (Zhang et al. 2005).

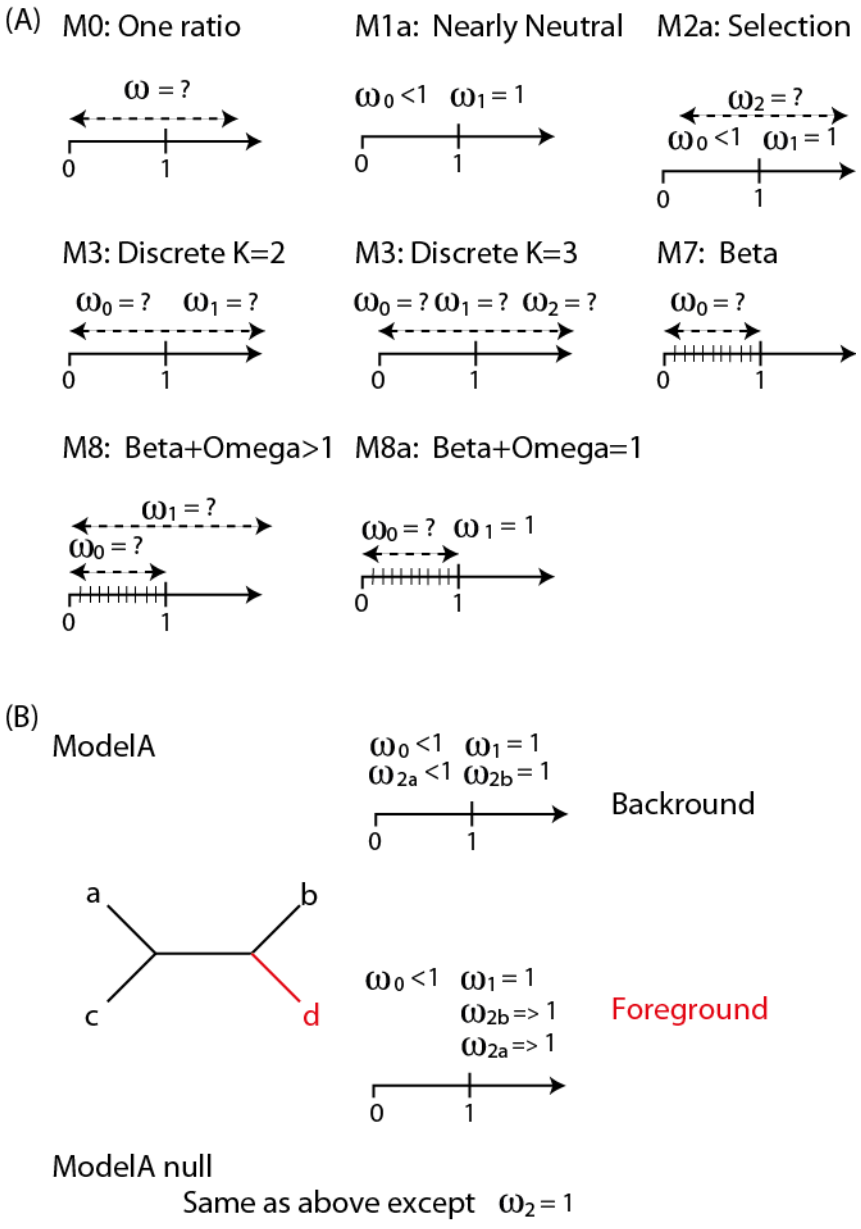


Figure 1.4 Codon models of Substitution implemented in CodeML from the PAML package (Yang 1997, Yang 1998).

The codon models of substitution employed from the CodeML software are illustrated. Codon models for site-specific analysis are shown in (A) and those for lineage-site are shown in (B). ω is as described in the main text. M0 is the simplest model, models differ from the previous by the addition of parameters in the order they appear in the figure (left to right) until M8, which is the most complex model site model. Foreground species (d) is shown in red and background species (a, b and c) are shown in black.

The codon models above are implemented in an ML framework and therefore could report results from a local minimum on the likelihood plane, for this reason all CodeML analyses performed in this thesis employ a variety of starting omega values (i.e., 0, 1, 2, 10) as in previous publications (Yang et al. 1998, Yang 1997, Loughran et al. 2008, Morgan et al. 2010). To assess the significance of more parameter rich models compared to their less parameter rich counterparts, a likelihood ratio test (LRT) was proposed (Nielsen and Yang 1998, Yang et al. 2000). The LRT follows a chi-squared distribution and the difference in lnL values (Δl) between two models is multiplied by 2 in most comparisons. The M3 Discrete (K=3) cannot be tested against any of the other models using LRT, but through comparisons with M3 Discrete (K=2), it can provide interesting results if the more parameter rich M3 Discrete (K=3) model has a better likelihood score. The degrees of freedom between the two models are then used to determine if the more parameter rich model is a statistically better fit than the less complex nested model. The degrees of freedom for each of the CodeML modes are as follows: M0 (df=1) , M1a (df=2), M2a (df=4), M3Dk2 (df=3), M3Dk3 (df=5), M7 (df=19), M8 (df=21), M8a (df=20), ModelA (df=10), Model A null (df=9). The comparison of M8 to M8a model follows a 50:50 mix of a χ^2 distribution with df=1 and therefore is significant at the 5% level if twice the Δl is $\geq$ 2.71. The models compared in the CodeML selective pressure analyses and summaries of the statistics are described in Table 1.1.

Table 1.1 Likelihood Ratio Test (LRT) Calculations

Comparison	df	Δl	Critical χ2 values
M0 v M3k2	2	X2	≥ 5.99
M3Dk2 v M3Dk3	-	X1	≥ 1.00
M1a v M2a	2	X2	≥ 5.99
M7 v M8	2	X2	≥ 5.99
M8 v M8a	1	X2	≥ 2.71 (@5%) ≥ 5.41 (@1%)
M1a v Model A	2	X2	≥ 5.99
Model A v Model A null	1	X2	≥ 3.84 (@5%)

The comparison of nested models permitted in the analysis, degrees of freedom (df), the amount the difference in the lnL scores between models (Δl) is multiplied by the value in the third column, the χ2 critical values are shown in the last column. Table adapted from (Morgan et al. 2010).

The posterior probability (PP) of a specific amino acid site belonging to the positively selected category is calculated by applying Empirical Bayes estimates (EB) to the ω ratio on a per site basis. There are two EB methods employed: (i) Naïve empirical Bayes (NEB) (Yang et al. 1998), and (ii), Bayes empirical Bayes (BEB) (Yang et al. 2005). The NEB method is sensitive to error in small datasets where ML estimates may have large sampling error and thus can result in false positive inference of sites under positive selection (Anisimova et al. 2002). The BEB is more robust as it assigns a prior to the model parameters and integrates over uncertainties which reduces the rate of false positive detection when analysing small sample sizes (Yang et al. 2005).

Analysis by Friedman and Hughes (2007) seemed to illustrate that signatures of positive selection detected using CodeML were influenced entirely by GC3 content and the underlying Ds rate, rather than accurately reflecting the ω ratio. This study was later disputed and it was determined that it was in fact a misunderstanding of how the codon models work - the false positive rate of LRTs in CodeML is in fact now estimated to be < 5% (Zhai et al. 2012).

1.1.7 Factors affecting mutational rate variation in mammals

The rate of mutation is not consistent in all lineages across evolution, this is true for mammals also. Species-specific rates of evolution can be affected by life traits such as longevity and body size (Peto et al. 1975), metabolic rate (Martin and Palumbi 1993), germ-line generation time (Wilson et al. 1977), and variations in species-specific DNA repair mechanisms (Britten 1986, Hart and Setlow 1974). Each of these factors is dealt with individually below but they are all tied together in a specific lineage and therefore an observed increase in mutation rate in a given species could be due to a number of these factors in combination.

The cells of larger mammals with long life spans undergo more rounds of cell division and are exposed to more mutagens over their life time, however they are not more prone to cancer – this is known as Peto's paradox (Peto et al. 1975). It is proposed that larger mammals have adapted their DNA repair networks to be more efficient in coping with the cancer risk associated with their increased longevity and higher number of cell divisions (Peto et al. 1975). Recent improvements in phylogenetic models that assess the relationship between life-trait variations and substitution rates, have shown a negative correlation between the rate of substitution and mass and longevity (Lartillot and Poujol 2011). Empirical evidence is still relatively sparse for cancer protecting mutations/mechanism but there is some support from the literature for lineage specific adaptation. Different anti-cancer mechanisms that control cell proliferation have evolved across rodents in correlating to their body size (Seluanov et al. 2007). Mechanisms such as replicative senescence controls cell proliferation by successively shortening telomeres after each round of cell division (expanded on in Chapter 4), and "cell-to-cell contact inhibition" which control cell proliferation through initiation of apoptosis when cells are in contact with one another (Seluanov et al. 2009). A recent selective pressure analyses of the BRCA/FANC DNA damage response pathway displays lineage-specific adaptations acting on mouse and human lineages (O'Connell 2010).

The metabolic rate hypothesis (Martin and Palumbi 1993) states that smaller-vertebrates generate higher levels of mutagenic oxygen radicals due to their higher mass-specific metabolic rates compared to larger vertebrates (Glazier 2005). Oxygen radicals are a by-product of metabolism (Barja 1999) and can cause mutations in DNA (Cooke et al.

2003). This suggests that mammals with high metabolic rates will produce more free radicals and would experience higher mutational rates than mammals with low metabolic rates (Martin and Palumbi 1993, Bleiweiss 1998). There has been empirical evidence for this through estimation of silent substitution rate from nuclear DNA and metabolic rates in primates (Martin and Palumbi 1993).

The germ-line generation time hypothesis, proposed by Wilson *et al*. (1977), states that mammals with short germ-line generation times will produce more offspring and will therefore undergo more rounds of meiosis per unit of time as compared to mammals with longer germ-line generation times (Wilson et al. 1977). Mammals with short germ-line generation times therefore will accumulate more DNA changes which has been observed by Martin and Palumbi (1993) in the analysis of rates of change occurring within the cytochrome b gene and globin data (Martin and Palumbi 1993). Studies show higher rates of nucleotide substitution in rodents compared to humans is strongly linked to the germ-line generation time (Wu and Li 1985, Li et al. 1987).

Species that have increased mutation rates due to any of these traits listed above are also observed to have a higher GC content due to GC-biased gene conversion (gBGC) in the DNA repair machinery (Galtier 2003, Escobar et al. 2011). The efficiency of DNA repair mechanism is reported to vary across species (Britten 1986, Hart and Setlow 1974) and therefore an increase in fixation of GC has been shown to increase in frequency in eukaryotes (Pessia et al. 2012) and also specifically in mammals (Lartillot 2012).

Factors such as diet (Yang 1998), intense sexual selection (Dorus et al. 2004) and pathogen load (Usanga and Luzzatto 1985, Fumagalli et al. 2011) can also influence lineage-specific variation in mutation rate. Different selective pressures acting on different species across the phylogeny cause lineage-specific heterogeneity in mutation rate that in turn influences base composition. These phenomena have a massive impact on phylogeny reconstruction (Foster 2004, Foster et al. 2009), as discussed in more detail in section 1.2.2, and in turn this impacts on accurate identification of positive selection using phylogeny-based methods.

1.1.8 Non-adaptive evolutionary signals mistaken as positive selection

Recombination is the exchange of genetic information between a pair of nucleotide sequences and has been shown to introduce variability to populations (Posada and Crandall 2001, Anderson and Kohn 1998, Feil et al. 2001) and influence the process of natural selection (Marais et al. 2001). The rate at which nucleotide substitutions occur and fix is impacted by the rate at which recombination occurs, and this can impair accurate phylogeny reconstruction (Posada and Crandall 2002) which in turn forms the foundation for LRT analyses of selective pressure variation (Anisimova et al. 2003). Nucleotide mismatches can occur at recombination breakpoints that are repaired by the DNA machinery, and if gBGC is prevalent, this results in large proportions of GC mutational fixations (Dreszer et al. 2007, Katzman et al. 2011). This increase in GC fixation can impact the calculation of substitution types and cause inflation of the ω ratio (Berglund et al. 2009, Galtier et al. 2009, Ratnakumar et al. 2010). It is therefore unsurprising that gBGC and the rate at which recombination occurs are correlated (Duret and Arndt 2008) and that both are critical in the assessment of selective pressure heterogeneity.

Both gBGC and recombination are correlated with N_e, as species with large N_e tend to be more affected by gBGC (Webster et al. 2006, Duret and Arndt 2008). A small N_e can cause a higher rate of fixation of weakly deleterious mutations (Eyre-Walker 2002, Woolfit and Bromham 2003). All of these scenarios can be misinterpreted as positive selection. The N_e varies significantly between species and within populations. Previous studies have shown that ω estimations tend to be higher in larger mammals, which may be a result of a smaller overall N_e (Popadin et al. 2007). The accuracy of N_e calculations is debatable or unknown for many mammal species, e.g. human N_e has been estimated to be between 3100 and 7500 (Tenesa et al. 2007). Analysis of mitochondrial and nuclear coding genes show that the ω estimation is inversely proportional to the N_e therefore a higher proportion of false positives would be expected from species with small N_e (Lartillot 2012).

The interactions between adaptive and non-adaptive events are complex and further development of methods are required to adequately tease apart the effects of these evolutionary processes (Lartillot 2012).

1.2 Approaches to phylogeny reconstruction using molecular data

Evolutionary relationships resemble a tree like structure, therefore terminology such as root, branch, node and leaf are used to describe the different parts of a phylogenetic tree, see Figure 1.5(A) (Salemi and Vandamme 2003). Extant taxa are referred to as leaves on a phylogenetic tree while internal nodes represent hypothetical ancestors of the extant taxa. In Figure 1.5(A), the taxa "a", "b" and "c" form a cluster and share a common ancestor (i). The ((a,b),c) cluster can be referred to as a monophyletic group and node "i" is their most recent common ancestor (MRCA). The order in which the nodes occur (what taxa are clustered and in what order) is referred to as the tree topology. A tree is rooted if a taxon or set of taxa is the most distantly related to the rest of the taxa on the tree, e.g. in Figure 1.5(A) taxon "g" is the most distantly related and the remaining taxa form the ingroup. A rooted tree has directionality; whereas, for an unrooted tree the directionality is unknown, see Figure 1.5(B). It is worth noting that for prokaryotes "tree-thinking" is not explaining the observed data, the high levels of horizontal gene transfer are so disruptive to the pattern of vertical descent that alternative frameworks for describing the evolution of these entities is now necessary (McInerney et al. 2011). The mammalia are the major focus of this thesis, and these can be represented by a traditional phylogeny of vertical descent.

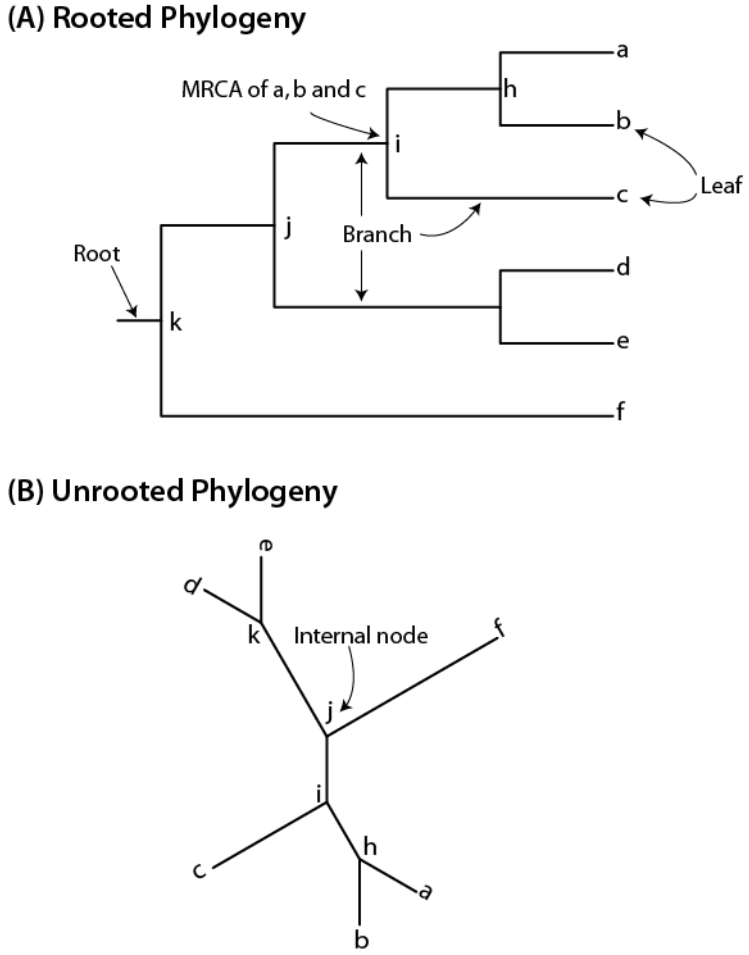


Figure 1.5 Comparison of rooted and unrooted tree structures.

A rooted phylogenetic tree (A) and an unrooted phylogenetic tree (B) are displayed. Both (a) and (b) display the same topological information, the only difference between the trees is the presence of a root. Adapted from (Salemi and Vandamme 2003).

1.2.1 Modelling Evolution

Several phylogenetic reconstruction methods employ an evolutionary model, which can clarify the features of how sequences have changed over time. Methods that are not model based also have assumptions on how the sequences are related but they are not explicit (traditionally Parsimony has been used as one such example although this may now be considered controversial (Steel 2011)). An evolutionary model is composed of two parts - the tree and the process. The process is subdivided into the exchange rate of characters and the composition vector. The first evolutionary model was the Jukes-Cantor (JC) model (Jukes 1969) which assumes that the exchange rate between characters (A, C, G and T) are equally likely, as are the base frequencies in the composition vector. Kimura expanded on the Jukes-Cantor (JC) model (Jukes 1969) by allowing the exchange rate of characters to have two different substitution types, transitions and transversions, giving the Kimura 2 parameter model (K2P) (Kimura 1980b). Felsenstein expanded on the Jukes-Cantor (JC) model by allowing unequal base frequencies, giving the Felsenstein (F81) model (Felsenstein 1981). Therefore, nested within both F81(Felsenstein 1981) and K2P(Kimura 1980b) models is the original Jukes-Cantor (JC) model (Jukes 1969). In this way the process of model development has proceeded through incremental parameterisation of base frequencies and/or exchange rates among characters and has lead to several models of increased complexity (Kimura 1981, Zharkikh 1994, Hasegawa et al. 1985, Tamura and Nei 1993), up to and including the general time reversible (GTR) model that allows for estimated base frequencies and all possible substitutions to differ (Tavaré 1996).

Not all characters in a sequence evolve at the same rate, and models that account for associated site rate variation (ASRV) using a gamma distributed rate variation among sites (+Γ) are a better fit to biological data than their simpler alternatives (Yang 1996). It is also known that not all sites in an alignment will vary and so the +I parameter was developed to account for these unchanging or invariable sites (Reeves 1992). The use of the +I and +Γ parameter together is not recommended (Yang 2006) as it has been shown that the gamma distribution already accounts for the invariance and when both parameters are used simultaneously it is difficult to accurately estimate each parameter as +I leads to a change in the estimation of +Γ and visa versa (Mayrose et al. 2005).

In addition to developing models for nucleotide data, a separate set of evolutionary models were developed for amino acid data that account for the 20 character states in

this data type. The simplest amino acid model is the Poisson model (Bishop and Friday 1985), which is sampled from the JC model but extended to 20 characters instead of 4. The Poisson model assumes that all changes between amino acids are equal as well as the probability of change between states being equally likely. The Poisson model (Bishop and Friday 1985) does not take into account the different chemical properties of amino acids. Dayhoff (1978) developed an amino acid model that could account for the probability of change among amino acids (Dayhoff 1978). This model was used as a stepping-stone to create standardized amino acid models for phylogeny reconstruction. The amino acid models frequently employed are empirical and are based on experimental data. One such example is the Jones Taylor Thornton (JTT) model which is an extension of the Dayhoff model but is based on a mutation data matrix estimated from transmembrane proteins (Jones et al. 1992). The mtRev model is based on the observed mutations in complete mitochondrial sequences of 20 vertebrate sequences (Adachi and Hasegawa 1996). The WAG model is based on mutation data matrix estimated from globular proteins (Whelan and Goldman 2001). The Blosum62 rate matrix differs from the other empirical amino acid rate matrices as it was estimated using blocks of sequence data from groups of proteins rather than being estimated from the entire multiple sequence alignment (Henikoff and Henikoff 1992). Blosum62 is therefore quite a general model that is good at identifying distantly related proteins and is used in the sequence similarity search program BLAST (Altschul et al. 1990). Frequently employed empirical amino acid models and the datasets from which they were derived are listed in Table 1.2.

Table 1.2 Empirical Amino Acid Models of Evolution

Model	Dataset Used	Reference
Poisson	None	(Bishop and Friday 1985)
WAG	Globular proteins	(Whelan and Goldman 2001)
Blosum62	Blocks from related proteins	(Henikoff and Henikoff 1992)
CPREV	Plastid genomes	(Adachi et al. 2000)
Dayhoff	Protein sequences > 85% identical	(Dayhoff 1978)
JTT	Transmembrane proteins	(Jones et al. 1992)
mtRev	mtDNA vertebrate species	(Adachi and Hasegawa 1996)
mtMam	mtDNA mammal species	(Yang et al. 1998)
rtRev	Retroviral Pol proteins	(Dimmic et al. 2002)
LG	Protein families	(Le and Gascuel 2008)
VT	Protein families	(Muller and Vingron 2000)

The gamma distribution of rates across sites parameter (Yang 1996) can also be applied to empirical protein models, however, they still assume homogeneity of composition and exchange rates across sites and across lineages, and have previously been shown to cause systematic errors (Foster 2004). In Chapter 2 these models are explored using previously published data and specifically assembled datasets.

Dataset partitioning has been employed to accommodate heterogeneity in nucleotide datasets by allowing different substitution rates at different codon positions (Nylander et al. 2004). In amino acid data it has been implemented to subdivide a gene or a set of genes into different partitions allowing for different rates of change, or different evolutionary models, to account for heterogeneity within the data (Nylander et al. 2004). Choosing the correct partition for the dataset requires *a priori* knowledge about the data and/or assumptions about the data and in this way can be problematic (Brandley et al. 2005, Shapiro et al. 2006). Mixture models have an advantage over data partitioning methods as they are able to determine the parameter and model distribution directly from the data (Pagel and Meade 2004).

The development of methods that account for compositional heterogeneity and exchange rate heterogeneity across the phylogeny and the dataset were a major improvement to available models (Foster 2004, Lartillot and Philippe 2004). These models are defined as node-discrete composition heterogeneity (NDCH) and node-discrete rate matrix heterogeneity (NDRH) (Foster et al. 2009). Lartillot and Philippe (2004) developed a mixture model that allows each site class to have a specific biochemical profile that is a probability vector over the 20 amino acids (CAT) (Lartillot and Philippe 2004). Each profile defines a simple amino-acid replacement process where each time a substitution event occurs, a new amino acid is chosen at random according to the probabilities defined by the profile, this is referred to as an amino acid replacement process. The likelihood at each site of the alignment is then averaged over all available processes defined by the mixture. It has been shown that mixture models such as CAT and the combination of CAT with the previously described GTR model, known as CAT-GTR, perform far better than one matrix or empirical models on datasets where saturation or long branch attraction is a problem (Lartillot et al. 2007). While these models have many important and significant improvements on previous models they do have specific requirements. To accommodate the more parameter rich models it is necessary to have a dataset of sufficient size, for CAT and CAT-GTR (Lartillot and Philippe 2004) it has been demonstrated that an alignment of longer than 1,000 and 10,000 amino acids respectively are necessary. While CAT works well at incorporating mixture models at the amino acid substitution level it does not allow for compositional heterogeneity among lineages (Lartillot and Philippe 2004). A extension of the CAT model called CAT-BP (Blanquart and Lartillot 2008) was developed to accommodate composition bias between lineages by introducing "break points" along the branches of the topology. Application of the CAT-BP model is extremely computationally demanding and convergence of parameters under this process has been shown to be difficult to achieve (Nesnidal et al. 2010). Therefore when data requires the use of heterogeneous models it is important to explore methods that can account for heterogeneity of exchange rates and composition over the phylogeny (Foster 2004), and over the data (Lartillot and Philippe 2004). These heterogeneous models have been applied to mammal datasets in Chapter 2.

1.2.2 Problems affecting phylogeny reconstruction

Phylogeny reconstruction given a set of homologous characters should be relatively straight forward assuming the characters are reliable. The addition of more characters should result in convergence towards the correct answer (Felsenstein 1988). However, when discordance among gene trees is present, species tree reconciliation is not improved by the addition of more taxa (Degnan and Rosenberg 2009). The contributing factors to gene trees not reconciling with the species tree have been discussed in section 1.2.1.1. When methods are inconsistent (i.e. are unable to recover the same tree through repeat experiments), and when data contains systematic biases, there can be increase in the support for wrong phylogeny with the addition of more data (Felsenstein 1978, Hendy and Penny 1989, Huelsenbeck 1995, Huelsenbeck and Hillis 1993). These biases include homoplasy, compositional biases, long-branch attraction (LBA) and heterotachy and are each detailed in turn in the following sections.

1.2.1.1 Gene Tree Species Tree Discordance

Discordance across gene trees has frequently been observed, and examples include; plant species such as pines (Syring et al. 2007), hominoids (Chen and Li 2001) and Australian grass finches (Jennings and Edwards 2005). One of the major contributing factors is incomplete lineage sorting (ILS) which is the failure of two or more gene lineages in a population to coalesce. ILS occurs when the expected time for gene lineages to coalesce, based on the effective population size, occurs prior to the speciation event itself (Degnan and Rosenberg 2009). Therefore, ILS occurs more frequently on shallow parts of a species tree (short branches) and coalescence can occur between lineages that are not closely related causing incongruence between gene trees, and between gene and species trees (Degnan and Rosenberg 2009). ILS has been hypothesised to be the primary cause of discordance on the intra-order placement within the Laurasiatheria Superorder in placental mammals (Hallstrom and Janke 2010) this is discussed in Chapter 2. Concatenated datasets, also known as Supermatrices (see section 1.3.3.1), are particularly susceptible to ILS as a single phylogeny is generated to represent all the data (Philippe et al. 2004).

Horizontal gene transfer (HGT) is a major cause of gene tree discordance in prokaryotic species (Philippe and Douady 2003), also mentioned in section 1.1.4. HGT is extremely rare in eukaryotic species, but introgression of gene flow due to hybridization does

occur and can disrupt gene tree species tree reconciliation (see section 1.4.1) (Rieseberg et al. 2000). Recombination events, which are discussed in detail in section 1.1.8, can cause different segments of a gene to have a different coalescence time and this has been shown to cause discordance between gene trees and low node support (Posada and Crandall 2002). The concept of differential retention and loss of duplicates is an important complicating factor in phylogeny reconstruction as it can lead to the incorrect identification of 1:1 orthologous relationships across species and can result in discordance among gene trees and between gene and species trees (Philippe et al. 2011). Finally the process of natural selection, as described in section 1.2, as a driving force of speciation also plays a role in gene tree distributions differing from the multispecies coalescence (Degnan and Rosenberg 2009).

1.2.1.2 Lack of a molecular clock and phylogeny reconstruction

A molecular clock refers to the regular accumulation of mutations over time. This concept came from the neutral theory of evolution (Kimura 1968) where mutations accumulated at a relatively constant rate across lineages through the process of genetic drift. If the molecular clock held, then the number of mutations that have occurred in a lineage could be used as a proxy to calculate the divergence date of a particular species or conversely having the date from the fossil record would allow us to estimate the number of mutations expected. The molecular clock hypothesis was proposed following the observation of the α-globin gene accumulates mutations at a rate proportional with time (Zuckerkandl 1962), depicted in Figure 1.4(A). This is one of the few cases where a protein evolves in this clock-like way. Since its initial proposal the molecular clock has been disproven in its purest form and the current literature illustrate how mutations do not occur in a clock-like fashion: examples are seen from invertebrates (Thomas et al. 2006), mammals (Li et al. 1996, Gu and Li 1992, Yang and Nielsen 1998), arthropods (Ayala 1997, Rocha-Olivares et al. 2001) and plants (Bousquet et al. 1992). The significance of the abandonment of the molecular clock for phylogeny reconstruction was illustrated by Yang in 1996 when he showed that if the rate of change of characters is assumed to be constant, then the incorrect topology will be found (Yang 1996).

1.2.1.3 Compositional bias and phylogeny reconstruction

Compositional bias occurs when species incorrectly cluster together on a phylogeny due to shared similarity of composition. A phylogenetic analysis of the relationships between *Thermus* and *Deinococcus* for example showed that high GC composition in *Aquifex* (73%) and *Thermus* (72%), erroneously grouped these two species together with the exclusion of *Deinococcus* (52%) and *Bacillus* (50%) (Embley et al. 1993), see Figure 1.6(B). There is a large amount of evidence from the literature (mostly from the early 1990's) for the impact of compositional heterogeneity on incorrect phylogeny reconstruction (Loomis and Smith 1990, Penny et al. 1990, Hasegawa and Hashimoto 1993, Embley et al. 1993, Sidow and Wilson 1990, Lockhart et al. 1992a, Lockhart et al. 1992b). Since then compositional heterogeneity has been observed across the Metazoa (Nesnidal et al. 2010), *Drosophila* (Carulli et al. 1993) and *Mammalia* (Romiguier et al. 2010). Translating nucleotide sequences into amino acids can ameliorate the problems of compositional bias (Loomis and Smith 1990, Hasegawa and Hashimoto 1993, Hashimoto et al. 1995). Bias at the DNA level can affect amino acid content as has been shown for mitochondrial encoded genes (Foster et al. 1997), and compositional bias from both DNA and amino acid characters have been reported (Foster and Hickey 1999). This is of particular interest in the study of mammals where, as seen in section 1.1.7 and 1.1.8, biased gene conversion is known to occur.

1.2.1.4 Long Branch Attraction and phylogeny reconstruction

Long branch attraction (LBA) occurs when sequences with higher mutational rates, that do not share a MRCA, are clustered together erroneously; see Figure 1.6(C). This systematic error was originally described by Felsenstein and is also referred to as the "Felsenstein Zone" (Felsenstein 1978). LBA can equally be thought of as short branch attraction, as the slower evolving taxa can also be incorrectly drawn together by their shared ancestral traits (symplesiomorphy) (Philippe et al. 2005b). LBA significantly impacts phylogeny reconstruction (Philippe 2000, Bergsten 2005) and has been frequently observed in mammalian datasets (Reyes et al. 2000, Loughran et al. 2008).

1.2.1.5 Homoplasy and phylogeny reconstruction

Homoplasy is the similarity between sequences or species which is not the result of descent from a common ancestor, and can be further described as either (i) reversal, (ii)

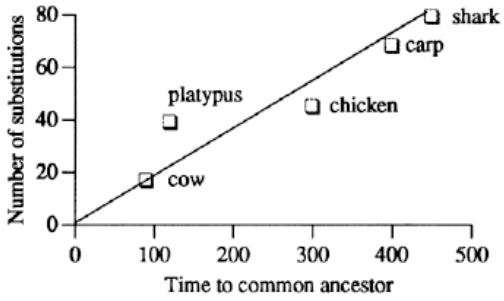
parallelism, or (iii), convergence of character states, see Figure 1.6(D) (Sanderson and Hufford 1996). Reversal occurs when a character or feature present in an extant taxon was also present in a distant lineage but not in the recent ancestor of that taxon. This has been observed in angiosperms where characteristics similar to the ancestral state (brightly coloured, long petals) have reoccurred, i.e. a switch in pollination strategy from the derived state of insect pollinated (scented short white petals) to the ancestral state of humming bird pollinated (Thomson 2008). Parallelism is the independent evolution of similar traits or characters from an ancestral condition. Parallel evolution has been observed between marsupial and placental mammals where species such as marsupial and placental moles, tasmanian wolf and the European wolf, the European sabre-tooth tiger (*Smilodon*) and the South American sabre-tooth tiger (*Thylacosmilus*), have independently evolved similar morphological features from their Therian ancestor. Rokas (2008) observed extremely high levels of parallel evolution in eukaryotic proteins, and estimated the frequency of homoplastic characters to be 2-fold higher than expected under the neutral model of protein evolution (Rokas and Carroll 2008). Convergent evolution occurs when independent species evolve the same character or trait from different ancestral states. The first example of this was recognised in the adaptation of lysozymes in foregut fermenters (Kornegay et al. 1994, Stewart and Wilson 1987, Stewart et al. 1987), phylogenetic reconstruction using this character placed cows within the primate clade, in direct conflict with multiple studies and a variety of methods that place Cow it within the Cetartiodactyla Order of the Laurasiatheria. At the molecular phylogenetic level, homoplastic characters and particularly convergent evolution has been shown to give radical inconsistencies between phylogenies inferred from mitochondrial genomes (Castoe et al. 2009).

1.2.1.6 Heterotachy and phylogeny reconstruction

Heterotachy is the heterogeneous rate at which a site in a sequence evolves over time (Philippe and Lopez 2001). As functional constraints and positive selection can act on different parts of a gene sequence, the rate of substitution along a sequence is not uniform, as shown in Figure 1.6(E). This is a common occurrence in sequence alignments (Lopez et al. 2002, Ane et al. 2005, Philippe et al. 2005b, Taylor et al. 2006) and when it is not modelled correctly, heterotachy can cause inconsistencies in phylogeny inference (Kolaczkowski and Thornton 2004). Concatenated datasets are particularly prone to having incongruence's in branch lengths which can result in

overall incorrect topological reconstruction (Matsen and Steel 2007). The mixture of branch lengths (MBL) approach was developed by Kolaczkowski and Thornton (2004) and attempted to account for heterotachy. Increased parameter requirements to adequately model independent rates of site substitution across all taxa is however, computationally demanding (Kolaczkowski and Thornton 2004). Covarian models were developed to accommodate heterotachy by allowing sites switch from variable to invariable states (Tuffley and Steel 1998), but this model is limited as it assumes the rate at which sites shift are site-independent (Zhou et al. 2007).

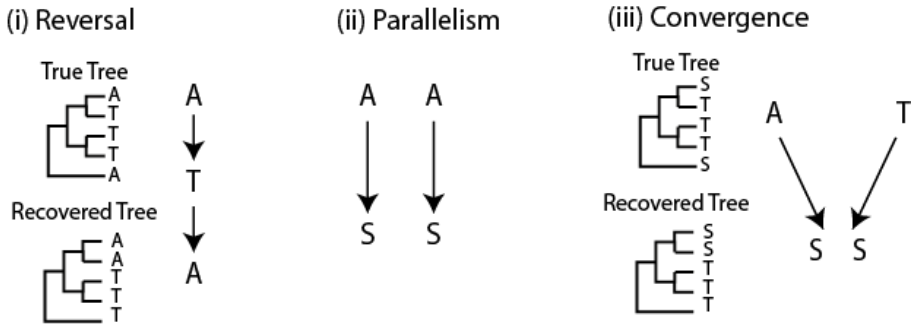
(A) The molecular clock hypothesis, adapted from (Zuckerkandl 1962)



(B) Composition Bias

(C) Long Branch Attraction

(D) Homoplasy



(E) Heterotachy

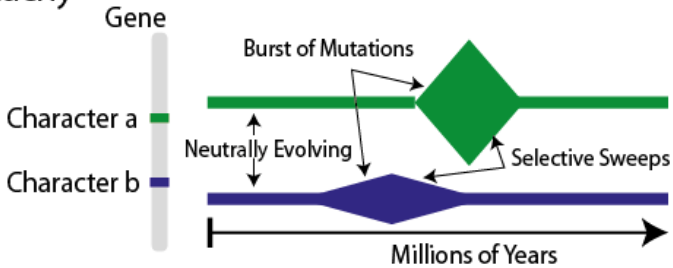


Figure 1.6 Systematic Errors that affect Phylogeny reconstruction

Figure 1.6 legend: Factors that affect phylogeny reconstruction (A to E) are shown. The molecular clock for α-globin (A) shows the time to the common ancestor in millions of years on the x-axis. The number of amino acid substitutions in the α-globin gene sequence is recorded for species with respect to human and these values are on the Y-axis. This image has been adapted from the original publication (Zuckerkandl 1962). Panel (B) illustrates the effects of compositional bias causing the resolved tree to incorrectly group *Aquifex* with *Deinococcus*. Panel (C) shows an example of LBA causing the reconstructed phylogeny to erroneously cluster "Long1" with "Long2" to the exclusion of "Short1" and "Short2". Panel (D) displays the three types of homoplasies at the amino acid level: (i) reversal, (ii) parallelism, and (iii), convergent evolution. The ancestral character state and the derived character state along with the impact on the phylogeny reconstruction are depicted. Panel (E) depicts Heterotachy acting on character "a" and "b". On the right is a diagram depicting the evolutionary history of these genes over millions of years. The thin horizontal lines represent periods of conservation, while the diamonds represent periods of rapid evolution followed by selective sweeps. These evolutionary changes occur at different times and at different rates for both character "a" and "b".

1.2.2 Methods of Molecular Phylogenetic Analysis

Resolution of phylogenetic relationships involves reconstructing the evolutionary history of species based on homologous characters. The earliest methods were distance matrix methods and involved the percent sequence distance calculated for all pairwise combinations of sequences. Taxa were subsequently arranged on a tree based on their sequence distance from one another (Baldauf 2003). Algorithms such as neighbor-joining (NJ) (Saitou and Nei 1987), minimum evolution (ME) (Kumar 1996), Fitch-Margoliash (Fitch and Margoliash 1967) and UPGMA all work under variations on this theme. These methods are consistent if sequences are very similar, however they are less efficient than maximum likelihood (ML) and Bayesian inference (BI) based methods (Kuhner and Felsenstein 1994).

Discrete data methods include parsimony, maximum likelihood (ML) and Bayesian inference (BI), and these approaches allow for tree searching under a model of evolution. Parsimony works by determining the minimum number of steps that are needed to fit alternative character states onto a tree or set of competing topologies (Figure 1.7). Parsimony methods exclude invariable sites and sites that are unique to individual taxa. There are two forms of parsimony, (1) Fitch parsimony, or (2), weighted parsimony. Fitch parsimony involves an equal weighted matrix where every character change is equally possible - this method is susceptible to signal issues from homoplastic characters leading to biases in the phylogenetic analysis. Weighted parsimony attempts to overcome this bias by incorporating matrices that account for physiochemical changes of amino acids or weighted changes of transitions (purine-purine) versus transversions (purine to pyrimidine, and, vice versa) at the nucleotide level. Felsenstein showed that parsimonious methods are extremely inconsistent in the presence of LBA (Felsenstein 1978). Both distance based methods and parsimony methods have become outdated. Improvements in computational power and models of evolution have lead to frequent use of likelihood (including Bayesian) based methods that have been shown to correct for inconsistencies such as LBA when the correct model is employed (Huelsenbeck et al. 2001). Likelihood based methods also have the desirable feature of using all the available data in their analyses (unlike parsimony, where large portions of data are removed). The assumptions of likelihood are explicit, not implicit, and ancestral states are estimated using a probabilistic framework (Huelsenbeck et al. 2001). Bayesian and maximum likelihood are the two likelihood-

based methods employed throughout this thesis, both have been expanded on further in sections 1.2.2.1 and 1.2.2.2.

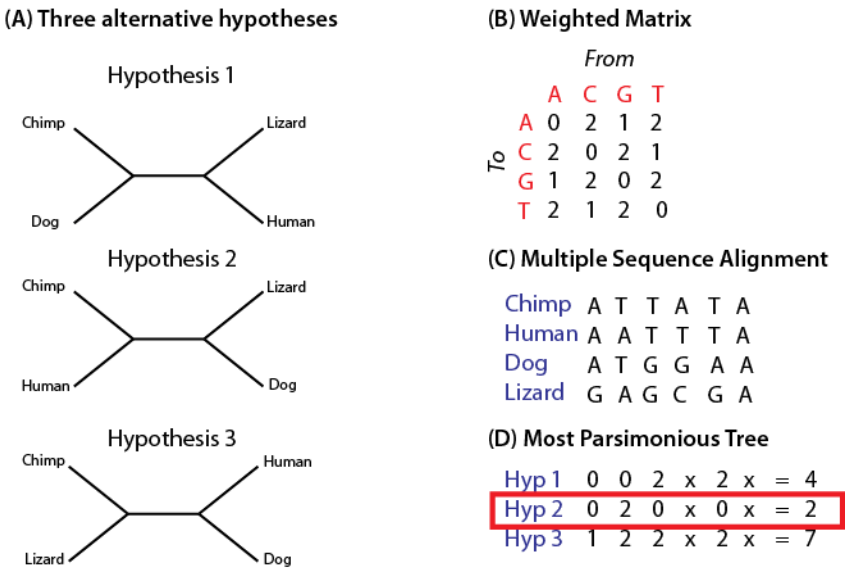


Figure 1.7 Parsimony Reconstruction Methods.

The 3 possible unrooted topologies for 4 taxa: chimp, human, dog and lizard are shown in (A). The weighted matrix applied in this case is shown in (B) and scores nucleotides staying the same as 0, a transition as 1 and a transversion as 2. The multiple sequence alignment for the 4 taxa is shown in C. Using these alternative hypotheses for the relationships between the 4 taxa and using weighted parsimony, values for all hypotheses in (D) are shown and the most parsimonious tree is the one with the fewest number of changes required, i.e. hypothesis number 2 is the favoured tree.

1.2.2.1 Maximum Likelihood

Maximum Likelihood (ML) for phylogenetics was first introduced for gene data in 1964 (Huber 1964), but it was through the work of Felsenstein and the application of his "pruning algorithm" that allowed ML to be applied to sequence data (Felsenstein 1981). The maximum likelihood approach for phylogeny resolution is based on an aligned sequence dataset (X) that is fixed, and, a topology (τ) that is free to vary. The likelihood of the observation is based on the probability of observing the aligned sequence data (X) given the model of evolution, which is composed of a tree (τ) including branch lengths (υ), and a model of evolution with a set of parameters (θ), as shown in Eqn. 4.

Eqn. 4 Maximum Likelihood Equation

$$L(\tau,\upsilon,\theta)=Pr(X|\tau,\theta)$$

To illustrate the mechanisms of likelihood estimation the following example has been adapted from (Foster 2001), and follows the illustration in Figure 1.8. If a pairwise alignment of two sequences Seq1 and Seq2, as shown in Figure 1.8(A), and a model that is composed of an exchange rate matrix (R) and a composition vector (π) is considered, as shown in Figure 1.8(B), then the likelihood of the observed process from Seq1 to Seq 2 would be:

$$L= \pi_A . R_{(A \rightarrow A)} . \pi_T . R_{(T \rightarrow T)} . \pi_T . R_{(T \rightarrow C)} . \pi_C . R_{(C \rightarrow C)}$$
$$L=(0.1)(0.976)x(0.3)(0.979)x(0.3)(0.01)x(0.4)(0.983)$$
$$L=0.000033813375552$$

Branch lengths are accounted for by altering the character exchange rate matrix. In this example the branch length is short and therefore there is a high chance of a character remaining the same. If branch lengths were longer, i.e. evolving faster or separated by greater evolutionary distance, then the probability of the character remaining the same would be lower, this is illustrated in Figure 1.8(C). The Certain Evolutionary Distance (CED) of this model is 1. If the CED were increased to 2, then it would raise the power of the rate matrix to 2, see Figure 1.8(C). To compute the ML of the data the number of CED units is increased until the ML is reached, this tipping point effect is shown in Figure 1.8(D). After this point the likelihood begins to decrease again further increase

in CED units overestimate the branch lengths. The maximum likelihood model in this example is the number of CED units needed to reach the highest possible score for the likelihood, in this case 0.00018 at a CED unit of ~18.

ML methods have been shown to outperform neighbor-joining methods (Huelsenbeck 1995). ML is also more effective at resolving the phylogeny than parsimony based methods for datasets with LBA (Kuhner and Felsenstein 1994) and heterogeneous rates (Gadagkar and Kumar 2005), assuming the correct model is employed. Maximum likelihood is however a frequentist based approach and as such does not account for uncertainties within phylogeny (Huelsenbeck et al. 2002).

(A)

Seq 1: ATTC
Seq 2: ATCC

(B)

$$P = \begin{bmatrix} 0.976 & 0.01 & 0.007 & 0.007 \\ 0.002 & 0.983 & 0.005 & 0.01 \\ 0.003 & 0.01 & 0.979 & 0.007 \\ 0.002 & 0.013 & 0.005 & 0.979 \end{bmatrix}$$

$$\pi = [\,0.1, 0.4, 0.2, 0.3] \quad (A\ C\ G\ T)$$

(C)

$$\begin{bmatrix} 0.976 & 0.01 & 0.007 & 0.007 \\ 0.002 & 0.983 & 0.005 & 0.01 \\ 0.003 & 0.01 & 0.979 & 0.007 \\ 0.002 & 0.013 & 0.005 & 0.979 \end{bmatrix}^2 = \begin{bmatrix} 0.953 & 0.02 & 0.013 & 0.015 \\ 0.005 & 0.966 & 0.01 & 0.02 \\ 0.007 & 0.02 & 0.959 & 0.015 \\ 0.005 & 0.026 & 0.01 & 0.959 \end{bmatrix}$$

(D)

CED	Likelihood
1	0.00003
2	0.0000559
3	0.0000782
10	0.000162
15	0.000177
20	0.000175
30	0.000152

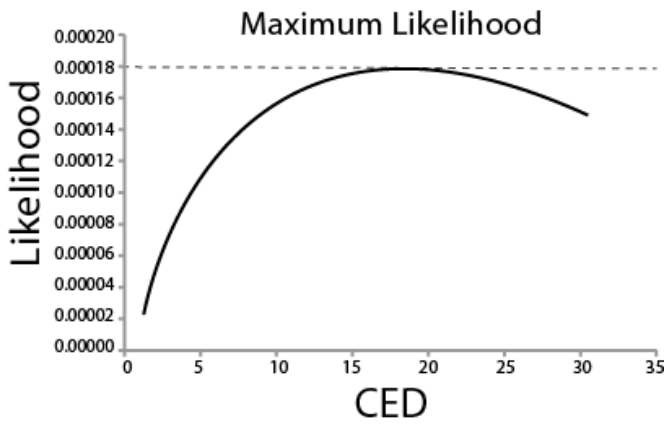


Figure 1.8 Cartoon describing a Maximum Likelihood approach for determining the optimum CED of a dataset.

The process of calculating the maximum likelihood of MSA (A) under the model of evolution (B) where the order of the nucleotides is A, C, G and T is shown. The affect of raising the power of the rate matrix to 2 to increase to 2CED is shown in (C), and panel (D) shows the likelihood for a set of CED values computed, the likelihood is plotted versus the CED to illustrate the maximum likelihood value which lies at the tipping point of the plateau.

1.2.4.5 Bayesian Inference

The application of Bayesian inference (BI) in phylogeny reconstruction is relatively recent and was proposed by three separate groups concurrently (Mau 1996, Li 1996, Rannala and Yang 1996). BI is based on the likelihood function, however, instead of calculating the most probable tree as is the case in ML, a Bayesian analysis searches for a set of credible trees (Huelsenbeck et al. 2002). BI has a clear advantage over maximum likelihood as it incorporates a prior hypothesis on how the data has evolved. The Bayes theorem, as applied to estimate the phylogeny of a dataset, calculates the posterior probability distribution (Pr) of the hypothesis (H_i) against the probability of observing all alternative hypotheses (H_j).

Eqn. 5 Bayes Theorem

$$\Pr[H_i|X] = \frac{\Pr[X|H_i] \times \Pr[H_i]}{\sum_{j=1}^{n} \Pr[X|H_j] \times \Pr[H_j]}$$

Bayesian inference requires all possible hypotheses to be tested and as such is extremely computationally intensive. By incorporating a Markov Chain Monte Carlo (MCMC) approach the approximation of the joint posterior probability density based on constructing a Markov chain is allowed and a more efficient exploration of parameter space is possible (Metropolis et al. 1953, Hastings 179), such as better exploration of tree-space (Huelsenbeck et al. 2001). The distribution of tree space is often described as a hill, and the MCMC approach works sampling a hypothesis by walking around this hill (Metropolis et al. 1953, Hastings 179). The MCMC starts with a hypothesis (random tree + model) and calculates the likelihood of that hypothesis. An alternative hypothesis is proposed; this can be generated through a change in parameter or change to the tree, and is compared against the previous hypothesis. If the new hypothesis "moves uphill" i.e. improves in likelihood units/fit of the model to the data, the step will automatically be taken, otherwise the new hypothesis moves down the hill and that the step is rejected and another hypothesis is proposed. Sometimes downward steps only make a small decline in fit and they are also accepted; this is an important feature of the approach as it helps to minimise the possibility of entering zones of local minima during the MCMC run.

At the start of any analysis with Bayes theorem, almost all newly suggested hypothesis are accepted, and this is referred to as the "burnin". After numerous iterations, the acceptances stabilize to a "plateau" and this region of acceptable hypotheses are used to create a consensus tree (Hall 2007), see Figure 1.9 for demonstration of "burnin". It is recommended that two or more analyses are run simultaneously so that convergence upon a topology can be assessed (Felsenstein 2004). For the first few generations the topologies sampled from each run will differ, but after a number of generations, when convergence is reached (i.e. when a good sample from the posterior probability distribution is achieved), the tree samples should be very similar. If the runs are dissimilar to one another convergence has not been achieved and this is suggestive of a local minimum issue or the chains needing to run for longer (Felsenstein 2004).

Metropolis coupling (MC) was introduced to improve mixing in the MCMC run and improve sampling of the distribution (Geyer 1992). It involves splitting each run into a number of chains, which run in parallel using MCMC. Throughout the run these chains propose to swap information exchanging some or all of their parameters. The decision to accept or reject new parameters is made by the Metropolis-Hastings algorithm. Typically, n-1 chains are heated and the remaining chain is cold, and it is from the cold chain that parameters and trees are sampled. Heating chains flattens out the posterior probability, as the heated chains will more easily find isolated peaks in the posterior distribution and can help the cold chain move more rapidly between isolated peaks (Geyer 1992). A consensus tree summarizes the information contained in the set of trees sampled from the cold chain post burnin. The methods by which a consensus tree is made are traditionally majority rule, where clades with less than 50% support are collapsed, but they can also be made using strict consensus or reduced majority rule consensus (Felsenstein 2004). The posterior probability (PP) is the measure of support given for a node on a tree based on the model, the priors and the data (Huelsenbeck and Rannala 2004). It has been reported that PP can be over inflated (Erixon et al. 2003) and this is especially true when using complex models (Huelsenbeck and Rannala 2004). Both ML and BI have been employed in this thesis for phylogeny reconstruction in Chapters 2 and 3, and for selective pressure heterogeneity analyses and estimation of codon positions under positive selection in Chapter 4. While BI can incorporate more complex models of evolution (Lartillot and Philippe 2004), it is much slower than ML and therefore experimental design and method choice is often governed by the size of the data.

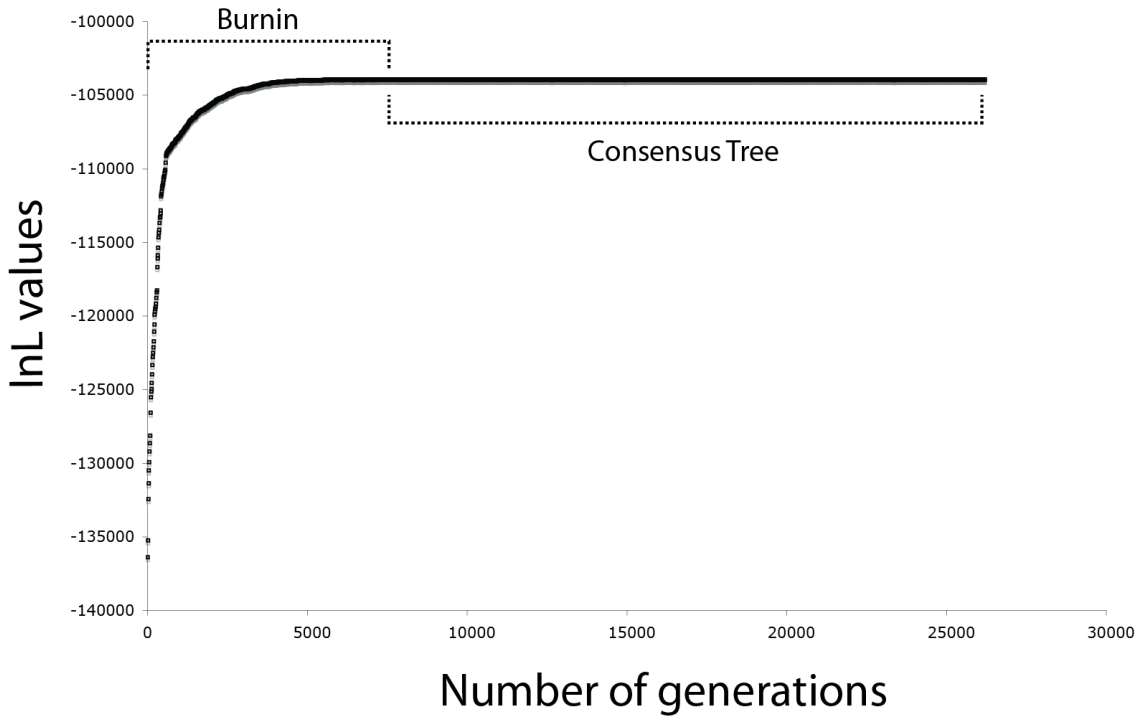


Figure 1.9 Plot of the number of generations versus the lnL values in the MCMC chain demonstrating "burn-in".

The graph demonstrates a standard MCMC chain, where the first 7,500 alternative hypotheses are discarded; these are termed "burnin". From the remaining runs and samples of the plateau - a majority rule consensus tree is built. In a standard Bayesian, analysis two separate MCMC chains are used. This ensures that local minima are not represented on the final stable state. Convergence is achieved when both MCMC chains agree across topology and parameter estimators.

1.3 Desiderata for Dataset and Method Choice

Congruence between multiple data types and phylogenetic approaches is preferential when resolving the history of a set of species and, in theory convergence should be achieved when more characters are considered (Felsenstein 1978, Felsenstein 1988). Clashes between the application of morphological characters and molecular characters (Jenner 2004) along with incongruence between Supertree and Supermatrix methods (Bininda-Emonds 2004, Song et al. 2012) are frequently observed. Not all data types are suitable for the phylogenetic question at hand and it is necessary to customise each approach based on dataset availability along with the most suitable method. In this section, the power and pitfalls of a variety of popular datasets and methods are detailed.

1.3.1 Determining the best characters for phylogeny reconstruction

Regardless of whether molecular or morphological characters are employed for phylogeny reconstruction there are a number of desirable features a character should have, these include: (i) discontinuous variation between characters states and (ii) non-reversed character states (Poe and Wiens 2000). Historically morphological characters have been applied in phylogeny reconstruction. Since Zuckerkandl and Paulings seminal paper on the use of molecular data for evolutionary analyses (Zuckerkandl and Pauling 1965), the volume of data available has increased substantially and is increasingly popular in phylogeny reconstruction (Pagel 1999). Morphological data are more prone to homoplasy than molecular data (Hedges and Maxson 1996, Wiens et al. 2003) and therefore is more likely to support the incorrect phylogeny. Even though morphological characters are more easily observed and sequencing molecular data was once expensive to generate (Hillis 1987), however, now it is very cheap with Next Generation Sequencing (NGS) technology, even for non-model organisms. Molecular datasets yield a far greater number of characters overall and therefore have more statistical power to resolve phylogenies (Hillis 1987, Donoghue and Sanderson 1992). Molecular sequence data has been used in different formats for mammal phylogeny reconstruction; nuclear data, mitochondrial data, coding sequences and non-coding sequence data have been used separately and in combination (Murphy et al. 2001a, Prasad et al. 2008).

The phylogenetic goal, i.e. resolution of shallow relationships versus deep divergences, should inform the choice of data type. It is known that mitochondrial genes and non-coding DNA sequences accumulate synonymous mutations at a faster rate than nuclear

genes and coding DNA (Brown et al. 1982, Burger et al. 2003). DNA sequence data are more prone to saturation than amino acid data (Kosiol et al. 2007). Therefore, comparative genomic analyses tend to use nuclear coding sequences (Hallstrom et al. 2007), and population studies tend to use mtDNA control regions (Rosenbaum et al. 2009).

The age of phylogenomics has arisen as a result of increased genome availability, and now features of whole genomes are being applied to phylogeny reconstruction such as gene order (Korbel et al. 2002, Sankoff et al. 1992), intron positions (Roy and Gilbert 2005), protein domain structure (Lin and Gerstein 2000, S. Yang et al. 2005) and gene content (Snel et al. 1999). Gene content and gene order methods require homology assessment and assignment but do not require multiple sequence alignment (Delsuc et al. 2005). Gene-content methods reconstruct phylogenetic trees using a distance matrix which is estimated from the proportion of shared orthologous genes between genomes (Snel et al. 1999). Gene-order methods reconstruct phylogenetic trees by minimizing the number of breakpoints between genomes or by analyzing the presence and absence of orthologous genes across genomes (Korbel et al. 2002, Blanchette et al. 1999). Rare genomic events (RGEs) such as insertions and deletions or retrotransposed elements have held much promise for phylogenomics as theory suggests they are less likely to contain homoplastic signal (Cantrell et al. 2001, van de Lagemaat et al. 2005). However, so far it seems that RGEs are too rare and stochastic error resulting from small sample sizes makes it difficult to resolve short branches (Nishihara et al. 2005), this approach also requires high quality genomic sequence data that is not always available (Philippe et al. 2005a). The use of presence or absence of microRNA data has gained popularity in recent phylogenetic studies (Dolgin 2012, Campbell et al. 2011, Rota-Stabelli et al. 2011, Lyson et al. 2012). Congruence between microRNA generated phylogenies and other data types are rare (Campbell et al. 2011, Rota-Stabelli et al. 2011), and therefore its benefit of using miroRNA in phylogeny reconstruction needs further exploration.

1.3.2 Increased taxon sampling versus increased sequence length

There are ongoing debates concerning whether an increase in taxa number or an increase in sequence data is better for phylogeny inference (Rosenberg and Kumar 2001, Hedtke et al. 2006, Zwickl and Hillis 2002, Pollock et al. 2002). A greater

sampling of taxa improves estimates of model parameters, which improves lineage sorting among clades (Hillis 1996, Hedtke et al. 2006). By increasing taxon sampling the chance of distantly related species being drawn together on a phylogeny (LBA) is reduced (Hillis 1996). The support for a phylogeny has also shown to be increased with denser taxon sampling, leading to a more robust phylogeny (Hedtke et al. 2006). The greater the number of taxa sampled, the more complex the phylogenetic analysis becomes but this is not a linear increase. The number of possible tree topologies increases with the addition of taxa by 2n-5 (where n is the number of taxa). For example, there are 3 possible unrooted topologies for a 4-taxon tree, but 2,027,025 possible unrooted topologies for a 10 taxon tree. As the number of taxa increases so too does the likelihood of encountering homoplasy (Sanderson 1989).

Rosenberg and Kumar (2001) argue the benefits of increased sequence length over taxon sampling (Rosenberg and Kumar 2001), but others have opposing views with simulation studies determining that extensive taxon sampling lead to more accurate phylogenetic inference (Zwickl and Hillis 2002, Pollock et al. 2002). In response to these simulation results, Rosenberg and Kumar conducted further simulation studies and concluded that when data is limited it is more beneficial to have more character data than more taxon data (Rosenberg and Kumar 2003). Theoretically, if enough genes are used and sequence length is substantial, phylogenetic conflict present should be overcome and the true species phylogeny should be recovered (Rokas et al. 2003). When inconsistent methods such as maximum parsimony, or models that do not adequately describe the data, are applied - an incorrect phylogeny with high bootstrap support can be found despite extensive sequence length (Delsuc et al. 2005, Felsenstein 1978, Phillips et al. 2004). Low taxon sampling is associated with a decrease in bootstrap support for the correct topology and in some cases bootstrap values can drop as the gene number increases, this is the result of taxon sampling being insufficient (Hedtke et al. 2006). In general increased taxon sampling has been shown to have a more positive effect on phylogeny reconstruction compared with increased sequence data (Hedtke et al. 2006), see Figure 1.10.

Taxa choice is also important, particularly when choosing outgroups. An outgroup that is too distantly related to the ingroups can cause rapidly evolving ingroup species to be more sensitive to the effects of LBA (Hillis 1998). To overcome this it is necessary to sample an outgroup that is closely related to the ingroup species, and also to increase

the number of outgroup species (Heath et al. 2008). In summary, phylogenetic inference from molecular data is dependant on evolutionary models that fit the data, appropriate character selection, sufficient taxon sampling and enough sequence data to obtain a robust and repeatable analyses (Hedtke et al. 2006). There were 39 taxa analysed in Chapter 2 with an alignment length of 27,220 aa and 455 taxa analysed in Chapter 3 with an alignment of 3,906 aa, making it appropriate to discuss increased taxon sampling versus increased sequence length.

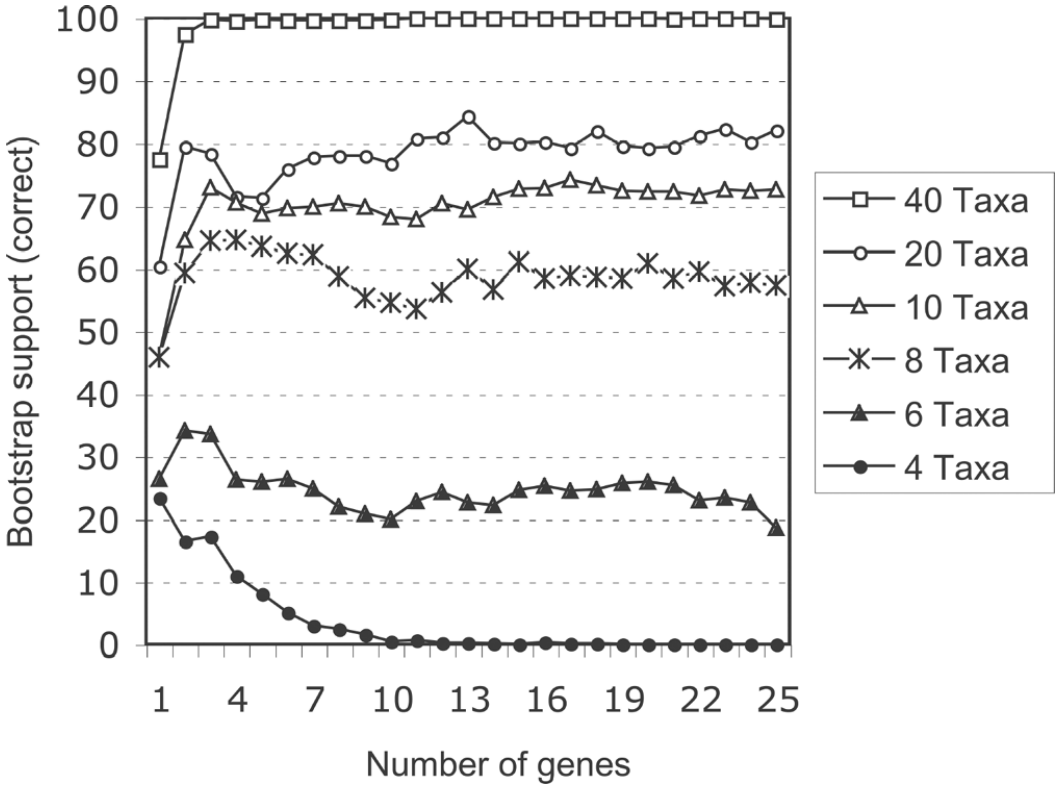


Figure 1.10 The relationship between the number of genes and taxa and the Bootstrap support obtained.

Results of the effect on bootstrap support values from randomly sampling across 1 to 25 genes (x-axis) for between 4 and 40 taxa. Figure is taken from (Hedtke et al. 2006) with permission.

1.3.3 Data assembly and methods for phylogenomics

A limiting factor of phylogenomics is the availability of sequence data, which means that taxon sampling is unavoidably limited for some studies. Large scale genomics projects have caused a shift from single gene phylogenetic studies to phylogenomic studies which encompass multiple genes helping reduce the effects of stochastic error (Heath et al. 2008). Projects such as the Genome 10K project (http://www.genome10k.org/) mean that availability of whole genome data for taxa is steadily improving. Individual gene trees can vary in both topology and branch lengths, and Supermatrix and Supertree (Sanderson et al. 1998) methods have been developed to overcome the topological variations between gene trees so that the true species phylogeny can be recovered.

1.3.3.1 The Supermatrix Approach to phylogeny reconstruction

A Supermatrix (SM) is the concatenation of alignments from several genes and can also be called the "Total Evidence" approach (Delsuc et al. 2005, Kluge 1989). Supermatrix datasets can encompass missing data from taxa, and there have been reports that this can lead to incorrect branching and low resolution between nodes (Wiens 2006, Kearney 2002). Recent studies have shown that this does not significantly affect the reconstruction of the phylogeny (Philippe et al. 2004) and as long as there is sufficient phylogenetic signal for the available characters then phylogenetic resolution is not a problem (Delsuc et al. 2005). In certain cases species specific sequence data contains large proportions of systematic error, removing these sequence regions from taxa can improve node support in phylogeny reconstruction (Grant and Kluge 2003).

A standard Supermatrix approach assumes that all characters share the same branching history, it does not take into account hybridization, incomplete lineage sorting and horizontal gene transfer, all of which can be observed at the individual gene level (de Queiroz and Gatesy 2007). Heterotachy and LBA can be accounted for with more sequence data (Philippe et al. 2004). The longer alignment length afforded by a Supermatrix allows for application of parameter rich models that accommodate lineage and dataset heterogeneity (Lartillot and Philippe 2004, Foster 2004, Foster et al. 2009).

1.3.3.2 The Supertree approach to phylogeny reconstruction

Analysing data using a Supertree approach involves generating a single phylogeny from each individual gene set and summarizing these trees (which contain partially

overlapping leaves/OTUs) into one consensus tree (Delsuc et al. 2005, Cotton and Wilkinson 2007). One of the main advantages of Supertree methods is that its reconstruction is a step removed from the sequence or morphological data and it is therefore possible to summarise results obtained from different characters (Delsuc et al. 2005). Supertree methods while taking into account individual gene histories can ignore dataset inconsistencies supported by low bootstrap values (Ren et al. 2009). Supertree methods have also been shown to report relationships for taxon groupings that are not observed in any of the fundamental gene trees. A variety of Supertree methods have been developed; matrix representation by parsimony (MRP) (Baum 2002, Ragan 1992), MinCut (Semple and Steel 2000), semi-strict Supertree (Goloboff and Pol 2002), and the most similar Supertree method (Creevey and McInerney 2005). These methods vary widely in performance and lack the ability to account for uncertainties in the fundamental tree data (Creevey and McInerney 2005, Eulenstein et al. 2004, Ren et al. 2009). In summary, the decision to use a Supertree or Supermatrix method should be based on data availability.

1.3.4 Applications of Phylogeny in Evolutionary Medicine.

The practical application of phylogeny and evolutionary theory in medicine is evident (Nesse et al. 2006). Genetic drift and natural selection have left mosaic patterns of nucleotide substitution within genomes, and these single nucleotide variations (SNPs) have been identified in thousands of disease-associated genes (Li and Agarwal 2009, Hindorff et al. 2009, Roberts et al. 2010). Understanding the association between these genetic variants and disease is the underpinning of genetic medicine, but given there are between 6000-10,000 non-synonymous SNP in a given individual (Li and Agarwal 2009, Hindorff et al. 2009, Roberts et al. 2010) making medicine personalised will be a time consuming and costly venture (Kumar et al. 2011). By taking an evolutionary medicine approach, and applying phylogeny and selective pressure variation detection methods (described in section 1.1 and 1.2), it is possible to determine functionally important disease causing SNPs in a more cost effective manner (Kumar et al. 2011). Phylogenetic methods have been employed to compare clonal subtypes of tumours and determine driver mutations that cause resistance to treatment and cause metastasis (Egan et al. 2012, Watters and McLeod 2003, Greaves and Maley 2012). In the field of infectious disease; it is possible to trace the origins of infectious strains as was carried out for the H1N1 influenza epidemic, more famously known as the “Swine Flu”

epidemic (Smith et al. 2009). Phylogeny was used to determine the source of the SARS epidemic (H5N1 influenza) and identified it as a corona virus similar to one endemic in bats (Li et al. 2004). The application of phylogeny also determined that sooty mangabeys were the source of HIV-2 in human populations (Gao et al. 1992, Gao et al. 1994) and chimpanzees were the source of HIV-1 (Gao et al. 1999). The use of phylogeny in cancer pathways has allowed identification of lineage-specific adaptations in tumour suppressor and oncogenes (O'Connell 2010, Morgan et al. 2012). Evolutionary approaches like these have been hypothesised to uncover key adaptations that cause aging and cancer (Antolin et al. 2012). The practical uses of phylogenetics in the field of evolutionary medicine, are extensive and with the advancement of *de novo* sequencing the practical applications continue to grow.

1.4 Speciation and the Diversification of Mammals

Ancient mammals arose on the Synapsid branch of the Amniote tree (Figure 1.11) after the divergence of Synapsid and Sauropsid lineages during the Paleozoic age, ~325 Million Years Ago (MYA) (Lefevre et al. 2010). Both the Cretaceous Terrestrial Revolution (KTR) ~93 MYA, and the Cretaceous-Paleogene (KPg) mass extinction event (~65 MYA), influenced mammal diversification by opening up ecospace and promoting inter-ordinal and intra-ordinal diversification (Lloyd et al. 2008, Meredith et al. 2011). This thesis examines mammals in the context of their phylogenetic relationships and their molecular adaptation. In the following section, I have briefly discussed the various modes of speciation, emergence of placental mammal Superorders and various life traits pertaining to extant placental mammals.

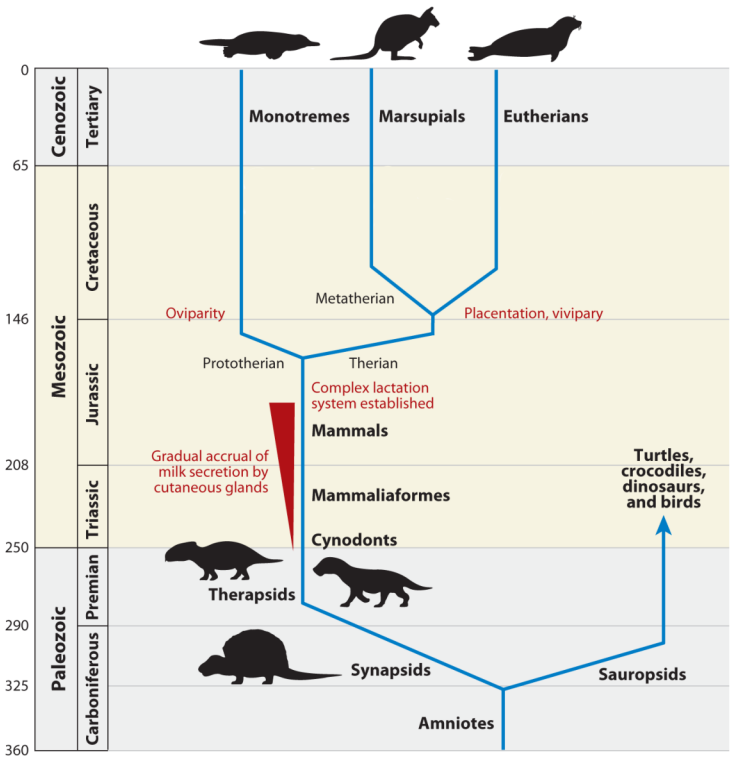


Figure 1.11 Divergence of Synapsid and Sauropsid lineages and the emergence of mammals

The dates in time from Paleozoic Era (360 MYA to 250 MYA) to the Cenozoic Era (65 MYA to present) are shown bottom to top along with the approximate diverge times of lineages leading to extant mammal groups (Monotremes, Marsupials and Eutherians). This image was taken from Lefevre et al (2010) with permission.

1.4.1 Modes of Speciation

Ernst Mayr was the first to define a species as a population, or a group of populations, whose members have the potential to interbreed with one another in nature and to produce viable fertile offspring but cannot produce viable fertile offspring with members of another species (Mayr 1970). Mayr proposed the concept of Allopatric speciation resulting from a physical barrier separating populations, such as sea level change, glacier formation or a new mountain range. This physical barrier to gene flow along with enough time for genetic drift and adaption to occur in each population independently resulted in the populations not being able to interbreed (Butlin et al. 2008, Mayr 1970). There are other modes by which speciation can arise: Peripatric, Parapatric, and Sympatric speciation (Butlin et al. 2008). Peripatric speciation is similar to allopatric speciation as it involves geographic barriers that prevent gene flow between populations, but differs as it is initiated by small population separating from the main population. This population bottleneck causes a higher rate of mutational fixations resulting in a shorter period for new species to arise. Peripatric speciation has been observed in reef hermit crabs that have been geographically isolated (Malay and Paulay 2010). Parapatric speciation, originally described by Fisher (1958), takes place when partial geographic barriers arise and sexual selection assures that no gene flow occurs between populations (Fisher 1958). Tennessee cave salamanders are an example of parapatric speciation where gene flow is impeded between surface dwelling species and cave dwelling species whose ancestors descended to cave habitats as a result of climate change (Niemiller et al. 2008). Sympatric speciation occurs in a population sharing the same habitat and is the result of strong sexual selection. The sympatric speciation as a theory is still debated as there is very little empirical support to date (Mallet et al. 2009). One of the few examples of sympatric speciation has been observed in the three-spinned sticklebacks a freshwater fish (Gow et al. 2008) where there are two populations of three-spinned sticklebacks interbreeding in the same location: (i) a large species that feeds on large prey such as snails, flat worms and tiny crustaceans, and (ii), smaller species that feed on small plankton. Evidence suggests that these species have experienced disruptive selection caused by natural selection favouring fish that were very big or very small, along with preferential mating with similar size species (Gow et al. 2008).

If reproductive isolation is not maintained between populations, the two populations can interbreed and develop hybrids (Sadava 2006). Postzygotic barriers can operate after fertilization, (i) to prevent the hybrid zygote from fully maturing, (ii) to make it less fit than the non-hybrid offspring, or (iii) in the case of a physically healthy or viable outcome - to ensure that resultant hybrid organism is infertile (Sadava 2006). Hybrid zones may form if reproductive isolation is incomplete where previously separated populations come into contact (Rieseberg et al. 2000) and can introduce genetic novelty and diversity into a gene pool (Noor et al. 2000) through a process called "genetic introgression" (Anderson 1949). Despite publications on these type of hybridization events receiving some criticism in the past (Yamamichi et al. 2012, Presgraves and Yi 2009, Currat and Excoffier 2004), scientific evidence suggests that hybrids of human ancestors occurred with literature supporting a human-chimp hybrid isolation event after the initial speciation of humans and chimps (Patterson et al. 2006) and humans and Neanderthals (Green et al. 2010).

1.4.2 Divergence time estimates and life traits of mammals

Mammals are subdivided into placental (eutherian) and non-placental (non-eutherian) clades. Non-placental mammals include Monotremes and Marsupials, whereas placental mammals consist of Xenarthra, Afrotheria, Laurasiatheria and Euarchontoglires (Asher and Helgen 2010). Dating the divergence of mammals with molecular data involves a phylogeny of known relationship, calibration of nodes using dates obtained from the fossil record, and a model of evolution (Benton et al. 2009). There are conflicts surrounding the date estimation pertaining to the emergence of mammals (Benton et al. 2009, Meredith et al. 2011, Bininda-Emonds et al. 2007, Hedges et al. 1996) and the dates given here are from the most recent publication on dating mammal nodes (Meredith et al. 2011).

The Monotremes are the least diverse of all the extant mammal orders with only two species in this group, the platypus and the echidna. Previous studies have estimated the emergence of mammals to be ~217.8 MYA (Meredith et al. 2011). Monotremes are the deepest diverging group on the mammal tree (van Rheede et al. 2006) and are exclusively found in Australia (Augee 2007). Monotremes possess the majority of mammalian features: lactation, hair, single bone in the lower jaw and endothermic temperature regulation, however their reproductive mechanisms differ substantially.

Monotremes are egg-laying mammals and secrete a high fat substance through sweat as they do not have mammary glands.

At birth Marsupials are not fully developed and they must be protected and suckled in their mothers pouch until they reach maturity. They are estimated to have split from the monotreme ancestor ~190 MYA (Meredith et al. 2011). The majority of marsupials are Australian and include kangaroos, koalas, wombats, the tasmanian devil and bandicoots. The only non-Australian mammals are opossums, that are native to America (Augee 2007).

Placental mammals are estimated to have emerged ~101.3 MYA (Meredith et al. 2011) which corresponds with the hypothesis that the KTR and KPg events contributed to the placental mammal diversification (Lloyd et al. 2008). Compressed cladogenesis, shown to have a major negative impact on the resolution of the metazoan phylogeny for example (Rokas et al. 2005), is so severe in mammals that the phylogenetic signature has proven very difficult to decipher. There is much confusion concerning the order in which the Superorders arose (Murphy et al. 2001b, Murphy et al. 2007, Prasad et al. 2008, Kriegs et al. 2006, Amrine-Madsen et al. 2003, Springer et al. 2007, Reyes et al. 2004, Nishihara et al. 2007, Tobe et al. 2010, Song et al. 2012). It is now accepted that when the supercontinents split, Laurasia (northern hemisphere) was where the Boreoutheria arose, this group subsequently split to the Euarchontoglires (including for example primates and rodents) and the Laurasiatheria (including for example cows, whales, bats). Gondwana (southern hemisphere) is understood to be where the ancestor of the Afrotheria and the Xenarthra arose but the ordering of the divergence of these major clades is still uncertain (Wildman et al. 2007). The Afrotheria Orders are; Afrosoricida (tenrecs and golden moles), Macroscelidea (elephant shrews), Tubulidentata (aardvark), Hyracoidea (hyraxes), Proboscidea (elephants) and Sirenia (dugongs and manatees). The extant members of the xenarthran superorder can only be found in the Americas. The Xenarthra are divided into (i) the Cingulata Order, which has shelled mammals such as the armadillo, and (ii), the Pilosa Order (sloths and anteaters). The Laurasiatheria are comprised of Insectivora (hedgehogs and shrews), Perissodactyla (horse and rhinoceroses), Cetartiodactyla (whales, cow, pig), Carnivora (dogs, bears, cats and seals), Chiroptera (bats) and the Pholidota (pangolin and scaly anteaters). The Euarchontoglires contain the Glires (rodents and rabbits), Primates (humans, macaques), Dermoptera (cologus or flying lemurs), Scandentia (tree shrews)

and Strepsirrhini (lemurs and tarsiers). The approximate placement of the Superorders, orders and intra orders is summarised in Figure 1.12.

Today there are ~5400 extant mammals displaying a huge range in life traits that directly impact on variation in mutational rates and nucleotide composition (Romiguier et al. 2010, Li et al. 1987). The observed heterogeneity in evolutionary rates in mammal genomes can be influenced by many factors including diet (Yang 1998), disease (Usanga and Luzzatto 1985) and intense sexual selection (Dorus et al. 2004). Variations are observed in body size, longevity, metabolic rate and germ-line generation time across mammals, and as detailed in section 1.1.6 these impact upon the rate of mutation (Leroi et al. 2003, Peto et al. 1975, Martin and Palumbi 1993, de Magalhaes and Costa 2009, Bleiweiss 1998, Caulin and Maley 2011). The variation in life traits such as body weight, longevity, metabolism and age at which sexual maturity is reached are detailed in Figure 1.13 for the mammals analysed in this thesis.

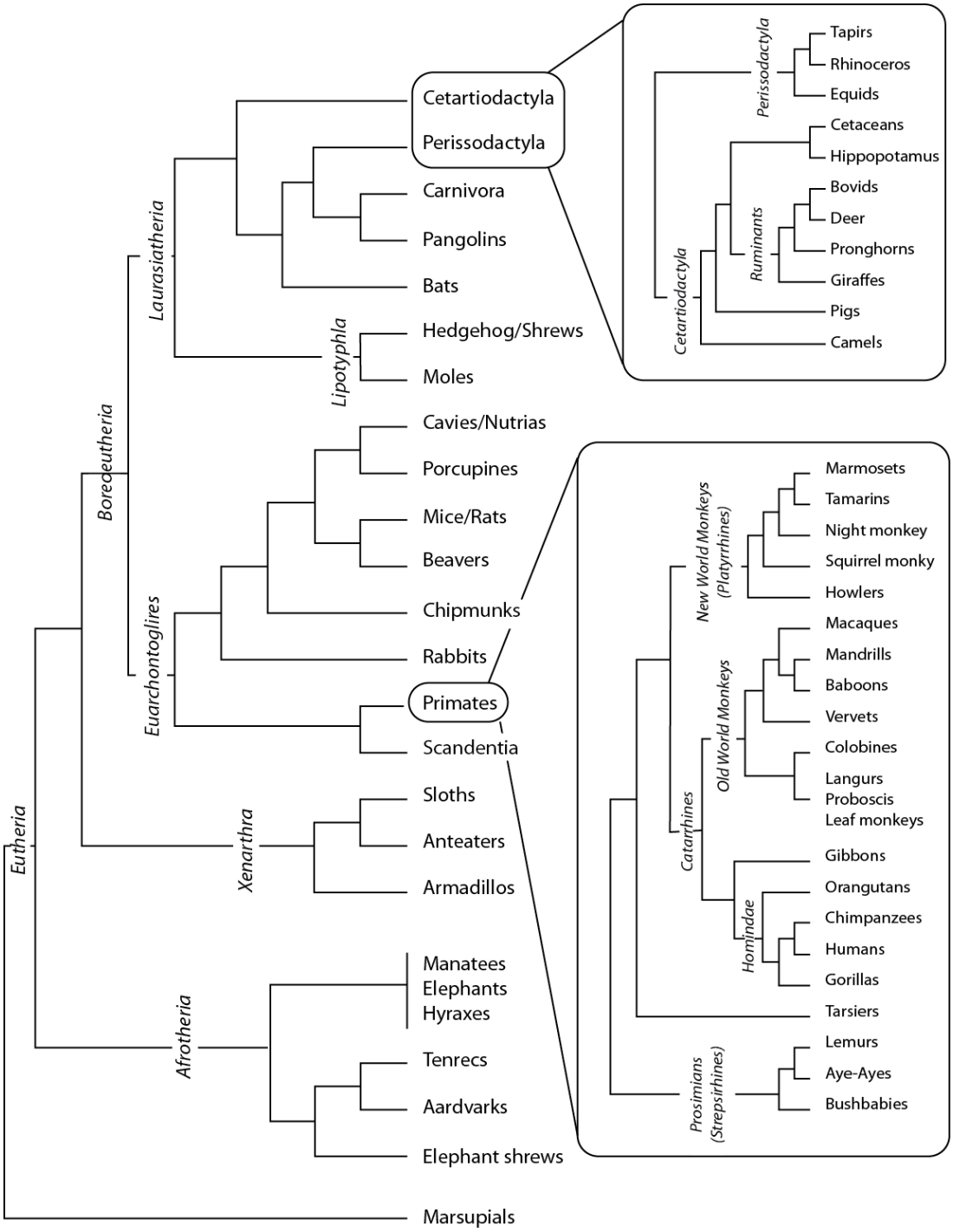


Figure 1.12 Approximate placement of all major groups in the Eutherian Phylogeny according to the Exafroplacentaila hypothesis

The nomenclature, approximate placement of orders and intra-orders is based on a study by Nishihara et al (2006). The Cetartiodactyla and the Perissodactyla clades have been expanded upon showing more detailed previously published analysis (Murphy et al. 2001a). The Primate clade has been expanded upon using a primate specific dataset (Perelman et al. 2011).

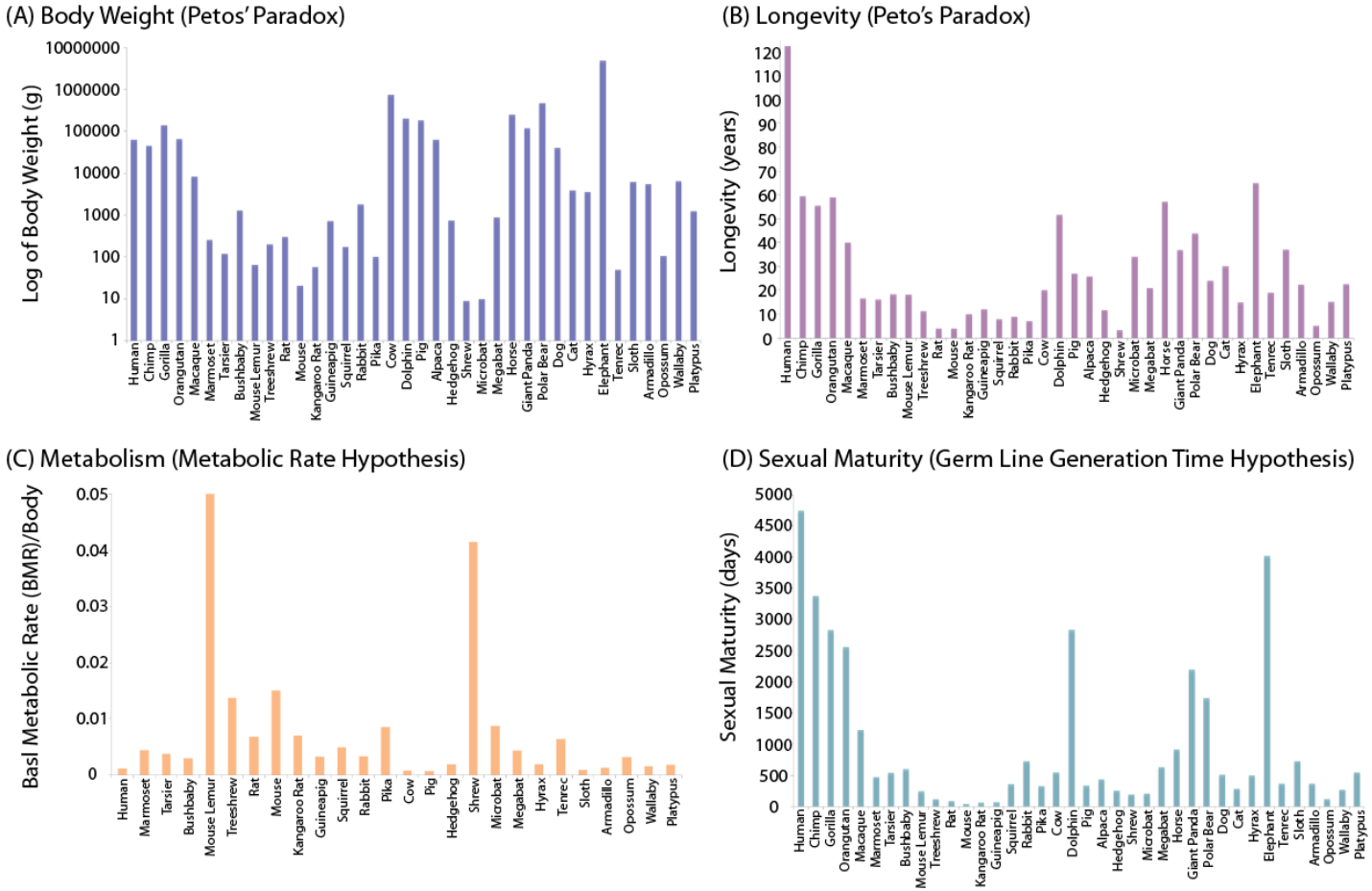


Figure 1.13 Life trait variations observed across mammals.

Variations in (A) Body weight, (B) Longevity, (C) Metabolism and (D) Sexual Maturity are shown for mammal species listed on the x-axis. Data obtained from AnAge Database (de Magalhaes and Costa 2009).

1.5 Aim of Thesis

The application of a robust mammal phylogeny to molecular evolutionary studies has prominence in the field of evolutionary medicine. In this thesis I address the fundamental issues that have prevented the reconstruction of a well-resolved mammal phylogeny and will show how this robust phylogeny is crucial for the accurate detection of adaptive evolutionary events.

The objective of Chapter 2 is to assess whether heterogeneous models were better at describing mammal data compared to homogeneous models, and most importantly, were the models generated adequate to describe both the compositional and rate exchange heterogeneity in the data. Parameter rich models require data that is suitable in size and phylogenetic informativeness to accurately estimate parameters, therefore it was necessary to assess dataset suitability and determine whether the data tested was capable of accommodating parameter rich models and if these data had enough phylogenetic information to distinguish between alternative branching hypotheses. The position of the root of the placental mammal phylogeny and the intra-order placements within the Laurasiatheria are explored in Chapter 2 using nuclear data.

To overcome the taxon deficit imposed by the use of nuclear data in Chapter 2, mitochondrial sequence data was employed in Chapter 3. Mitochondrial data has been shown to saturate faster than nuclear data and produce phylogenies that differ to those inferred with nuclear data. Previous phylogeny reconstruction efforts using mitochondrial data did not attempt to tease apart the usable phylogenetic signal from phylogenetic conflict. Therefore, in Chapter 3 I set out to test the effects of removing phylogenetic conflict from the mitochondrial data along with different partitioning of the data to improve overall phylogenetic signal and assess whether these data can be applied to the questions pertaining to the resolution of mammal phylogeny.

Finally, following these studies it was evident that model adequacy and data suitability were fundamental in phylogeny inference, but how critical is that phylogeny for selective pressure analyses using CodeML? The consequence of using a star or random phylogeny for selective pressure analyses has been explored, but what are the effects of employing a gene tree generated from data that contained phylogenetic conflict and whose composition was not adequately modelled? In Chapter 4 the implications of

applying improper phylogeny on selective pressure analyses are briefly assessed along with an analysis of the adaptive and non-adaptive evolutionary pressures acting on critical telomere interacting genes.

The overall goal of this thesis is to demonstrate the impact of data suitability and model choice in the resolution of mammal phylogeny and to apply the resolved phylogeny to the assessment of selective pressure variation in a small set of genes.

Chapter 2

2 Comparison of Heterogeneous and Homogeneous Models in the Resolution of the Placental Mammal Phylogeny

2.1 Introduction

2.1.1 Conflict in the Mammalian Phylogeny

Following a large number of studies to date using different datasets, datatypes and approaches there are now numerous competing hypotheses concerning the position of the root of the placental Superorders (Murphy et al. 2001b, Murphy et al. 2001a, Murphy et al. 2007, Prasad et al. 2008, Kriegs et al. 2006, Amrine-Madsen et al. 2003, Springer et al. 2007, Reyes et al. 2004, Nishihara et al. 2007, Tobe et al. 2010), the branching of extant placental orders (Stanhope et al. 1998, Waddell et al. 1999, Song et al. 2012) and the precise placement of many individual placental species (Janecka et al. 2007). The four major competing hypotheses for the position of the root of the placental phylogeny are illustrated in Fig. 2.1 and detailed in Table 2.1. In summary these alternative hypothesis place the following as the earliest diverging placental mammal group: (A) the Afrotheria, e.g. elephants and manatees (Afrotheria hypothesis), (B) the Xenarthra, e.g. armadillos and sloths (Epitherian hypothesis), (C) the common ancestor of Afrotheria and Xenarthra (Atlantogenata hypothesis), or (D) the Rodents, e.g. rat and mouse (Rodentia hypothesis).

It is not just the early diverging branches that are difficult to resolve; the placement of branching order in the Laurasiatheria is also contentious (Nishihara et al. 2006, Hallstrom and Janke 2008, Nikolaev et al. 2007). Within the Laurasiatheria the Chiroptera have been placed: (i) the second earliest diverging group after the Insectivora (Amrine-Madsen et al. 2003, Roca et al. 2004, Arnason et al. 2008), (ii) as a sister group beside the Perisodactyla-Cetartiodactyla clade (Asher 2007, Prasad et al. 2008), (iii) as a sister group to the Perissodactyla alone (Murphy et al. 2007, Waddell and Shelley 2003), and (iv), in the Pegasoferae position, placing them closer to the Ferae than the Cetartiodactyla (Nishihara et al. 2006). Laurasiatheria intra-order divergence dates have been estimated to be between 1 and 4 million years, which limits the time in which independent substitution's could have occurred in the respective Laurasiatherian genomes (Hallstrom and Janke 2008). These short divergence dates could make it difficult to tease apart the order in which the Laurasiatheria Orders arose. Species hybridization and incomplete lineage sorting has been suggested as a possible cause for the strong support shown in previous studies for conflicting phylogenetic hypotheses (Hallstrom and Janke 2008). Species hybridization leads to introgression

(the incorporation of genes from one species into the gene pool of another species), whereas incomplete lineage sorting produces a pattern of allele fixations from ancestral polymorphisms that does not reflect the species history. Both processes generate mixed phylogenetic signal at different loci that results in strong support for discordant mammalian relationships (Hallstrom and Janke 2008, Churakov et al. 2009). There have been numerous attempts to resolve the Laurasiatheria and competing hypotheses from previous publications have been outlined in Figure 2.2 and Table 2.1.

To date there have been many datasets applied to the resolution of the position of the placental root (Hallstrom and Janke 2010). These data have included nuclear and mitochondrial genes (with protein coding genes treated either as nucleotide or amino acid sequences), non-coding DNA, morphological data and rare genomic events (Murphy et al. 2001a, Murphy et al. 2007, Kriegs et al. 2006, Asher 2007, Hallstrom and Janke 2010). It is now clear, however, that not all datasets are equally suitable for all phylogenetic questions at all phylogenetic depths as shown most recently by Song *et al* (2012) in a reanalysis of the mammal dataset reported by Meredith et al (2011). For example the work of Brown and co-workers have shown that mitochondrial genes and non-coding DNA sequences accumulate mutations at a faster rate than nuclear genes (Brown et al. 1982), making them less useful for deep phylogenetics. In addition DNA sequences are more prone to mutational saturation than amino acid datasets (Kosiol et al. 2007), and might be more strongly affected by biases related to lineage-specific codon usage preferences (Rota-Stabelli et al. 2012).

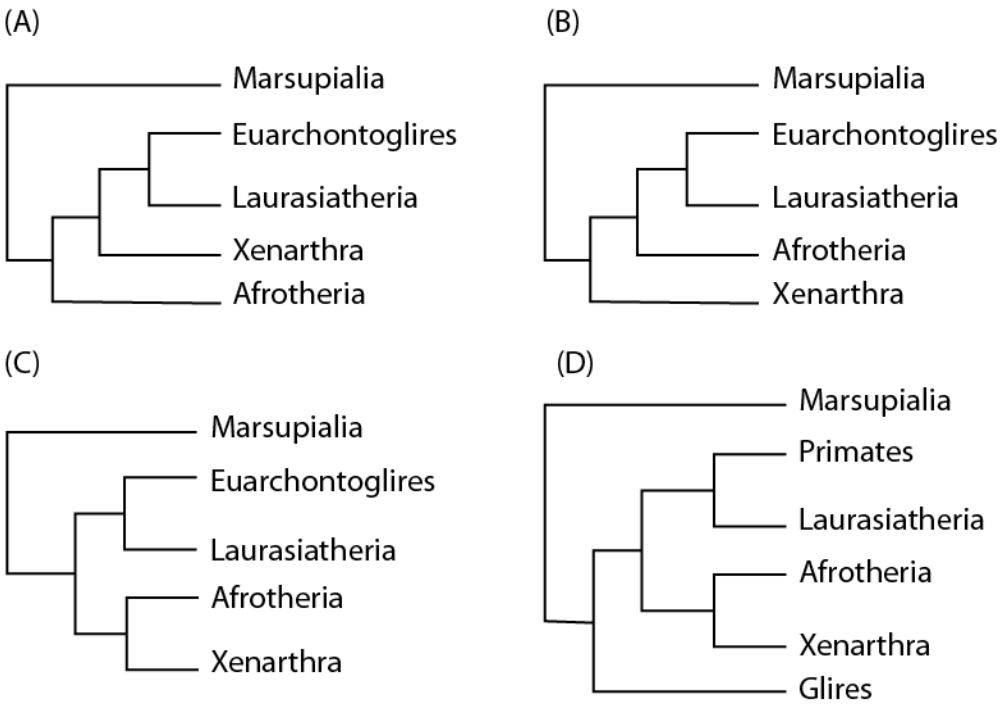


Figure 2.1 Conflict in Placental Rooting Hypotheses.

(A) Exafroplacentalia hypothesis, (B) Epitherian hypothesis, (C) Atlantogenata hypothesis, and (D) Rodentia hypothesis.

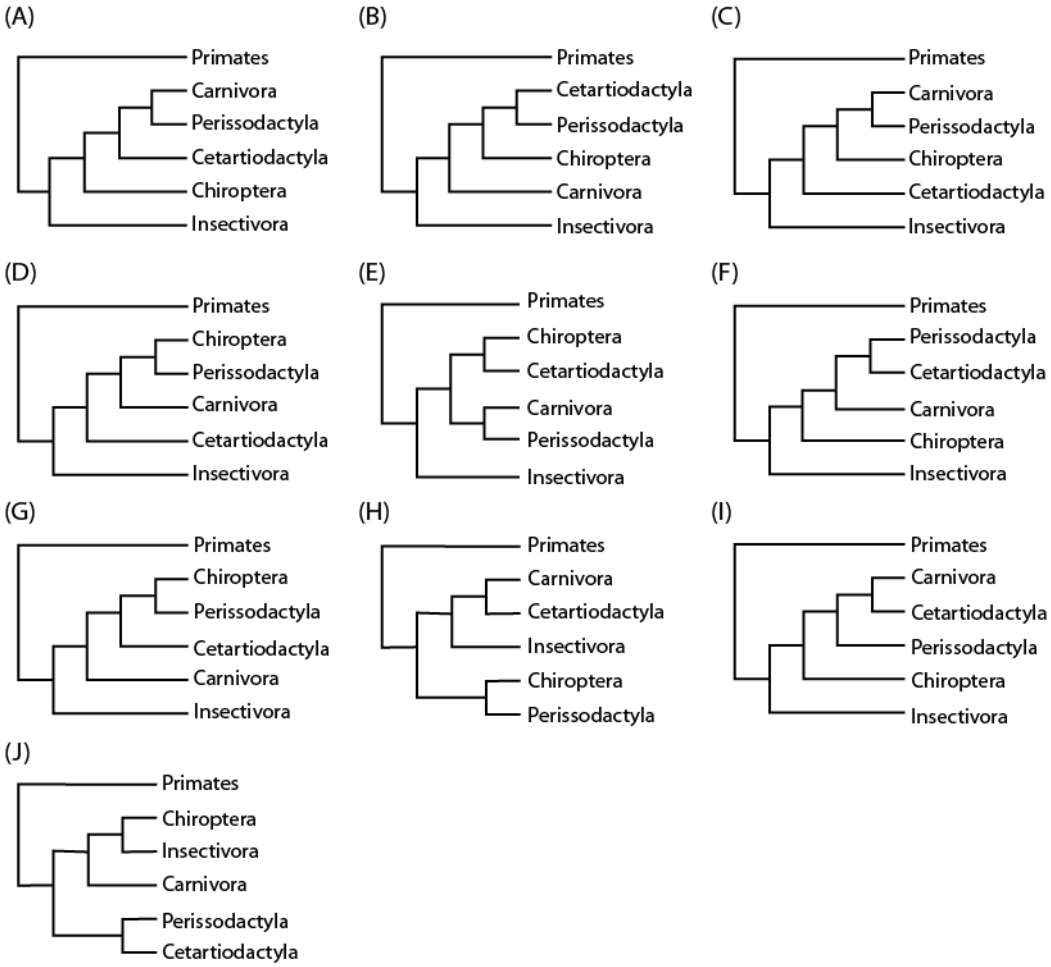


Figure 2.2 Alternative Hypotheses for Laurasiatheria Intra-order Placements

The 10 different current hypotheses describing the placement of orders within the Laurasiatheria are represented graphically in A-J.

Table 2.1 Previous Publications on the placement of the placental root and the Laurasiatheria Phylogeny

Author, Year	Placental Mammals	Data Type	Data Size	Placental Root	Laurasiatheria Intra-order
(Waddell et al. 1999)	24	mtDNA	1,240 aa	D	---
(Murphy et al. 2001a)	64	Nuclear, mtDNA	9,979 bp	A	F
(Madsen et al. 2001)	28/51	Nuclear, mtDNA	5,708 bp/2947 bp	A,C	B
(Murphy et al. 2001b)	42	Nuclear, mtDNA	16,397 bp	A	A
(Scally 2001)	47	Nuclear	2,947 bp	C	B,A
(Delsuc et al. 2002)	47	Nuclear	5,130 bp	A	G
(Springer et al. 2003)	42	Nuclear, mtDNA	16,397 bp	A	A
(Amrine-Madsen et al. 2003)	59	Nuclear	1,342 bp	A	A
(Hudelot et al. 2003)	62	mtDNA	3,571 bp	C	F
(Waddell and Shelley 2003)	81	Nuclear, mtDNA	----	A,B,C	H
(Reyes et al. 2004)	63	Nuclear, mtDNA	6, 025 bp	---	A
(Gibson et al. 2005)	62	mtDNA	2,506 bp	A	I
(Kriegs et al. 2006)	32	RGC	-----	B	---
(van Rheede et al. 2006)	9	Nuclear,RGC	10773 bp	C	---
(Nishihara et al. 2006)	26	RGC	-----	A,B,C	C
(Springer et al. 2007)	40	Nuclear, mtDNA, Morphology	14,326 bp	A	F
(Murphy et al. 2007)	43	RGC	----	---	---

Author, Year	Placental Mammals	Data Type	Data Size	Placental Root	Laurasiatheria Intra-order
(Nikolaev et al. 2007)	16	Nuclear	429,675 bp	A	---
(Hallstrom et al. 2007)	11	Nuclear	2,168,859 bp	C	---
(Wildman et al. 2007)	11	Nuclear	1,443,825 bp	C	---
(Kjer and Honeycutt 2007)	78	mtDNA	14,740 bp	C	A
(Nishihara et al. 2007)	9	Nuclear	1,011,870 bp	C	---
(Asher 2007)	49	Morphology, DNA, RGC	15,389 bp	A,B,C	B, J, D,G
(Hallstrom and Janke 2008)	19	Nuclear	2,844,615 bp	C	---
(Prasad et al. 2008)	37	Nuclear	2,154,624 bp	C	B, I
(Arnason et al. 2008)	98	mtDNA	3617 aa	C	A
(Churakov et al. 2009)	6	RGC	----	A,B,C	---
(Tobe et al. 2010)	204	mtDNA (cyt b/ COI)	1,149bp/1,557 bp	-----	F
(Hallstrom and Janke 2010)	31	Nuclear	2,863,797 bp	A,C	E
(Meredith et al. 2011)	139	Nuclear	35,603 bp	C	B
(Song et al. 2012)	33	Nuclear	1,385,220 bp	C	A

Competing hypotheses are Labelled A-D for the placental rooting and A-J for the intra-order placements of Laurasiatheria as per Figures 2.1 and 2.2.

In this Chapter, heterogeneous models are defined as those that allow for compositional heterogeneity and exchange rate heterogeneity across the topology, defined by Foster (2009) as node-discrete composition heterogeneity (NDCH) and node-discrete rate matrix heterogeneity (NDRH) (Foster et al. 2009), and those that allow for heterogeneity across the dataset (Lartillot and Philippe 2004). All previous attempts to reconstruct the mammal phylogeny have used homogeneous models that do not account for variation in composition or exchange rates across the tree or dataset (Murphy et al. 2001b, Hallstrom and Janke 2008, Prasad et al. 2008). Given the known heterogeneities in the mammal data (discussed in section 1.3.7), it is essential to use sophisticated models that can adequately describe both exchange rate heterogeneity and compositional heterogeneity across the tree and the data.

The application of more sophisticated models places a greater emphasis on the size of the dataset, e.g. alignment lengths of greater than 1000 positions (i.e. concatenated datasets) have been shown to be necessary for the profile mixture models to reliably calculate the shapes of the profiles of these data (Quang et al. 2008). Therefore, the size of the dataset applied to the problem is critical, as is the quality of the data, if heterogeneous models are to be employed.

In this Chapter I approached the problem of the resolution of the mammal phylogeny at two phylogenetic depths, (i) the position of the placental root and (ii) the order of the major groups within the Laurasiatheria. I have applied and tested the fit of models that accommodate compositional heterogeneity and exchange rate heterogeneity over the phylogeny and over the data. I have compared how these heterogeneous models fit as compared to previously employed homogeneous models. With the increase in parameterization resulting from heterogeneous models, I was also interested in testing the suitability of datasets to the phylogenetic problem.

2.2 Materials and Methods

Three separate datasets were employed in the study of the position of the mammal root (66TaxonSet and 39TaxonSet) and the placement of orders within the Laurasiatheria (14TaxonSet). The assembly of these datasets is described below.

2.2.1 Assembly of Previously Published Datasets

A description of the data used in major publications to date to infer the placental mammal phylogeny is given in Table 2.1. Six of these publications were chosen based on size and scientific impact to test whether the datasets were compositionally heterogeneous and if the models originally employed adequately describe the composition of the data. The Nikolaev *et al* (2007) data was downloaded directly from online supplementary material and no further editing was required (Nikolaev et al. 2007). Access was granted to four of the datasets through personal communication with the authors (Hallstrom and Janke 2010, Murphy et al. 2001a, Meredith et al. 2011, Prasad et al. 2008). Finally, the dataset used in the first major publication on placental mammal phylogeny by (Murphy et al. 2001a) was obtained from online supplementary material. The dataset was not immediately applicable, therefore it was copied from a word document to UNIX readable format and the sequence information was extracted using a program called "MurphyPhy2Fasta.py" (Appendix A.1.1). This program was designed to read in information from the supplementary data and export it in FASTA and was called as follows:

```
python MurphyPhy2Fasta.py ZFX.txt ZFX.fasta
```

Editing of the alignment was carried out as per online supplementary material and methods (Murphy et al. 2001a). Based on the alignment length and taxon depth of these previously published datasets the Murphy *et al* (2001a) dataset was selected for further phylogenetic exploration under heterogeneous models. This dataset contains 66 taxa and has an alignment length of 9789bp - it is referred to as "66TaxonSet" for the remainder of this Chapter. The 66TaxonSet was analysed in three tiers, first as a nucleotide dataset containing 15 loci composed of 11 coding, 1 mtDNA and 3 3'UTR sequences totalling an alignment of 9,789 bp in length (66TaxonSet_nuc). Secondly as an amino acid dataset (66TaxonSet_aa), which contains 11 protein-coding genes and 2190 aa and finally with the amino acid characters recoded to their 6 Dayhoff categories

(66TaxonSet_day), all alignments have been made available in Appendix A.1.2. The following were the commands used in P4 to generate the three tiers of data for 66TaxonSet:

```
p4>var.doCheckForAllGapColumns=0
p4>read('66TaxonSet_nuc.nex')
p4>a=var.alignments[0]
p4>a.recodeDayhoff()
p4>a.writeNexus('39TaxonSet_day.nex')
p4>quit()
```

2.2.2 Assembly of Taxon Rich-Sequence poor Dataset

The Ensembl server (version 60) was mined through BioMart and coding sequences of canonical transcripts were downloaded for all available placental mammals, see Table 2.2 for list of taxa. The Polar Bear (*Ursus maritimus*) genome, kindly supplied by the Beijing Genomics Institute (BGI), China (B. Li et al. 2011) was obtained through FTP web server using the following access commands:

```
Host=
ftp://polar_bear_project:ft5poukue4ce@ftp.genomics.org.cn/20101216/annotation20101216/genepolar_bear.20101216.cds.gz
user = polar_bear_project
password= ft5poukue4ce
```

As 39 species were identified under the criteria of this dataset, it is referred to as "39TaxonSet" from this point on. The sequences of each species were quality checked to ensure that they were protein coding by establishing that the length of a sequence was a multiple of 3 and no internal STOP codons were present. This quality check was carried out using in house program called "QualityCheckCodingSequenceFASTA.pl" designed by Thomas Walsh (Appendix A.2.1.1) and is called as follows:

```
perl QualityCheckCodingSequenceFASTA.pl human.fasta qc_failed_human.fasta qc_passed_human.fasta
```

Only sequences that passed were used in further analyses. The coding DNA sequence was translated to amino acids using, "TranslateFASTA.pl" designed by Thomas Walsh (Appendix A.1.1.2) and is called as follows:

```
perl qc_passed_human.fasta human.prot
```

2.2.2.1 39TaxonSet – Ortholog Identification

An all versus all reciprocal mpiBLAST (Darling 2003) analysis with an e-value cut-off set to e^{-6} was performed on 39 genomes. The program "BlastParser.py" (Appendix A.2.2) was written to parse the BLAST output file, take the first 39 unique hits for each query and write them to individual FASTA formatted files. The program was called as follows:

```
python BlastParser.py
```

Only files that contained 39 unique species representatives were retained, any file with >1 species indicated the presence of paralogs that would bias results. The number of instances a species occurred was counted by first generating a list of all "BlastParser.py" output files using the following UNIX commands:

```
for i in *.fasta;
do
echo "$i" | cut -d '.' -f1 >> List
done
```

Using the "List" of genes generated, the number of times each taxon was represented was counted using the following UNIX commands:

```
for i in `cat List`;
do;
grep ">alp" $i.fasta | wc -l >> alp.count;
grep ">cat" $i.fasta | wc -l >> cat.count;
grep ">chi" $i.fasta | wc -l >> chi.count;
done
```

The above commands are a sample of commands used for 3 species, however during the analysis these commands were issued over the 39 individual species. Counts of species were placed together along with their corresponding file names using the paste command in UNIX as follows:

```
paste List alp.count cat.count chi.count > AllCounts.txt
```

The file "AllCounts.txt" was then opened in EXCEL, columns were sorted and only rows where each species of the 39 had a "1" were retained. For each of the 39 species, only 27 genes were identified as having one-to-one orthology.

2.2.2.2 Alignment Generation and Editing of 39TaxonSet

The 27 SGO's were aligned using MUSCLE v3.3 (Edgar 2004). The unique Ensembl identifiers for each species were changed into 3 letter codes, listed in Table 2.2. An example of how the UNIX commands were performed to alter the names of 2 of the 20 species names is shown:

```
for i in *.aln;
do;
sed -e 's />ENSGALG0/>Chk/g' $i | sed -e 's
/>ENSPTRG0/>Chi/g' >> $i.rename
done;
```

The program "MSAmaker.py" (Appendix A.2.3) was developed to concatenate SGO's into a Supermatrix. It works by issuing the following command in a directory containing datasets with ".fasta" file extension and is called as follows:

```
python MSAmaker.py
```

Alignments were manually inspected using Se-Al (Rambaut 2001) and misaligned sequences or spurious sequences as a result of sequencing errors were removed by eye. This gave a final alignment of 27,220 aa (Appendix A.2.4.1). Invariable sites were removed, leaving an alignment of 11,039 aa. The alignment was then recoded to dayhoff categories using same method described in section 2.2.1 (Appendix A.2.4.2).

Table 2.2 39TaxonSet list of species present in the dataset

Common Name	Species Names	Genome Coverage	Species code	Genome Version (Ensembl 60)
Alpaca	*Vicugna pacos*	2.51X	Alp	vicPac1
Armadillo	*Dasypus novemcinctus*	2X	Arm	dasNov2
Bushbaby	*Otolemur garnettii*	1.5X	Bus	otoGar1
Cat	*Felis catus*	1.87X	Fca	CAT
Chicken	*Gallus gallus*	7.1X	Chk	WASHUC2
Chimpanzee	*Pan troglodytes*	8X	Chi	CHIMP2.1
Cow	*Bos taurus*	7X	Cow	Btau_4.0
Dog	*Canis familiaris*	7X	Cfa	CanFam 2.0
Dolphin	*Tursiops truncatus*	2.59X	Ttr	turTru1
Elephant	*Loxodonta africana*	7X	Laf	Loxafr3.0
Giant Panda	*Ailuropoda melanoleuca*	6.8X	Gip	ailMel1
Gorilla	*Gorilla gorilla*	35X	Gor	gorGor3
Guineapig	*Cavia porcellus*	6.79X	Cts	cavPor3
Hedgehog	*Erinaceus europaeus*	1.86X	Hed	eriEur1
Horse	*Equus caballus*	6.79X	Eca	Equ Cab 2
Human	*Homo sapiens*	Deep	Hsa	GRCh37.p2
Hyrax	*Procavia capensis*	2.19X	Pca	proCap1
Kangaroo Rat	*Dipodomys ordii*	1.85 X	Kan	dipOrd1
Macaque	*Macaca mulatta*	6.1X	Mma	MMUL 1.0
Marmoset	*Callithrix jacchus*	6X	Mar	C_jacchus3.2.1
Megabat	*Pteropus vampyrus*	2.63X	Meg	pteVam1
Microbat	*Myotis lucifugus*	1.7X	Mic	myoLuc1
Mouse Lemur	*Microcebus murinus*	1.93X	Mol	micMur1
Mouse	*Mus musculus*	7X	Mmu	NCBI m37
Opossum	*Monodelphis domestica*	7.33X	Dvi	monDom5
Orangutan	*Pongo pygmaeus*	6X	Orb	PPYG2

Common Name	Species Names	Genome Coverage	Species code	Genome Version (Ensembl 60)
Pig	*Sus scrofa*	4X	Ssr	Sscrofa9
Pika	*Ochotona princeps*	1.93X	Pik	OchPri2.0
Platypus	*Ornithorhynchus anatinus*	6X	Ply	Ornithorhynchus_anatinus-5.0
Polar Bear	*Ursus maritimus*	Not Known	Pob	Version1
Rabbit	*Oryctolagus cuniculus*	7X	Ohy	oryCun2
Rat	*Rattus norvegicus*	7X	Rno	RGSC 3.4
Shrew	*Sorex araneus*	1.9X	Sar	sorAra1
Sloth	*Choloepus hoffmanni*	2.05X	Cho	choHof1
Squirrel	*Spermophilus tridecemlineatus*	1.9X	Squ	speTri1
Tarsier	*Tarsius syrichta*	1.82X	Tsp	tarSyr1
Tenrec	*Echinops telfairi*	2X	Ete	TENREC
Treeshrew	*Tupaia belangeri*	2X	Tre	tupBel1
Wallaby	*Macropus eugenii*	2X	Meu	Meug_1.0

2.2.3 Assembly of Laurasiatheria Dataset

Placental mammals were downloaded from the Ensembl server (version 67) and mined through BioMart. Improvements were made to several genome assemblies and updated genome versions are listed in Table 2.3. As this dataset is composed of 14 taxa, it is referred to for the remainder of the Chapter as 14TaxonSet. The Orders represented in this study were; the Cetartiodactyla (alpaca, dolphin, cow and pig), the Carnivora (panda, dog and cat), the Perisddodactyla (horse), the Insectivora (hedgehog and shrew), the Chiroptera (megabat and microbat) and the Primates (chimpanzee and human) which were the outgroup species used in this study.

Table 2.3 14TaxonSet list of species present in the dataset

Common Name	Species Names	Species code	Genome Coverage	Genome Version (Ensembl 66)
Alpaca	*Vicugna pacos*	Alp	2.51X	vicPac1
Dolphin	*Tursiops truncatus*	Dol	2.59X	turTru1
Cow	*Bos taurus*	Cow	7X	UMD3.1
Pig	*Sus scrofa*	Pig	4X	Sscrofa9
Cat	*Felis catus*	Cat	1.87X	CAT
Dog	*Canis familiaris*	Dog	7X	CanFam_2.0
Panda	*Ailuropoda melanoleuca*	Pan	6.8X	ailMel1
Horse	*Equus caballus*	Hor	6.79X	EquCab2
Hedgehog	*Erinaceus europaeus*	Hed	1.86X	eriEur1
Shrew	*Sorex araneus*	Shr	1.9X	sorAra1
Megabat	*Pteropus vampyrus*	Meg	2.63X	pteVam1
Microbat	*Myotis lucifugus*	Mic	7X	myoLuc2
Human	*Homo sapien*	Hum	Deep	GRCh37.p6
Chimpanzee	*Pan troglodytes*	Chi	8x	CHIMP2.1.4

2.2.3.1 14TaxonSet - Ortholog Identification

The clustering method orthoMCL (Li et al. 2003) was applied to 14TaxonSet data to identify orthologous families, the process that the software progresses through is summarised in Figure 2.3. The program "orthomclAdjustFasta" was used to apply 3 letter species codes to each of the 14 genome files listed in Table 2.3. Poor quality proteins were removed as per default quality settings. A concatenated file containing all 14 taxa called "goodProteins.fasta" was generated from the remaining data and contained 234,628 coding sequences. Reciprocal mpiBLAST (Darling 2003) was conducted on "goodProteins.fasta" where sequence similarity cut-off values were set at e^{-6}. The results were parsed using "orthomclBlastParser" and "OrthomclLoadBlast" and the ouput file "similarSequence.txt" was loaded into MySQL database. The pairwise relationships were then computed using "orthomclPairs".

Potential relationships were defined through orthoMCL (Li et al. 2003) using (i) reciprocal best similarity pairs to define preliminary orthologs, (ii) using pairwise similarity scores that were higher within a genome compared to between genomes as a way to define preliminary paralogs and, (iii), defining co-orthologous relationships if a protein identified across two species was connected through orthology and inparology. An expectation value of e^{-5} and a percent match of 60 were applied as cut off criteria for relationship identification. Initial relationship weights were computed as the average expectation value (e^{-5}) from BLASTp (Altschul et al. 1990) for each pair of sequences and were then normalised to correct for systematic differences of orthologous and paralogous relationships so that MCL (Enright et al. 2002) clustering was unbiased (Li et al. 2003). The MCL (Enright et al. 2002) program then considers all relationships simultaneously, and separates diverged paralogs and distant orthologs that were mistakenly assigned based on (weak) reciprocal best hits and sequences with different domain structures. The inflation value for MCL (Enright et al. 2002), ranges from 1.1 (most conservative) to 10.0 (most relaxed) (Enright et al. 2002). In this analysis the inflation value was set at 1.5 as it was determined as the optimal value from clustering of eukaryotic homologs (Li et al. 2003) and 18,555 gene families were identified. The program "FASTools_v41.py" written by Andrew Webb (see Appendix A.3.1) took the gene family information from MCL output file and generated FASTA formatted files.

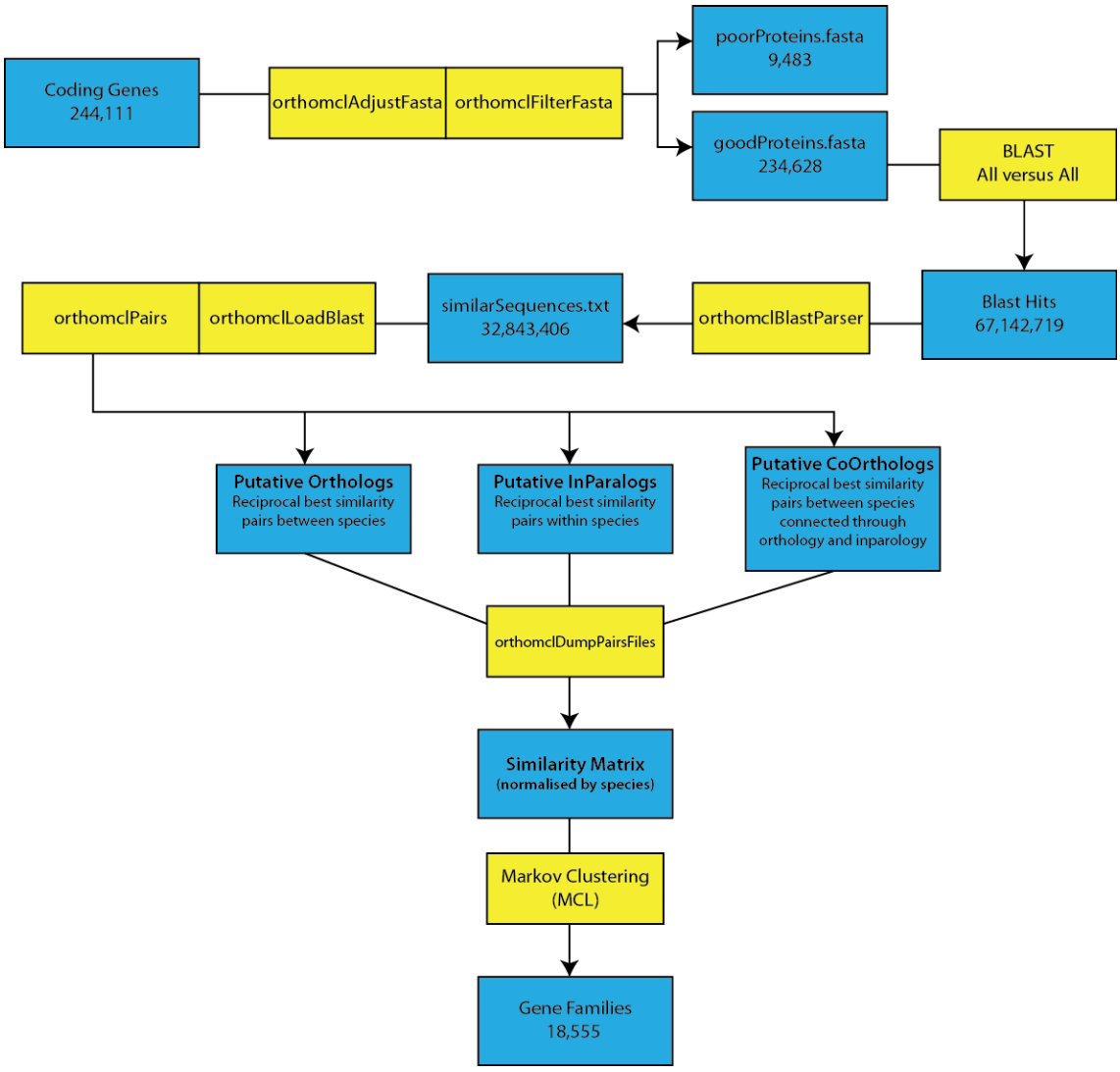


Figure 2.3 OrthoMCL pipeline for the identification of orthologous protein families

The blue boxes display dataset information and the yellow boxes show programs that were applied to the data. Adapted from Li et al (2003).

There were 18,555 homologous families identified using orthoMCL, and 11,942 of these families were SGOs that had both outgroups (human and chimpanzee) present. Of these only 2417 genes had sequences represented across all 14 taxa and these were brought forward for further testing.

2.2.3.2 14TaxonSet - Testing alternative alignment based approaches.

Previous studies have observed phylogenetic discordance as a result of mis-aligned sequence data (Morrison and Ellis 1997, Hall 2005). It was not feasible to manually check each of the 2417 genes for alignment quality; therefore a computational approach was employed. Alignments were generated using both "sequence based" and "evolutionary aware" methods. "Sequence based" methods were applied using AQUA (Muller et al. 2010), which encompasses MUSCLE (Edgar 2004) and MAFFT (Katoh and Toh 2008) alignment packages, and a refinement program called RASCAL (Thompson et al. 2003). The phylogenetically aware method, PRANK was employed with the '+F' option to account for insertion deletion events (Loytynoja and Goldman 2008). An alignment quality assessment method REFINER (Chakrabarti et al. 2006) was used to assess the quality of the resulting alignment using a norMD score (Thompson et al. 2001). The alignment with the highest norMD score was used. Where more than one alignment had an equal top score the following priority was given to alignment methods: MAFFT, MAFFT+RASCAL, MUSCE, MUSCLE+RASCAL and PRANK. The maximum norMD scores were placed into a single file using the following UNIX commands, where 'GeneList' is a list of all unique file names:

```
for i in `cat GeneList`; do
cat $i.MAFFT.norMD>> MAFFT.results
cat $i.MAFFT_RASCAL.norMD>> MAFFT_RASCAL.results
cat $i.MUSCLE.norMD>> MUSCLE.results
cat $i.MUSCLE_RASCAL.norMD>> MUSCLE_RASCAL.results
cat $i.PRANK.norMD>> PRANK.results
paste MAFFT.results MAFFT_RASCAL.results MUSCLE.results
MUSCLE_RASCAL.results PRANK.results > NorMD_alignments.txt
```

The program “Best_align.py” (Appendix A.3.2.1) was written to identify the maximum norMD scores in the “norMD_alignments.txt” file and is called as follows:

```
python Best_align.py norMD_alignments.txt > MaxnorMD_values.txt
```

The results of the maximum norMD (MAX norMD) scores for each alignment method are shown in Figure 2.9. Out of 2417 SGOs 1165 alignment methods had equal norMD scores. MAFFT had the highest norMD score for 145 SGOs, combined MAFFT+RASCAL for 107 SGOs, MUSCLE for 153 SGOs, combined MUSCLE+RASCAL for 120 SGOs and PRANK for 251 SGOs.

A program called “MapGapsFASTA.pl” (Appendix A.3.2.2) written by Thomas Walsh was used to align nucleotide FASTA files based on their corresponding protein alignments. This was carried out on the 2417 genes as follows:

```
for i in `cat GeneList`;
do;
perl MapGapsFASTA.pl $i.nuc $i.protal $i.nucal
done;
```

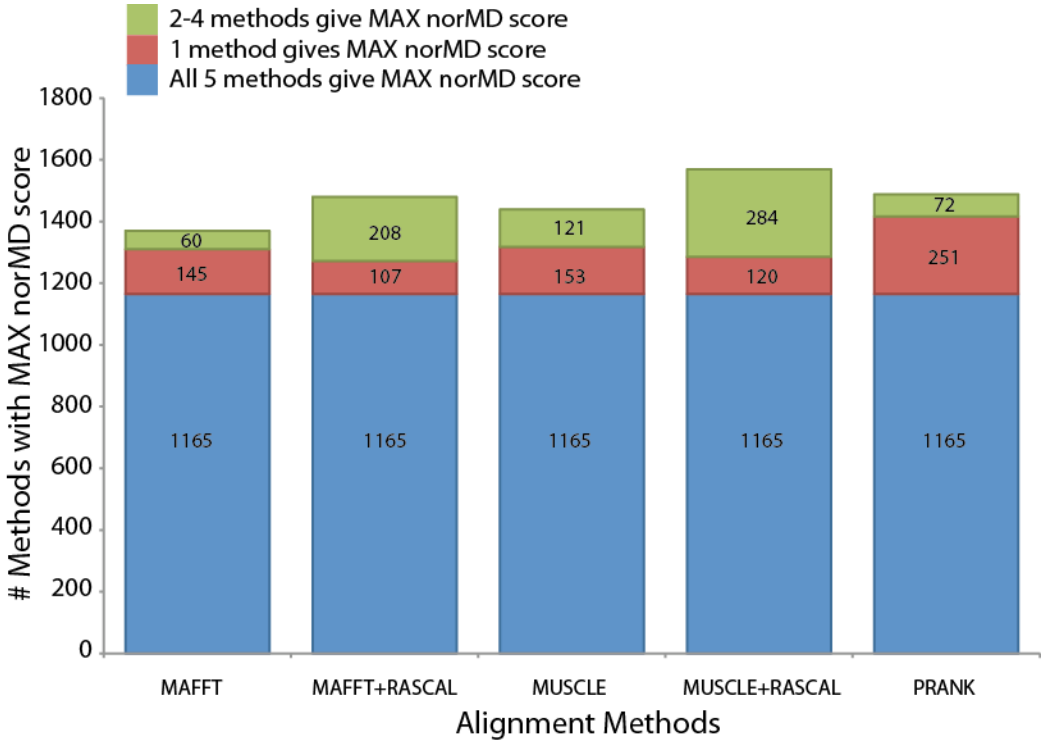


Figure 2.4 Different Alignment Methods Give Different Results

Alignment programs tested are shown on the x-axis and the number of alignments with the maximum norMDscore on the y-axis. The height of each bar represents the number of times a method had a high norMDscore. The cells are colour coded as follows: green = alignments where 2-4 program packages gave the max norMDscore; red = 1 alignment program gave the best norMDscore, blue = all 5 alignment packages resulted in the same norMDscore. The individual counts are given in each of the cells.

2.2.3.3 14TaxonSet - Alignment editing

The distribution of the mean percentage identity (Mean %ID) was calculated for each alignment using trimAl (Capella-Gutierrez et al. 2009), see Figure 2.5. Sequences that were incomplete or had high levels of mis sequenced regions were removed from the data by applying the following critera: All sequences must have at least 60% overlap with the entire MSA as well as a 0.6 minimum overlap of a position in the column to be considered a "good position". These cut offs were based on the Mean %ID, shown in Figure 2.6 and applied using trimAL (Capella-Gutierrez et al. 2009). All datasets were re-aligned using methods described in section 2.2.3.2. Only sequences that had 14 taxa were retained and following this filtering step the number of SGOs was reduced from 2417 to 1284.

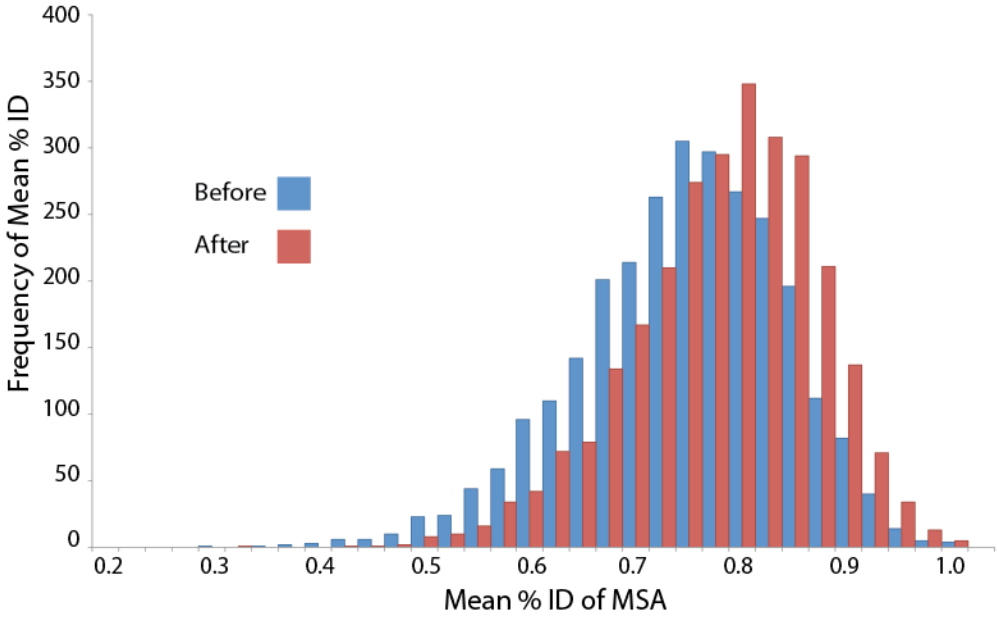


Figure 2.5 Distribution of Mean % ID of MSA's

The x-axis shows the mean % ID for each MSA and the y-axis the frequency for each mean % identity. The Blue distribution is the mean % ID score for each MSA before spurious sequences are removed and the red distribution is the mean % ID score for each MSA after spurious sequences are removed.

2.2.4 Selecting the Phylogenetic Model of Evolution.

2.2.4.1 66TaxonSet -Model of Evolution

There are currently two main approaches for modelling heterogeneity: one is to model heterogeneity across the phylogeny and the second is to model heterogeneity across the dataset. The compositional heterogeneity and exchange rate heterogeneity have been described by Foster (2009) as the NDRH and NDCH models. The NDRH model allows for exchange rates of sites to vary over the topology and the NDCH model allows for the composition to vary over the topology (Foster et al. 2009). The only model that can account for heterogeneity over the data, CAT, was developed by Lartillot and Philippe (2004) and is a probabilistic model that allows infinite mixture of site heterogeneity across the dataset (Lartillot and Philippe 2004). In this thesis, homogeneous models are defined as those that do not allow for compositional variation or exchange rate variation across the tree or dataset, for instance 1GTR+1C+I+4Γ and 1JTT+I+4Γ. The composition vector is not included in the empirical homogeneous amino acid model as it is a *Q* matrix. Substitution models were originally selected following model testing using ModelGenerator v85 (Keane et al. 2006) which tests the fit of 12 empirical amino acid instantaneous Q rate matrices and 14 nucleotide models to the dataset as well as the fit of gamma (+Γ), invariance (+I) and amino acid frequency parameters (+F) (Keane et al. 2006). Resultant models from this test were used as a starting point to generate models that incrementally increased in parameters and accommodated compositional heterogeneity and exchange rate heterogeneity across the phylogeny.

The logarithm of the marginal likelihood was calculated using the Newton Raftery equation 16 (Newton et al. 1994) in P4 (Foster 2004). Through comparison of Bayes factors (BF), the model that fits the data best is identified while simultaneously not overfitting the model to the data. Twice the difference in logarithms was calculated and then compared to the Kass and Raftery table, see Table 2.4 (Kass and Raftery 1995). If 2[lnL(Model2) – lnL(Model1)] is < 0 then Model 1 is a better fit to the data, but if 2[lnL(Model2) – lnL(Model1)] is > 2 then Model 2 is favored, see Table 2.4. The parameters used to build the composition and exchange rate heterogeneous models are shown in Table 2.5.

The CAT and CAT-GTR models (Lartillot and Philippe 2004) were compared to homogeneous models using Bayesian Cross Validation (Stone 1974) which is

implemented in PhyloBayes v3.2 (Lartillot et al. 2009). Cross Validation was calculated by splitting the dataset into two unequal parts, the learning set and the test set. The parameters of the model were estimated on the learning set, given the tree obtained from the MCMC run. The parameters were then used to calculate the likelihood of the test set. This was repeated 10 times for each model and the average of the overall lnL score was taken. The scores from each model were then compared and the top scoring model was chosen.

Table 2.4 Kass and Raftery table for Bayes Factor Comparisons.

2[lnL(Model2) – lnL(Model1)]	Evidence in favour of Model 2
<0	Negative
0-2.2	Not worth more than a bare mention
2.2-6	Positive
6-10	Strong
>10	Very Strong

Adapted from Kass and Raftery (Kass and Raftery 1995).

Table 2.5 Composition and exchange rate heterogeneous models applied to 66TaxonSet and 39TaxonSet

MODEL	RATES	COMP	+I	+Γ	PARAMETERS
66TaxonSet_nuc					
1GTR+1C+I+4Γ	1GTR(v)	1Estimate(v)	0.3(v)	0.7(v)	12
2GTR+1C+I+4Γ	2GTR(v)	1Estimate(v)	0.3(v)	0.7(v)	18
3GTR+1C+I+4Γ	3GTR(v)	1Estimate(v)	0.3(v)	0.7(v)	24
4GTR+1C+I+4Γ	4GTR(v)	1Estimate(v)	0.3(v)	0.7(v)	30
5GTR+1C+I+4Γ	5GTR(v)	1Estimate(v)	0.3(v)	0.7(v)	36
6GTR+1C+I+4Γ	6GTR(v)	1Estimate(v)	0.3(v)	0.7(v)	42
1GTR+2C+I+4Γ	1GTR(v)	2Estimate(v)	0.3(v)	0.7(v)	16
1GTR+3C+I+4Γ	1GTR(v)	3Estimate(v)	0.3(v)	0.7(v)	20
1GTR+4C+I+4Γ	1GTR(v)	4Estimate(v)	0.3(v)	0.7(v)	24
1GTR+5C+I+4Γ	1GTR(v)	5Estimate(v)	0.3(v)	0.7(v)	28
1GTR+6C+I+4Γ	1GTR(v)	6Estimate(v)	0.3(v)	0.7(v)	32
1GTR+7C+I+4Γ	1GTR(v)	7Estimate(v)	0.3(v)	0.7(v)	36
2GTR+5C+I+4Γ	2GTR(v)	5Estimate(v)	0.3(v)	0.7(v)	34
3GTR+5C+I+4Γ	3GTR(v)	5Estimate(v)	0.3(v)	0.7(v)	40
4GTR+5C+I+4Γ	4GTR(v)	5Estimate(v)	0.3(v)	0.7(v)	46
66TaxonSet_aa					
1JTT+I+4Γ	1JTT(f)	1JTT(f)	0.3(v)	0.7(v)	2
1JTT+1C+I+4Γ	1JTT(f)	1Estimate(v)	0.3(v)	0.7(v)	22
1JTT+2C+I+4Γ	1JTT(f)	2Estimate(v)	0.3(v)	0.7(v)	42
1JTT+3C+I+4Γ	1JTT(f)	3Estimate(v)	0.3(v)	0.7(v)	62
1JTT+4C+I+4Γ	1JTT(f)	4Estimate(v)	0.3(v)	0.7(v)	82
1JTT+5C+I+4Γ	1JTT(f)	5Estimate(v)	0.3(v)	0.7(v)	102
1JTT+6C+I+4Γ	1JTT(f)	6Estimate(v)	0.3(v)	0.7(v)	122
1JTT+7C+I+4Γ	1JTT(f)	7Estimate(v)	0.3(v)	0.7(v)	142
66TaxonSet_day					
1GTR+1C+I+4Γ	1GTR(v)	1Estimate(v)	0.3(v)	0.7(v)	23
2GTR+1C+I+4Γ	2GTR(v)	1Estimate(v)	0.3(v)	0.7(v)	38
3GTR+1C+I+4Γ	3GTR(v)	1Estimate(v)	0.3(v)	0.7(v)	53
4GTR+1C+I+4Γ	4GTR(v)	1Estimate(v)	0.3(v)	0.7(v)	68
5GTR+1C+I+4Γ	5GTR(v)	1Estimate(v)	0.3(v)	0.7(v)	83
6GTR+1C+I+4Γ	6GTR(v)	1Estimate(v)	0.3(v)	0.7(v)	98
1GTR+2C+I+4Γ	1GTR(v)	2Estimate(v)	0.3(v)	0.7(v)	29
1GTR+3C+I+4Γ	1GTR(v)	3Estimate(v)	0.3(v)	0.7(v)	35
1GTR+4C+I+4Γ	1GTR(v)	4Estimate(v)	0.3(v)	0.7(v)	41

MODEL	RATES	COMP	+I	+Γ	PARAMETERS
1GTR+5C+I+4Γ	1GTR(v)	5Estimate(v)	0.3(v)	0.7(v)	47
1GTR+6C+I+4Γ	1GTR(v)	6Estimate(v)	0.3(v)	0.7(v)	53
1GTR+7C+I+4Γ	1GTR(v)	7Estimate(v)	0.3(v)	0.7(v)	59
5GTR+2C+I+4Γ	5GTR(v)	2Estimate(v)	0.3(v)	0.7(v)	89
5GTR+3C+I+4Γ	5GTR(v)	3Estimate(v)	0.3(v)	0.7(v)	95
5GTR+4C+I+4Γ	5GTR(v)	4Estimate(v)	0.3(v)	0.7(v)	101
39TaxonSet_aa					
1JTT+4Γ	1JTT(f)	1Estimate(f)	0.0(f)	0.62(f)	0
1JTT+1C+4Γ	1JTT(f)	1Estimate(v)	0.0(f)	0.62(f)	20
1JTT+2C+4Γ	1JTT(f)	2Estimate(v)	0.0(f)	0.62(f)	40
1JTT+3C+4Γ	1JTT(f)	3Estimate(v)	0.0(f)	0.62(f)	60
1JTT+4C+4Γ	1JTT(f)	4Estimate(v)	0.0(f)	0.62(f)	80
1JTT+5C+4Γ	1JTT(f)	5Estimate(v)	0.0(f)	0.62(f)	100
1JTT+6C+4Γ	1JTT(f)	6Estimate(v)	0.0(f)	0.62(f)	120
1JTT+7C+4Γ	1JTT(f)	7Estimate(v)	0.0(f)	0.62(f)	140
1JTT+8C+4Γ	1JTT(f)	8Estimate(v)	0.0(f)	0.62(f)	160
39TaxonSet_day					
1GTR+1C+4Γ	1GTR(v)	1Estimate(v)	0.0(f)	0.62(f)	21
2GTR+1C+4Γ	2GTR(v)	1Estimate(v)	0.0(f)	0.62(f)	36
3GTR+1C+4Γ	3GTR(v)	1Estimate(v)	0.0(f)	0.62(f)	51
4GTR+1C+4Γ	4GTR(v)	1Estimate(v)	0.0(f)	0.62(f)	66
5GTR+1C+4Γ	5GTR(v)	1Estimate(v)	0.0(f)	0.62(f)	81
1GTR+2C+4Γ	1GTR(v)	2Estimate(v)	0.0(f)	0.62(f)	27
1GTR+3C+4Γ	1GTR(v)	3Estimate(v)	0.0(f)	0.62(f)	33
1GTR+4C+4Γ	1GTR(v)	4Estimate(v)	0.0(f)	0.62(f)	39
1GTR+5C+4Γ	1GTR(v)	5Estimate(v)	0.0(f)	0.62(f)	45
2GTR+4C+4Γ	2GTR(v)	4Estimate(v)	0.0(f)	0.62(f)	53
3GTR+4C+4Γ	3GTR(v)	4Estimate(v)	0.0(f)	0.62(f)	69
4GTR+4C+4Γ	4GTR(v)	4Estimate(v)	0.0(f)	0.62(f)	84
5GTR+4C+4Γ	5GTR(v)	4Estimate(v)	0.0(f)	0.62(f)	99

The model codes used in this study; the number of GTR rate matrices applied to the data; the number of composition vectors estimated; the proportion of invariable sites (+I); the gamma distributed associated rate variation (+Γ), and the number of free parameters estimated for each model are given. V: parameter is free to vary, F: parameter is fixed.

2.2.4.2 39TaxonSet -Model of Evolution

Models of evolution were calculated as described in section 2.2.4.1. Invariant sites were removed and the estimate for +Γ value was optimized in P4 (Foster 2004) and was subsequently fixed for the remainder of the analyses. This reduced the computational requirement by removing the optimization of unnecessary parameters and also improved the speed of the analyses. Names and parameters of the compositional and exchange rate heterogeneous models applied to dataset are listed in Table 2.5.

2.2.4.3 14TaxonSet -Model of Evolution

Nucleotide and amino acid evolutionary models were calculated using ModelGenerator v85 (Keane et al. 2006), as described in section 2.2.4.1.

2.2.5 Testing for Compositional Homogeneity

2.2.5.1 Chi-squared (χ^2) test

All datasets were tested for compositional homogeneity using the χ^2 test. The χ^2 test calculated the distribution of expected base frequencies (Exp) from these data and tested whether the observed base frequencies (Obs) fell within the data (Eqn. 6). This test does not take phylogenetic relationships into account and therefore was unable to accommodate lineages that deviated from the average compositional distribution (Rzhetsky and Nei 1995).

Eqn. 6 Chi-Squared Equation

$$\chi^2 = \sum \frac{(\text{Obs-Exp})^2}{\text{Exp}}$$

2.2.5.2 Model fit test

Tree- and model-based composition fit tests were applied to all datasets with the exclusion of the individual gene sets in 14TaxonSet. This test is known as the model fit test and was devised by Foster et al. (Foster 2004). The model fit test is similar to the χ^2 test except that the Exp comes from the model, not the data. One hundred datasets were simulated on the tested model and phylogenetic tree and χ^2_m was sampled (where m is the expected value from the model). The goodness-of-fit measure, or χ^2-value, from the actual data was then compared to the simulated data (χ^2_m). The model parameters and

branch lengths were optimized in a ML framework, and the data was considered homogeneous if the calculated value is greater than 95% of the null distribution (Foster 2004).

2.2.5.3 Chi-squared test specific to Tree Puzzle

The χ^2 test implemented in Tree-Puzzle (Schmidt et al. 2002) compared the composition of each sequence to the frequency distribution that was assumed in the ML model and labelled the compositionally heterogeneous species that did not fall within the 95% null distribution. This test was less computationally demanding than the model fit test in P4 (Foster 2004) therefore it was better able to accommodate the 1284 nucleotide and 1284 amino acid alignments from the 14TaxonSet.

2.2.6 Phylogeny reconstruction Methods

2.2.6.1 Phylogenetic Analysis using MrBayes

Phylogeny reconstruction of the 14TaxonSet was carried out using hybrid MrBayes v.3.1.2h (Huelsenbeck and Ronquist 2001). Two independent MCMC chains were run for 1.5 million generations sampling every 40 generations. The first 375,000 trees (25%) were discarded as "burnin" and the remaining 75% was used to construct a majority rule (MR) consensus tree. Convergence was determined by analysing the average standard deviation of split frequencies. If the average standard deviation of split frequencies between MCMC runs remained below 0.01, then it was an indication of convergence being reached. The nexus blocks used for the 14TaxonSet are available in Appendix A.4.1. The trees from 14TaxonSet are available in appendix A.4.2.

2.2.6.2 Phylogenetic Analysis using P4

Phylogenetic searches using models that accommodate heterogeneity over the phylogeny were performed in P4 (Foster 2004) on 66TaxonSet and 39TaxonSet. For each of the heterogeneous models tested, 10 independent MCMC runs were used and these were permitted to run long after the likelihood values of the chain converged. An example of the sMcmc.py file for the 3GTR+5C+I+4Γ model is shown in appendix A.4.3. Only runs that had converged were used in further analyses, see Figure 2.6 for example of run. Convergence was determined by finding agreement between topologies generated from the independent MCMC runs that were in agreement with one another.

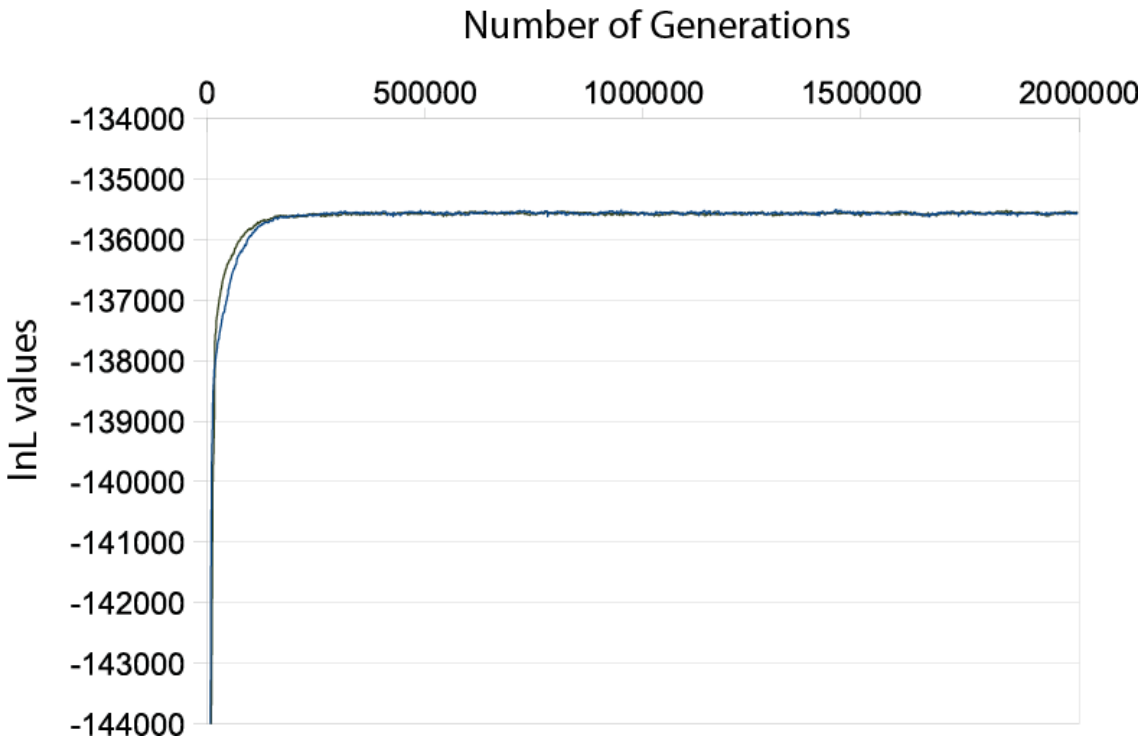


Figure 2.6 Two independent MCMC chains for the 3GTR+5C+I+4Γ model on 66TaxonSet_nuc

The plot of lnL values (y-axis) shown for two independent MCMC chains that were run for 2,000,000 generations (x-axis) for the 3GTR+5C+I+4Γ model on the 66TaxonSet_nuc dataset.

2.2.6.3 Phylogenetic Analysis using PhyloBayes

Phylogenetic searches for 66TaxonSet and 39TaxonSet were performed using heterogeneous models CAT and CAT-GTR (Lartillot and Philippe 2004) along with homogeneous JTT+4Γ and GTR+1C+4Γ (+I parameter added for 66TaxonSet) in PhyloBayes v3.2 (Lartillot et al. 2009). Two separate MCMC runs were performed each with four MCMC chains (1 cold, 3 hot) and sampling from the cold chain every 100 trees until convergence was reached. Convergence was assessed by observing the maximum difference (maxdiff) of splits between chains as per the literature; if the maxdiff < 0.1 it was a good run, if the maxdiff < 0.3 it was an acceptable run and if the maxdiff was between 0.3-1 then the sample was not long enough and the MCMC chain needed to keeping running. If the maxdiff was > 1 then this indicated that at least one of the runs was stuck at a local maximum. Phylogenies obtained from dataset heterogeneous models are provided in appendix A.4.4.

2.2.7 Posterior Predicative Simulations

Each model was tested for goodness-of-fit using posterior predictive simulations (Foster 2004). These tests were applied to the models that accommodate compositional heterogeneity across the phylogeny in P4 (Foster 2004) and models that accommodate heterogeneity across the data in PhyloBayes v3.2 (Lartillot et al. 2009). Using the given model for a dataset and the parameters estimated from the MCMC run for that dataset, simulated datasets were created. The fit of the model was estimated by the tail-area probability (Gelman et al. 1995). If the real data fell within the distribution of the simulated dataset then the composition of the data was well described by the given model. Using posterior predictive simulations in PhyloBayes v3.2 (Lartillot et al. 2009) an assessment was made of the compositional deviation for each taxon and the overall fit of the model to the composition. The output for posterior predictive simulations in PhyloBayes v3.2 (Lartillot et al. 2009) were computed and given in the form of Z-scores. A Z-score > 2 indicated that the model has failed to fit the data.

2.2.8 Likelihood Mapping

To determine if there was adequate phylogenetic signal within each dataset, likelihood mapping (LM) was performed using TreePuzzle v5.2 (Schmidt et al. 2002). The

following is a brief description on how LM works: Each tree is broken into quartets and support for each quartet was assessed. If a strong signal was present in these data then all three possible relationships for the quartets were equally likely. For ease of interpretation this is generally displayed visually on a triangle by placing the signal at the vertices; see Figure 2.7(B). If there was conflict in the signal among the quartets then the quartet will either be in the centre, which means there was a star phylogeny, or at the edges resulting in a network, see Figure 2.7(A). If 10% or more of the quartets were placed in the regions numbered 4-7 of Figure 2.7(C), the dataset is discarded because of having inadequate phylogenetic signal.

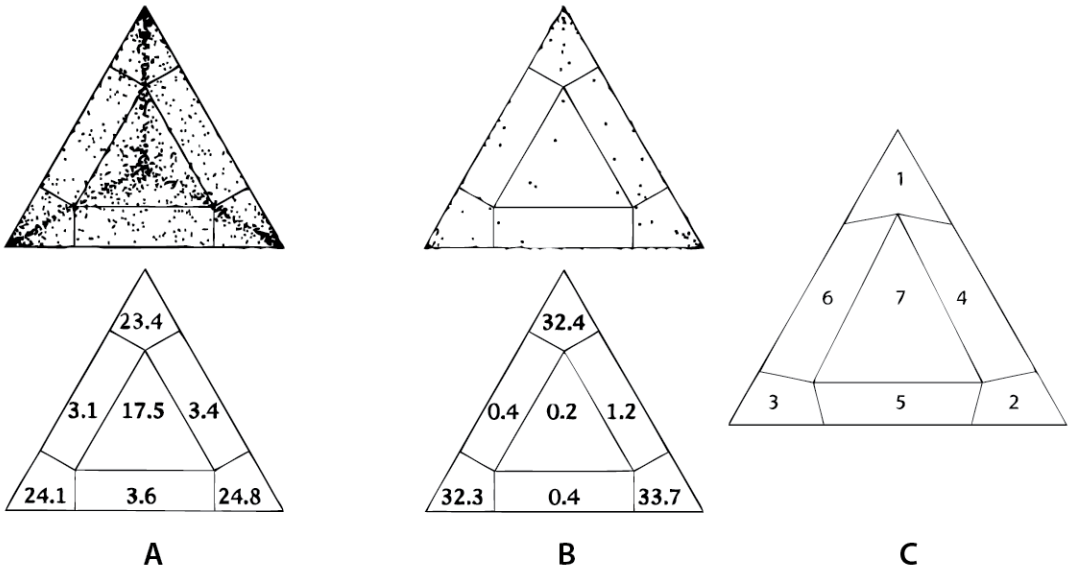


Figure 2.7 Example of Likelihood Mapping Results

Likelihood-mapping results for two biological datasets, (A) shows poor phylogenetic signal, and (B) shows good phylogenetic signal. The upper triangles correspond to the triangles A and B and show the same information but in two possible ways, the top triangle shows the distribution pattern of the individual quartets represented by dots and the lower triangles show the percentage of quartets that belong in each of the 7 sections of the triangle. Figure C shows the numbering of the individual quartets. This image was adapted from (Strimmer and von Haeseler 1997), with permission.

2.3 Results

Three main datasets have been used in this study; 66TaxonSet, 39TaxonSet and 14TaxonSet, all of which are detailed in Table 2.6. Datasets 39TaxonSet and 14TaxonSet were designed and assembled to address specific phylogenetic questions relating to the position of the root of placental mammals and the ordering of the groups within the Laurasiatheria respectively. 66TaxonSet was assembled to test the quality and suitability of data previously applied to the resolution of the mammal phylogeny. The results for each dataset are detailed in sections 2.2.1 to 2.2.3.

Table 2.6 Summary of Data Sets Analysed.

Datasets	Dataset Type	MSA Type	Alignment Length
66TaxonSet_nuc	Nucleotide	Supermatrix	9,789 bp
66TaxonSet_aa	Amino Acid	Supermatrix	2,190 aa
66TaxonSet_day	Dayhoff Recoded 6	Supermatrix	2,190 characters
39TaxonSet_aa	Amino Acid	Supermatrix	23,900 aa
39TaxonSet_day	Dayhoff Recoded 6	Supermatrix	23,900 characters
14TaxonSet_nuc	Nucleotide	1284 individual genes	
14TaxonSet_aa	Amino Acid	1284 individual genes	

2.3.1 Phylogeny reconstruction using homogeneous models

2.3.1.1 Test for Compositional Homogeneity on Previously Published Datasets

All previous studies on the placental mammal phylogeny have used homogeneous models of evolution that have not accounted for compositional and exchange rate heterogeneity across the tree and the data, see Table 2.1. Six of these previous publications were analysed in a likelihood framework to test if the composition of their datasets were homogeneous under the χ^2 test of homogeneity. The model fit test (Foster 2004) was performed to see if the models employed in the original publications described the composition of the data. The results of these tests are shown in Table 2.7

It was not possible to model all the data using the model fit test (Foster 2004). The Hallstrom and Janke (2010) dataset is the largest phylogenomic dataset for placental mammals published to date emcompassing 2,863,797 bp and 31 taxa. This dataset could not be tested under the model fit test (Foster 2004) as P4 was unable to accommodate the large alignment length. The recently published Meredith *et al*., 2011 dataset contained 169 taxa and had topologies inferred using amino acid (11,010 aa) and nucleotide datasets (35,603 bp) (Meredith et al. 2011). A model fit test (Foster 2004) was attempted on Meredith's data, however simulation of the homogeneous models (JTT+4Γ & GTR+1C+4Γ) over both amino acid and nucleotide dataset resulted in a decrease in log likelihood scores during optimisation of branch lengths so the model fit test (Foster 2004) could not be completed. As a result it was consequently not possible to test for compositionally deviating taxa in the Meridith's amino acid or nucleotide datasets. Previously published data were either not well described by the homogeneous model of evolution according to the χ^2 test of homogeneity or individual taxa were not described well by the composition of the model, according to the model fit test. Therefore, it was evident that more sophisticated models that account for heterogeneity in rate exchange and in composition over these data and the phylogeny were needed.

Table 2.7 Testing for Compositional Heterogeneity in 20TaxonSet and Previously Published Datasets

Dataset	Sequence Data	Evolutionary Model	Data size	Taxa #	Chi-Squared	Model Fit Test	
						Overall	Taxa P < 0.05
66TaxonSet (Murphy et al. 2001a)	Coding and Non Coding DNA	GTR+4Γ	9,979 bp	66	P=0.000000	P=1.000	13
(Murphy et al. 2001b)	Coding and Non Coding DNA	GTR+I+4Γ	16,397 bp	42	P=0.000000	P=1.000	2
(Nikolaev et al. 2007)	Coding DNA	GTR +I+4Γ	204,786 bp	18	P=0.000000	P=1.000	8
	Non Coding DNA	GTR+I+4Γ	429,675 bp	18	P=0.000000	P=1.000	7
(Prasad et al. 2008)	Coding DNA	GTR+I+4Γ	21,510 bp	37	P=0.000000	P=1.000	29
	Non Coding DNA	GTR+I+4Γ	132,423 bp	37	P=0.000000	P=1.000	0
(Hallstrom and Janke 2010)	Coding DNA	GTR+I+4Γ	2,863,797 bp	31	P=0.000000	NA	NA
(Meredith et al. 2011)	Amino Acids	JTT+4Γ	11,010 aa	169	P=1.000000	NA	NA
	Coding DNA	GTR+4Γ	35,603 bp	169	P=0.000000	NA	NA

NA is where chi-squared calculations were not possible.

2.3.2 66TaxonSet – A re-analysis of a previously published dataset

The purpose of this analysis was to determine if heterogeneous models that allow for compositional heterogeneity and exchange rate heterogeneity across the phylogeny and the data were a better fit to data than homogeneous models. The Murphy *et al* (2001a) dataset, which was renamed to "66TaxonSet" was chosen for indepth study above all other previously published datasets as it was one of the most highly cited mammal phylogeny studies and was a suitable size for analysis using P4 (Foster 2004). The 66TaxonSet was also tested to see if heterogeneous models were adequate in describing the composition of the dataset, and if the 66TaxonSet had enough information to conclude upon one topology and reject competing hypotheses. The dataset was analysed at three levels, first at the nucleotide level (66TaxonSet_nuc), then at the amino acids (66TaxonSet_aa) level. It was not possible to test NDRH models on amino acid data as each additional rate matrix had 189 free parameters and therefore made the analysis computationally prohibitive. Instead, the dataset was recoded into 6 Dayhoff categories (66TaxonSet_day), thereby allowing testing of NDRH models, see Table 2.6. The 66TaxonSet_nuc and 66TaxonSet_aa showed compositional heterogeneity where 13/66 and 47/66 taxa respectively did not fit the homogenous model (P values < 0.05, not corrected for multiple comparisons), while the 66TaxonSet_day was compositionally homogeneous. These results support the need for models that are able to accommodate compositional heterogeneity over the tree and the data for 66TaxonSet.

2.3.2.1 Testing the fit of Heterogeneous Models to 66TaxonSet

The next step was to investigate if models that account for heterogeneity over the phylogeny and data were a better fit to mammal datasets than previously employed homogeneous models. First models that account for compositional heterogeneity and exchange rate heterogeneity across the phylogeny were applied to 66TaxonSet. The best fitting homogeneous model was determined and was compared against heterogeneous models with increasing numbers of rate matrices and compositional vectors (i.e. NDRH and NDCH models). The best fitting homogeneous model was identified using ModelGenerator v85 (Keane et al. 2006), and for 66TaxonSet_nuc it was GTR+1C+I+4Γ, and for 66TaxonSet_aa it was JTT+I+4Γ (Jones et al. 1992). Tree heterogeneous models were evaluated by progressively increasing the number of rate exchange matrices and composition vectors (or both) until there was no further

improvement from increased parameterisation. This was determined through Bayes Factor (BF) comparison, where the criterion was that the difference between the fit of the two models to the dataset must have 2ln(BF) > 6 (see Table 2.4 for the list of models used to analyse 66TaxonSet).

BF analysis indicated that there was strong evidence in favouring the heterogeneous model that contained 3 GTR matrices, 5 composition vectors, an invariable sites category, and 4 categories of gamma distributed associated rate variation, i.e. 3GTR+5C+I+4Γ over the homogeneous model (1GTR+1C+I+4Γ), 2ln(BF) = 1097.1. The largest change in BF scores was observed in the shift between the compositionally homogeneous model (1GTR+1C+I+4Γ) and the model that allows for compositional heterogeneity across the tree (1GTR+2C+I+4Γ) with 2ln(BF) = 727.7. All BF comparisons are detailed in Table 2.8. Posterior predictive simulations on 66TaxonSet_nuc showed that homogeneous models did not fit the data, see Figure. 2.8(A), while the best fitting model that accounted for heterogeneity across the phylogeny (3GTR+5C+I+4Γ) did fit the data, see Figure 2.8(B). This was significant as it demonstrated that the model employed in the original publication of the 66TaxonSet did not adequately describe the data (Murphy et al. 2001a).

Table 2.8 Comparing the fit of tree heterogeneous models applied to 66TaxonSet.

66TaxonSet_nuc														
	1GTR+1C+I+4Γ													
2GTR+1C+I+4Γ	113.77	**2GTR+1C+I+4Γ**												
3GTR+1C+I+4Γ	176.77	57.35	**3GTR+1C+I+4Γ**											
4GTR+1C+I+4Γ	201.56	82.15	14.53	**4GTR +1C+I+4Γ**										
5GTR+1C+I+4Γ	216.36	96.95	29.33	0.95	**5GTR +1C+I+4Γ**									
6GTR+1C+I+4Γ	232.85	113.44	45.82	17.44	-7.61	**6GTR +1C+I+4Γ**								
1GTR+2C+I+4Γ	727.73	608.32	540.7	512.32	487.27	481.43	**1GTTR+2C+I+4Γ**							
1GTR+3C+I+4Γ	879.13	759.71	692.09	663.72	638.66	632.82	140.91	**1GTR +3C+I+4Γ**						
1GTR+4C+I+4Γ	904.72	785.31	717.69	689.31	664.26	658.42	166.51	17.74	**1GTR+4C+I+4Γ**					
1GTR+5C+I+4Γ	937.86	818.45	750.82	722.45	697.4	691.55	199.64	50.87	18.26	**1GTR+5C+I+4Γ**				
1GTR+6C+I+4Γ	950.55	831.14	717.69	735.15	710.09	704.25	212.34	63.57	30.95	5.08	**1GTR+6C+I+4Γ**			
1GTR+7C+I+4Γ	956.41	837	769.38	741	715.95	710.1	212.34	69.43	36.81	10.94	-17.98	**1GTR+7C+I+4Γ**		
2GTR+5C+I+4Γ	1068.96	949.55	881.93	853.55	828.5	822.65	218.19	181.98	149.36	123.49	94.57	79.7	**2GTR+5C+I+4Γ**	
3GTR+5C+I+4Γ	1097.11	977.69	910.07	881.7	856.64	850.8	358.89	210.12	177.51	151.63	122.72	107.84	11.53	**3GTR+5C+I+4Γ**
4GTR+5C+I+4Γ	1118.31	998.89	931.27	902.9	877.84	872	380.09	231.32	198.7	172.83	129.04	129.04	32.73	3.17

66TaxonSet_aa				
	1JTT+I+4Γ			
1JTT+1C+I+4Γ	217.3	**1JTT+1C+I+4Γ**		
1JTT+2C+I+4Γ	283.3	63.1	**1JTT+2C+I+4Γ**	
1JTT+3C+I+4Γ	326.6	106.5	31.6	**1JTT+3C+I+4Γ**
1JTT+4C+I+4Γ	369.4	149.3	74.4	39.8

66TaxonSet_day														
	1GTR+1C+I+4Γ													
2GTR+1C+I+4Γ	67	**2GTR+1C+I+4Γ**												
3GTR+1C+I+4Γ	102.7	29.4	**3GTR+1C+I+4Γ**											
4GTR+1C+I+4Γ	121.9	48.6	8.4	**4GTR+1C+I+4Γ**										
5GTR+1C+I+4Γ	152.3	79	38.7	16.7	**5GTR+1C+I+4Γ**									
6GTR+1C+I+4Γ	171	97.6	57.4	35.4	7.1	**6GTR+1C+I+4Γ**								
1GTR+2C+I+4Γ	16.4	-56.9	-97.1	-119.2	-147.5	-165.3	**1GTR+2C+I+4Γ**							
1GTR+3C+I+4Γ	28.9	-44.5	-84.7	-106.7	-135	-152.8	9	**1GTR+3C+I+4Γ**						
1GTR+4C+I+4Γ	44.8	-28.6	-68.8	-90.8	-119.1	-136.9	25	8.9	**1GTR+4C+I+4Γ**					
1GTR+5C+I+4Γ	55.7	-17.6	-57.8	-79.9	-108.2	-126	35.9	19.8	6.3	**1GTR+5C+I+4Γ**				
1GTR+6C+I+4Γ	64.2	-9.1	-49.4	-71.4	-99.7	-117.5	44.4	28.3	14.8	1	**1GTR+6C+I+4Γ**			
1GTR+7C+I+4Γ	71.5	-1.9	-42.1	-64.1	-92.4	-110.2	51.7	35.6	22.1	8.3	-1.6	**1GTR+7C+I+4Γ**		
5GTR+2C+I+4Γ	169.3	96	55.8	33.7	5.4	-12.4	149.5	133.4	119.9	106.1	96.2	88.3	**5GTR+2C+I+4Γ**	
5GTR+3C+I+4Γ	184.5	111.1	70.9	48.9	20.6	2.8	164.7	148.6	135.1	121.3	111.4	103.5	1.6	**5GTR+3C+I+4Γ**
5GTR+4C+I+4Γ	199.8	126.5	86.3	64.2	35.9	18.1	180	163.9	150.4	136.6	126.7	118.9	17	-7.8

Model 1 is given on the top row and Model 2 is on the left column. The results of 2[lnL(Model2) – lnL(Model1)] are shown in the intersecting cell between the 2 models. The calculation of 2ln(BF) was carried out using the Kass and Raftery table (shown in inset) as a guide. 2ln(BF)>6 strongly supports Model2.

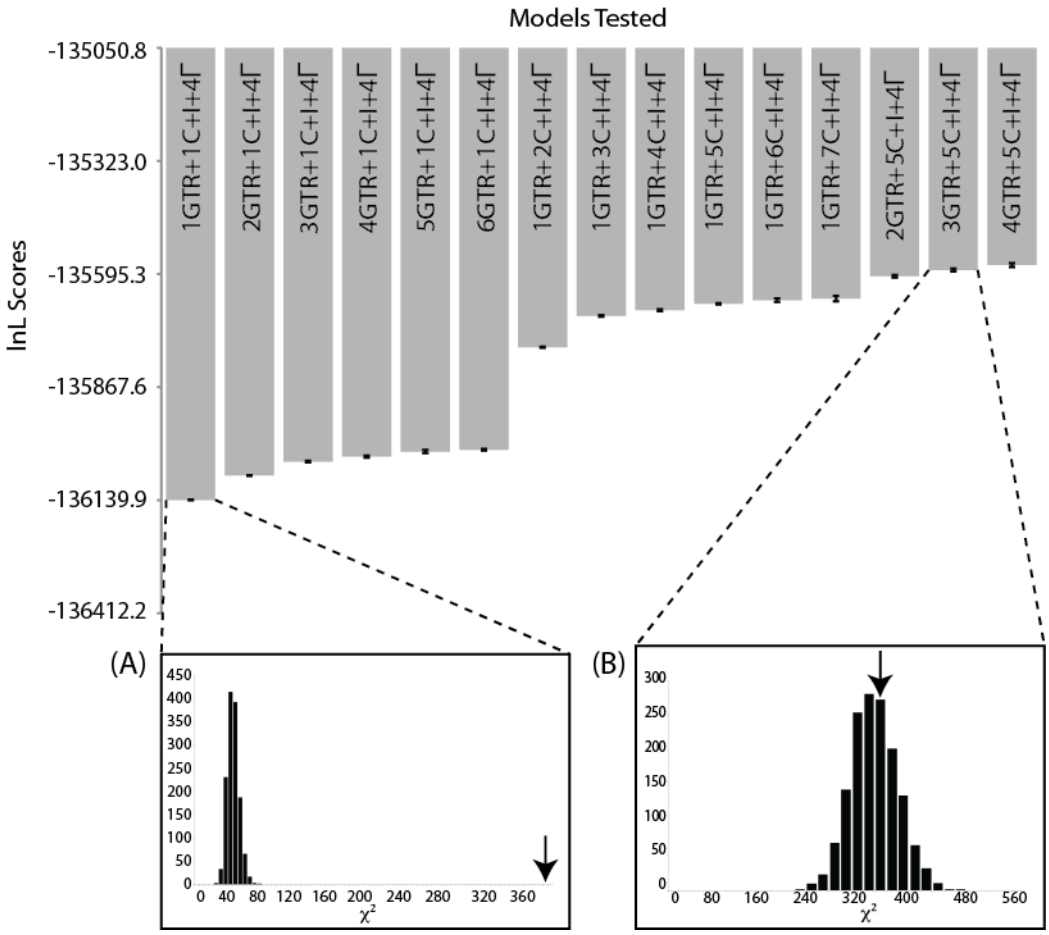


Figure 2.8 Fit of alternative models applied to 66TaxonSet_nuc, and posterior predictive simulations for the (A) homogeneous model and (B) heterogeneous model of best fit following BF analysis for 66TaxonSet_nuc.

Models used in the analysis are shown on x-axis. Inside each grey bar is the composition of the individual model, i.e. the number of rate matrices and composition vectors. The Y-axis shows the lnL score, the higher the grey bar the worse the fit of the model. Inset (A) shows posterior predictive simulations for the homogeneous model (1GTR+1C+I+4Γ). Inset (B) shows posterior predictive simulation for the best-fit heterogeneous model (3GTR+5C+I+4Γ). The black bar graph in inset (A) and (B) represents the posterior predictive simulations of these models on 66TaxonSet, and the arrows represent the $\chi 2$ position in the simulation for the real data.

The number of composition vectors were increased incrementally on the 66TaxonSet_aa to evaluate the improvement of fit of the application of compositionally heterogeneous models over the phylogeny. Convergence was assessed by ensuring the at least 6 of the 10 independent MCMC runs were in agreement on parameters and topology following generations of majority rule consensus tree (described in section 2.2.6.2). It was found that models 1JTT+5C+I+4Γ, 1JTT+6C+I+4Γ and 1JTT+7C+I+4Γ did not reach convergence during the MCMC run, which was not surprising given each of these models contained over 100 free parameters. BF comparisons indicated that the 1JTT+4C+I+4Γ model provided best fit to the data, and is better than the compositionally homogeneous model 1JTT+1C+I+4Γ by 2ln(BF) = 369.4. Posterior predictive simulations showed that both homogeneous (1JTT+1C+I+4Γ), and the best fit model allowing for compositional heterogeneity over the phylogeny (1JTT+4C+I+4Γ), fitted the data with tail area probabilities of 0.79 and 0.97 respectively, see Figure 2.9 (A) and (B). In summary, the posterior predictive simulation study showed that the compositionally homogeneous model were satisfactory in describing the 66TaxonSet_aa, however BF comparisons show that the model that accounted for compositional heterogeneity across the phylogeny described the data better.

To determine the effect of multiple rate matrices on the amino acid dataset and remove a layer of compositional heterogeneity, the 66TaxonSet_aa was recoded to 6 Dayhoff categories (66TaxonSet_day). It was found that increasing the number of rate matrices had a greater effect on improvement of model fit, as shown in comparison of model 1GTR+1C+I+4Γ to model 2GTR+1C+I+4Γ where 2ln(BF) = 67.0. This was compared to the impact of increasing composition vectors, where model 1GTR+1C+I+4Γ compared to model 1GTR+2C+I+4Γ gave 2ln(BF) = 16.4. BF comparisons identified model 5GTR+2C+I+4Γ as the best fitting model, and when it was compared against the homogeneous model (1GTR+1C+I+4Γ) it resulted in a large improvement in fit, i.e. 2ln(BF) = 169.3. Posterior predictive simulations showed that both the homogeneous model 1GTR+1C+I+4Γ and the combined NDCH and NDRH model of best fit, i.e. 5GTR+2C+I+4Γ, were both adequate in describing the composition of the data, see Figure 2.9 (C) and (D).

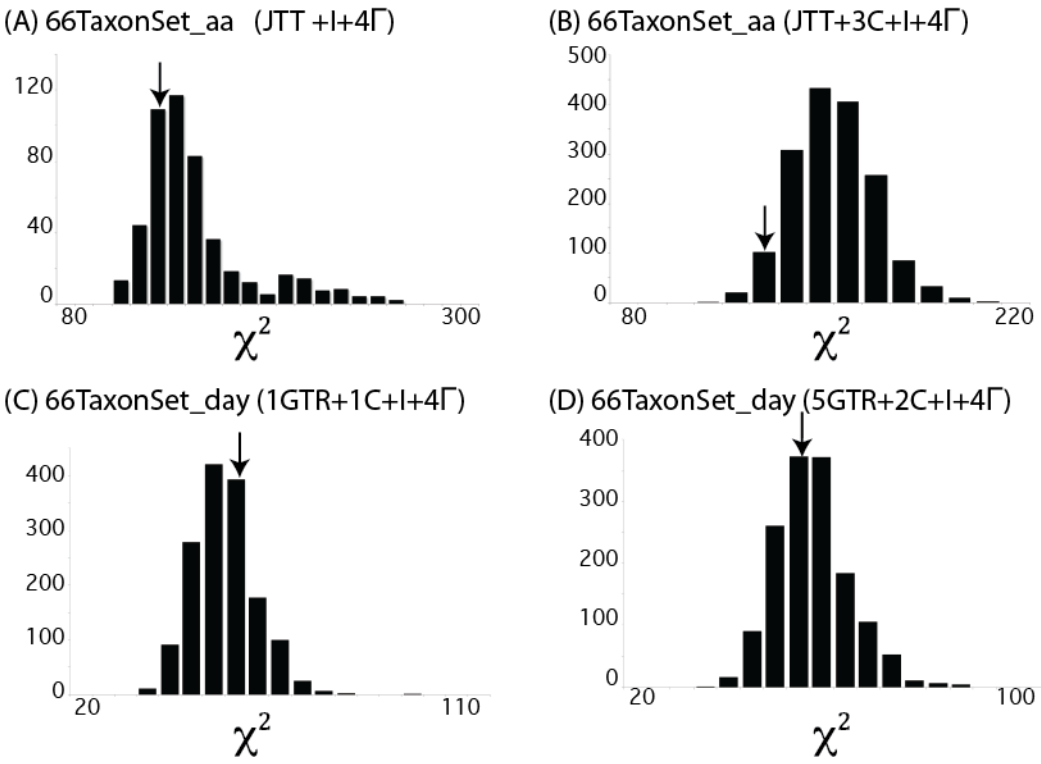


Figure 2.9 Posterior predictive simulations in P4 for 66TaxonSet

Bar graphs (A) and (C), represent the posterior predictive simulations of homogeneous models (JTT+I+4Γ and 1GTR+1C+14Γ) for 66TaxonSet. Bar graphs (B) and (D) represent the posterior predictive simulations of the heterogeneous models of best fit. The arrows in all cases represent the χ^2-position for the real data.

After testing the impact of models that accommodate heterogeneity over the phylogeny, the next step was to explore the impact of models that accommodate heterogeneity across the dataset by applying profile mixture model CAT (Lartillot and Philippe 2004). It was not possible to test the fit of the CAT (Lartillot and Philippe 2004) model to the dataset using BF comparisons, instead 10-fold Bayesian cross validation was employed. The fit of dataset heterogeneous models CAT and CAT-GTR (Lartillot and Philippe 2004) were compared against the fit of homogeneous models GTR+1C+I+4Γ and JTT+I+4Γ. The results indicated that for both 66TaxonSet_nuc and 66TaxonSet_day a GTR+I+4Γ model fitted the data best (see Table 2.9). For 66TaxonSet_aa CAT (Lartillot and Philippe 2004) is the best fitting model. The results of cross validation combined with the BF analyses showed that the model with the GTR rate matrix and compositional heterogeneity described 66TaxonSet_nuc and 66TaxonSet_day better than compositionally homogeneous models. Therefore, these heterogeneous models were used to derive optimal phylogenies for 66TaxonSet_nuc and 66TaxonSet_day, whilst the CAT (Lartillot and Philippe 2004) model was used to find the optimal tree for 66TaxonSet_aa.

Table 2.9 Cross Validation (CV) of models in PhyloBayes for 66TaxonSet

	GTR+1C+4Γ	**CAT**	**CAT-GTR**	**JTT+4Γ**
66TaxonSet_nuc				
GTR+1C+4Γ	------	-100.29 ± 526.11	-34.44 ± 490.49	------
CAT	100.29 ± 526.11	------	65.85 ± 420.09	------
CAT-GTR	34.44 ± 490.50	-65.85 ± 420.09	------	------
66TaxonSet_aa				
GTR+1C+4Γ	------	4.03 ± 282.58	3.28 ± 231.91	-96.41 ± 428.08
CAT	-4.027 ± 282.575	------	-0.75 ± 270.18	-100.44± 301.98
CAT-GTR	-3.278 ± 231.908	0.75 ± 270.18	------	-99.69 ± 400.85
JTT+4Γ	100.44 +/- 301.984	96.41 ± 428.08	99.69 ± 400.85	------
66TaxonSet_day				
GTR+1C+4Γ	------	-242.63 +/- 333.54	-101.26 ± 386.99	------
CAT	242.63 ± 333.54	------	141.38 ± 288.15	------
CAT-GTR	101.26 ± 386.99	-141.38 ± 288.15	------	------

The average of the CV log-likelihood scores compared across all models in PhyloBayes v3.2 (Lartillot et al. 2009). The standard deviation between the CV scores is shown after the "±". The reference model is shown on the horizontal and the test model is shown on the vertical. A positive score indicates that the tested model is better than the reference

model.

2.3.2.2 Phylogenetic Results of Best Fitting Models on 66TaxonSet

In contrast to the previously published topology for 66TaxonSet data, there were significant topological changes observed between the original published phylogeny (Murphy et al. 2001a) and the phylogenies obtained from the application of models accounting for heterogeneity over the phylogeny and the data. The posterior probability support values for placement of the placental root using the best fitting models and 66TaxonSet remained low: 66TaxonSet_nuc and 66TaxonSet_aa support the Atlantogenata hypothesis with a posterior probability of 0.46 (3GTR+5C+I+4Γ) and 0.51 (CAT) respectively. The 66TaxonSet_day supported the Epitherian hypothesis with a high posterior probability score of 0.98 (5GTR+2C+I+4Γ). Comparison of BF scores showed that 66TaxonSet_day was unable to reject the Atlantogenata hypothesis, 2ln(BF) = 1.96, the Exafroplacentalia hypothesis, 2ln(BF) = 5.24 or the Rodent hypothesis, 2ln(BF) = 0.34. See Table 2.10 for all BF comparison scores. The results indicated that while models that accommodate heterogeneity across the phylogeny and the data describe all treatments of 66TaxonSet better (66TaxonSet_nuc, _aa and _day), this dataset did not have sufficient phylogenetic information to support a single topology and reject all other competing phylogenetic hypotheses for the placement of the placental mammal root.

Further evidence for phylogenetic conflict in the 66TaxonSet dataset is found through LM analysis. There were 4.4% and 12.8% phylogenetic conflict identified in 66TaxonSet_nuc and 66TaxonSet_aa respectively using homogeneous models of evolution (GTR+I+4Γ and JTT+I+4Γ). While the nucleotide dataset passed the cut-off value of 10%, the amino acid dataset is nested within this dataset and failed the LM cut-off points. BF comparisons showed that the 66TaxonSet dataset was unable to support one topology over competing hypotheses and LM analysis shows that high levels of phylogenetic conflict are present in 66TaxonSet_aa. Combined these results suggests the dataset was not informative enough to resolve the placental mammal phylogeny, particularly with regards to the root. A summary of theses analyses is shown in Table 2.11.

Table 2.10 Testing alternative rooting hypotheses.

DataSet	Model of Best Fit (Model 1)	Alternative Rooting Hypotheses (Model 2)			
66TaxonSet_nuc	**Atlantogenata (3GTR+5C+I+4Γ)**	**Epitherian**	**Exafroplacentalia**	**Rodent**	**66TaxonSet Original**
2(lnLBF)	----------	6.20	6.98	0.44	3.48
66TaxonSet_aa	**Epitherian (1JTT+4C+I+4Γ)**	**Atlantogenata**	**Exafroplacentalia**	**Rodent**	**66TaxonSet Original**
2(lnLBF)	----------	8.44	7.64	4.40	23.24
66TaxonSet_day	**Epitherian (5GTR+2C+I+4Γ)**	**Atlantogenata**	**Exafroplacentalia**	**Rodent**	**66TaxonSet Original**
2(lnLBF)	----------	1.96	5.24	0.34	0.24

The 2ln(BF) are shown for the topologies generated under heterogeneous models of best fit compared with the competing rooting hypotheses and against the topology published in the original publication of the 66TaxonSet. 2ln(BF) are not calculated for 39TaxonSet_aa against competing hypotheses, as it failed the posterior predictive sampling test.

Table 2.11 Summary of data quality and phylogenetic tests applied to 66TaxonSet

Datasets	Compositional Heterogeneity Detected	Tree Heterogeneous	Dataset Heterogeneous	Likelihood Mapping	Root Position (Tree/Dataset)	Competing Hypothesis 2ln(BF) > 6
66TaxonSet_nuc	Yes	3GTR+5C+I+4Γ	GTR+I+4Γ	Pass	$C^{<0.5}$ /C	A,B
66TaxonSet_aa	Yes	1JTT+4C+I+4Γ	CAT	Fail	$B^{<0.5}$ /C	A,C
66TaxonSet_day	No	5GTR+2C+I+4Γ	GTR+I+4Γ	NA	B /$B^{<0.5}$	None

The results of the χ^2 test of compositional homogeneity are shown. "Yes" indicates there is compositional heterogeneity, "No" indicates that there is compositional homogeneity. The model of best fit from P4 and PhyloBayes analyses are shown. Pass/Fail is indicated with regards to LM tests. The placental root supported by each dataset is given: (A) Exafroplacentalia hypothesis, (B) Epitherian hypothesis, (C) Atlantogenata hypothesis and (D) Rodent hypothesis. The ability of each dataset to reject alternative rooting hypotheses is given, 2ln(BF) > 6 indicate alternative-rooting hypotheses can be strongly rejected. Superscript < 0.5 denotes posterior probability < 0.50.

2.3.3 Heterogeneous Modelling of the 39TaxonSet

Results from 66TaxonSet indicated that it was necessary to construct a novel dataset that (i) was large enough to accommodate the parameter rich heterogeneous models while being small enough to have parameters tested using P4, and (ii), had enough phylogenetic information to conclude on one topology and reject competing hypotheses.

The 39TaxonSet consisted of SGO families and contained 23,900 aligned amino acids (aa) across 39 taxa (35 placental mammals). Amino acid data were used as they are known to ameliorate problems with compositional biases and possible codon usage biases and also because amino acid sequences saturate more slowly (Philippe et al. 2011). The dataset was assembled using a strict best reciprocal blast hit approach (see materials and methods section 2.2.2.1). This dataset was recoded into 6 Dayhoff categories (39TaxonSet_day) for the same reasons applied to 66TaxonSet, see section 2.3.1.

2.3.3.1 Testing the fit of Heterogeneous Models to 39TaxonSet

There were 22 different models applied to the 39TaxonSet and these ranged from the homogeneous model with one composition vector and one exchange rate matrix up to phylogenetically heterogeneous models 1 JTT matrix and 7 composition vectors for 39TaxonSet_aa (amino acids). For 39TaxonSet_day up to 5 GTR matrices and 4 estimated composition vectors were applied to the data (see Table 2.5). χ^2 tests indicated overall compositional homogeneity, however the model fit test (Foster 2004) showed that 35/39 taxa fail to fit the expectation of the JTT+4Γ model in 39TaxonSet_aa and 21/39 taxa fail to fit the GTR+1C+4Γ model in 39TaxonSet_day.

BF analyses were employed to determine if there was strong evidence for the application of more parameter rich models over less parameter rich models for 39TaxonSet_aa. BF analyses determined that the models improved in fit as the number of composition vectors were increased over the dataset. The homogeneous model 1JTT+4Γ had lnL = -225184.25 while the heterogeneous model of best fit 1JTT+5C+4Γ scored lnL = -223954.76. Overall this BF analysis shows very strong support for the use of the heterogeneous model with 2ln(BF) = 2459.0, (see Table 2.12). It was observed that models allowing for compositional heterogeneity over the phylogeny did not adequately describe the data (P = 0.000), see Figure 2.10 (A) and (B).

Using more parameter rich models with the JTT rate matrix was not possible as the rate exchange matrix is a fixed. It was therefore necessary to test the effect of NDRH based models on the Dayhoff recoded dataset, 39TaxonSet_day. BF comparisons between models accounting for compositional and exchange rate heterogeneity across the phylogeny 2GTR+4C+4Γ (lnL = -103791.31) and the homogeneous model 1GTR+1C+4Γ (lnL = -103967.31) showed strong support for the heterogeneous model with 2ln(BF) = 352.0, see Table 2.12. The fit of the model 2GTR+4C+4Γ to the data was supported by posterior predictive simulations with a tail-area probability of 0.004, see Figure 2.10(D). While this model did not fit the data within the 95% confidence interval, it was a better description of the data than homogeneous model 1GTR+1C+4Γ that has a tail-area probability of 0.000, see Figure 2.10(C). The distribution of the parameters over the phylogeny for the 2GTR+4C+4Γ model are shown in Figure 2.11.

Table 2.12 Comparing the fit of tree heterogeneous models applied to 39TaxonSet.

39TaxaSet_aa												
	1JTT+4Γ											
1JTT+1C+4Γ	973.6	**1JTT+1C+4Γ**										
1JTT+2C+4Γ	2078.7	1101.0	**1JTT+2C+4Γ**									
1JTT+3C+4Γ	2347.6	1370.0	262.8	**1JTT+3C+4Γ**								
1JTT+4C+4Γ	2408.5	1430.9	323.7	51.8	**1JTT+4C+4Γ**							
1JTT+5C+4Γ	2459.0	1481.4	374.2	102.3	25.4	**1JTT+5C+4Γ**						
1JTT+6C+4Γ	2443.9	1466.3	359.1	87.2	10.2	-29.3	**1JTT+6C+4Γ**					
1JTT+7C+4Γ	2486.2	1508.6	401.4	129.5	52.6	13.1	-23.3	**1JTT+7C+4Γ**				
1JTT+8C+4Γ	2432.3	1454.7	347.5	75.6	-1.3	-40.9	-77.2	-78.5				

39TaxaSet_day												
	1GTR+1C+4Γ											
2GTR+1C+4Γ	53.2	**2GTR+1C+4Γ**										
3GTR+1C+4Γ	92.9	39.2	**3GTR+1C+4Γ**									
4GTR+1C+4Γ	109.0	55.3	8.9	**4GTR+1C+4Γ**								
5GTR+1C+4Γ	115.1	61.4	15.0	2.0	**5GTR+1C+4Γ**							
1GTR+2C+4Γ	195.3	141.6	95.2	82.2	72.4	**1GTR+2C+4Γ**						
1GTR+3C+4Γ	268.4	214.6	168.2	155.3	145.5	70.3	**1GTR+3C+4Γ**					
1GTR+4C+4Γ	293.1	239.3	192.9	180.0	170.2	95.0	21.3	**1GTR+4C+4Γ**				
1GTR+5C+4Γ	324.1	270.4	224.0	211.1	201.3	126.0	52.4	1.0	**1GTR+5C+4Γ**			
2GTR+4C+4Γ	352.0	298.3	251.9	267.2	229.1	153.9	80.3	28.9	5.6	**2GTR+4C+4Γ**		
3GTR+4C+4Γ	380.3	326.6	280.2	267.2	257.4	182.2	108.6	57.2	33.9	3.5	**3GTR+14C+4Γ**	
4GTR+1C+4Γ	399.9	346.2	299.8	286.8	277.0	201.8	128.2	76.8	53.5	23.1	-6.2	**4GTR+4C+4Γ**
5GTR+1C+4Γ	405.4	351.7	305.3	292.4	282.6	207.4	133.7	82.3	59.0	28.6	-0.7	-9.2

Model 1 is given on the top row and Model 2 is on the left most column. The results of 2[lnL(Model2) – lnL(Model1)] are shown in the intersect. 2(lnLBF) were carried out using the Kass and Raftery table (shown in inset) as a guide. 2ln(BF) > 6 strongly support Model2.

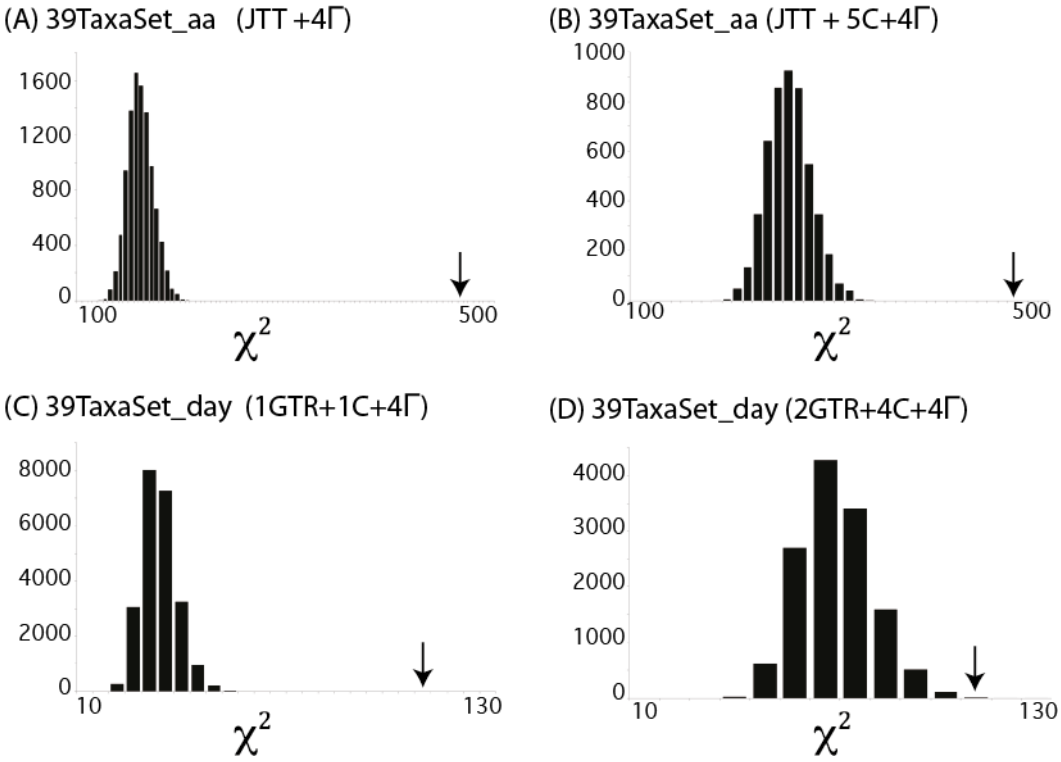


Figure 2.10 Posterior predictive simulations in P4 for 39TaxonSet

Bar graphs (A) and (C), represent the posterior predictive simulations of homogeneous models (JTT+I+4Γ and 1GTR+1C+4Γ) for 39TaxonSet. Bar graphs (B) and (D) represent the posterior predictive simulations of the tree heterogeneous models of best fit. The arrows in all cases represent the χ^2-position for the actual dataset.

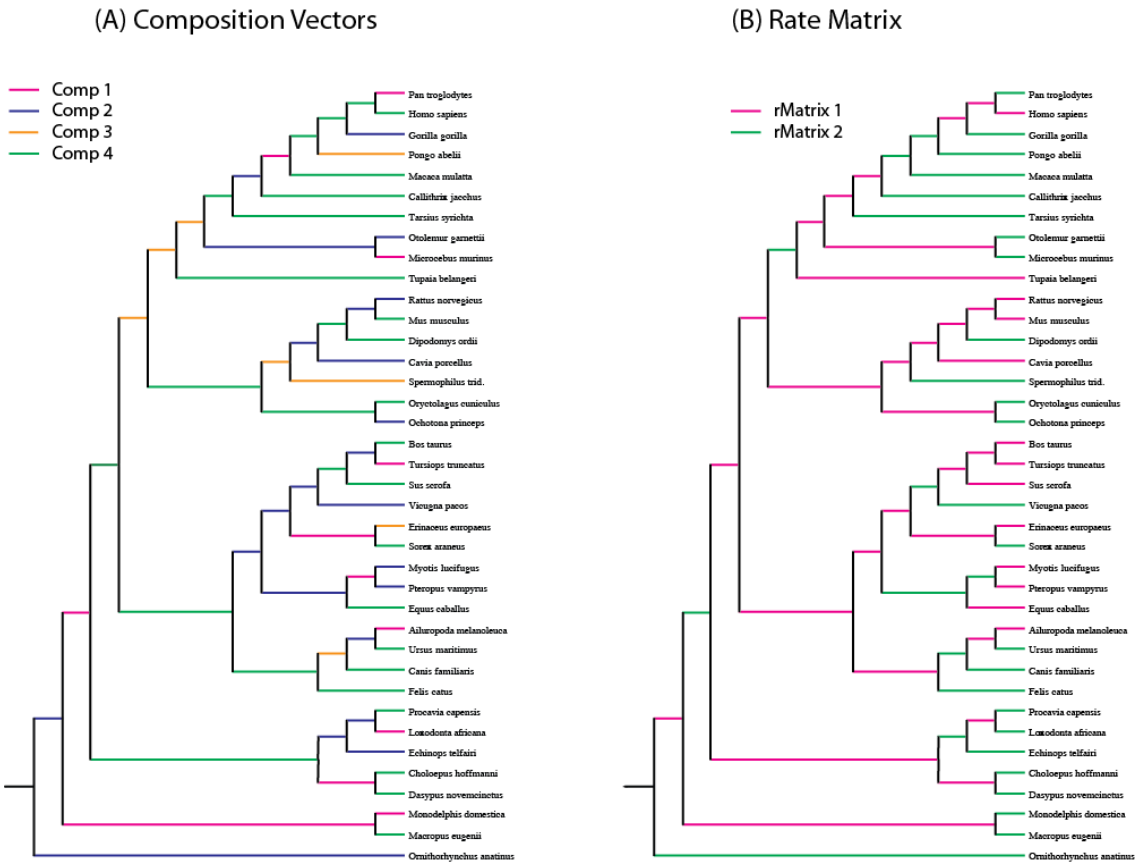


Figure 2.11 Parameter optimisation on 2GTR+4C+4Γ model generated topology.

The optimisation of (A) the 4 composition vectors and (B) the 2 rate exchange matrixes are shown over the topology obtained using the 2GTR+4C+4Γ model in P4.

Similarly to the analysis of the 66TaxonSet, the 39TaxonSet_aa was examined using heterogeneous profile mixture models across the data. The CAT and CAT-GTR (Lartillot and Philippe 2004) models were applied to the 39TaxonSet_aa to determine if profile mixture models fitted the data better than homogeneous GTR+1C+4Γ and JTT+4Γ based models. The fit was assessed using 10-fold Bayesian cross validation which identified CAT-GTR (Lartillot and Philippe 2004) as a better fit to the 39TaxonSet_aa. Posterior predicative simulations showed that this dataset was compositionally heterogeneous, both globally and more specifically for 9/39 taxa. The nine taxa that were compositionally heterogeneous (Z score > 2) from the compositional homogeneity test are highlighted in Figure 2.11. To determine the effect of the observed compositional heterogeneity on the phylogeny (i) the offending taxa were removed and repeated phylogenetic analysis was repeated, and (ii), the dataset was analysed at the dayhoff category level (39TaxonSet_day). When the compositionally heterogeneous taxa were removed the CAT-GTR (Lartillot and Philippe 2004) model on the reduced 39TaxonSet_aa (amino acids) dataset passed the global test of compositional homogeneity. Therefore analysis of the 39TaxonSet_day determined that CAT-GTR (Lartillot and Philippe 2004) was the best-fit to the 39TaxonSet_day as compared to CAT (Lartillot and Philippe 2004) and GTR+1C+4Γ models. Posterior predictive simulations of 39TaxaSet_day showed that while the CAT-GTR (Lartillot and Philippe 2004) model globally fits the data (Z score = 1.88555), 10 taxa were compositionally heterogeneous (Z score > 2). As recoding the 39TaxaSet_day data into Dayhoff categories did not ameliorate compositional heterogeneity under the CAT-GTR (Lartillot and Philippe 2004) model, it was determined that the NDCH and NDRH heterogeneous model 2GTR+4C+4Γ was a better description of the 39TaxonSet_day dataset and the CAT-GTR (Lartillot and Philippe 2004) model was a better choice for the 39TaxonSet_aa dataset.

2.3.3.2 Phylogenetic Results of Best Fitting Models on 39TaxonSet

Support was found for the Atlantogenata position of the root using a model that accounted for compositional and exchange rate heterogeneity across the tree (2GTR+4C+4Γ) and a model that accounted for heterogeneity across the dataset (CAT-GTR) for datasets 39Taxonset_day and 39Taxonset_aa respectively. The root of placental mammals was most strongly supported when placed on the branch leading to the common ancestor of Xenarthra and Afrotheria, see Figure 2.12. The removal of

compositionally heterogeneous taxa (identified in 39TaxonSet_aa through Posterior Predictive simulations and highlighted in Figure 2.12) did not alter the topology and the Atlantogenata root remained the preferred position compared to the three alternative positions considered. While the CAT-GTR (Lartillot and Philippe 2004) model was the best fitting model overall, it was not possible to compare alternative topologies under CAT (Lartillot and Philippe 2004) based models employed in PhyloBayes v3.2 (Lartillot et al. 2009). Therefore the best-fitting P4 (Foster 2004) heterogeneous model 2GTR+4C+4Γ was used (supports the Atlantogenata hypothesis with a posterior probability value of 0.96). Using the 2GTR+4C+4Γ model, the BF analyses indicated that 39TaxonSet_day was able to discriminate between alternative phylogenetic hypotheses with BF scores as follows: Exafroplacentalia 2ln(BF) = 16.08, Epitherian 2ln(BF) = 6.54 and Rodent 2ln(BF) = 298.46. While an improvement in fit of the model to the dataset was observed with the addition of compositional vectors and rate matrices, support for Atlantogenata as the sister group of all the other mammals was not eroded with more complex models.

The phylogeny produced using the CAT-GTR (Lartillot and Philippe 2004) model on the 39TaxonSet_aa was almost completely congruent with that obtained through the analysis of the 39TaxonSet_day using the 2GTR+4C+4Γ, see Figure 2.12. The only disagreement between these two topologies lay within the Laurasiatheria and relates to the placement of the 5 orders within this clade. The focus of this section of the analysis was the resolution of the root of the placental mammal phylogeny, to address the intra-order placement of the Laurasiatheria denser taxon sampling and ideally the complete genome sequencing of poorly represented Orders was necessary. This incongruence regarding the placement of Orders within the Laurasiatheria is explored in more detail in section 2.3.4. A summary of the results using the heterogeneous models described in this Chapter are given in Table 2.13.

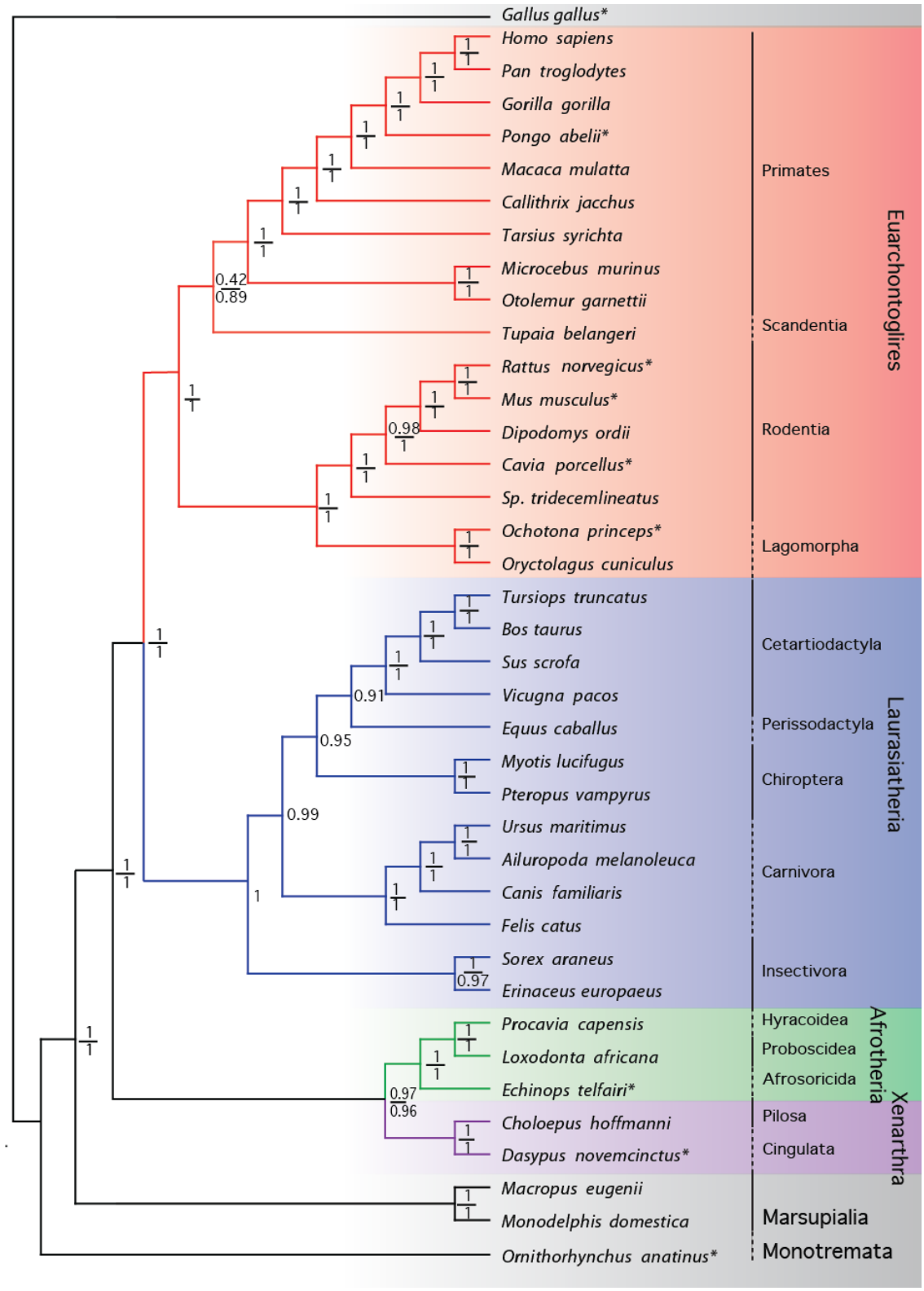


Figure 2.12 Phylogeny Reconstruction of 39TaxonSet

Figure 2.12 Legend : Phylogeny reconstruction carried out using CAT-GTR model on the 39TaxonSet_aa in PhlyoBayes v3.2 and on 39TaxonSet_day using the heterogeneous model 2GTR+4C+4Γ in P4. The support values for both methods are given at each node: the numerator is the Bayesian support value for nodes based on CAT-GTR in PhyloBayes, the denominator is the support value based on the 2GTR+4C+4Γ model in P4. Support values for the 2GTR+4C+4Γ model are shown where they are in agreement with the CAT-GTR model topology. The colour scheme represents the major groups of mammals: red = *Euarchontoglires*; blue = *Laurasiatheria*; green = *Afrotheria*; purple = *Xenarthra*; black = Monotremes, Marsupials and outgroup taxa. (*represents compositionally heterogeneous taxa).

Table 2.13 Summary of data quality and phylogenetic tests applied to 39TaxonSet

Datasets	Compositional Heterogeneity Detected	Tree Heterogeneous	Dataset Heterogeneous	Likelihood Mapping	Root Position (tree/dataset)	Competing Hypothesis 2ln(BF)=6-10
39TaxonSet_aa	Yes	$1JTT+5C+4\Gamma^{+}$	CAT-GTR	Pass	C/C	-----
39TaxonSet_day	Yes	$2GTR+4C+4\Gamma$	CAT-GTR	NA	C/C	A, B, D

The results of the χ^2 test of compositional homogeneity are shown. "Yes" indicates there is compositional heterogeneity, "No" indicates that there is compositional homogeneity. The model of best fit from P4 and PhyloBayes analyses are shown. Pass/Fail is indicated with regards to LM tests. The placental root supported by each dataset is given: (A) Exafroplacentalia hypothesis, (B) Epitherian hypothesis, (C) Atlantogenata hypothesis and (D) Rodent hypothesis. The ability of each dataset to reject alternative rooting hypotheses is given, 2ln(BF) > 6 indicate alternative-rooting hypotheses can be strongly rejected. Superscript + denotes that this model did not fit the data, Superscript < 0.5 denotes posterior probability < 0.50.

2.3.4 Assessing the suitability of data for the resolution of intra-order placement in the Laurasiatheria

There are potentially 2 contributing factors to the incongruence in the placement of Laurasiatherian Orders. The first is short divergence times for Orders within the Laurasiatheria (~1-4 MYA) - these dates may be too close for sufficient phylogenetic signal to fix (Hallstrom and Janke 2008). As nucleotide data saturates faster than amino acid data (Kosiol et al. 2007), there maybe more phylogenetic signal to determine the branching order between the Laurasiatheria Orders at the nucleotide level. Therefore both nucleotide and amino acid datasets have been employed in this study to assess whether one data type performs better than the other for this phylogenetic problem. Hallstrom and Janke discuss the possibility of species hybridization and incomplete lineage sorting as a possible cause to the disagreement on the placement of Orders within the Laurasiatheria (Hallstrom and Janke 2010). Therefore to assess whether individual genes give strong support for conflicting phylogenies, each gene was analysed individually instead of concatenating them into a Supermatrix. The limiting factor of phylogeny reconstruction on single genes (as opposed to in a Supermatrix framework) is that the smaller datasets are not able to accommodate parameter rich heterogeneous models described in section 2.3.2 and 2.3.3. RAxML (Stamatakis 2006) has the potential to perform heterogeneous analysis of individual gene sets using the CAT approximation model, however > 50 taxa are necessary to ensure that there is enough data per column to reliably estimate the per-site rate parameters (Stamatakis 2006).Therefore in this study individual genes were analysed and only datasets where the composition was adequately described by homogeneous models were employed.

The model of evolution was calculated for the 1284 SGO alignments at both nucleotide and amino acid levels. The amino acid model of evolution for the majority of SGOs (1202/1284) was JTT. The nucleotide model of best-fit were more varied with HKY achieving the best fit in 325/1284 cases followed closely by the K80 model in 238/1284 cases. The count of the models of best-fit for nucleotide and amino acid data are shown in Table 2.14.

Table 2.14 Evolutionary Models (Substitution Matrices)

Nucleotides		Amino Acids	
Substitution Model	**Counts**	**Substitution Model**	**Counts**
GTR	73	BLOSUM62	15
HKY	325	CPREV	5
K80	238	Dayhoff	6
K81	32	JTT	1202
K81uf	92	LG	36
SYM	20	MTREV24	1
TIM	60	MtArt	1
TIMef	10	MtMam	9
TVM	91	RtREV	1
TVMef	95	VT	3
TrN	200	WAG	5
TrNef	48		

Each dataset was tested for the percentage of phylogenetic conflict using LM as well as as compositional homogeneity using the model of best-fit. If a dataset contained more than 10% phylogenetic conflict it was deemed unsuitable for further phylogenetic studies. Overall there were extremely high levels of phylogenetic conflict in the Laurasiatheria datasets, with 1138/1284 nucleotide datasets and 1251/1284 amino acid datasets containing >10% phylogenetic conflict, see Figure 2.13(A). The nucleotide dataset had much less phylogenetic conflict than the amino acid datasets, with over 113 additional genes having < 10% phylogenetic conflict in the dataset, see Figure 2.13(A). The χ^2 test of compositional homogeneity in Tree-Puzzle v5.2 (Schmidt et al. 2002) was performed on both the nucleotide and amino acid datasets to compare the composition of each sequence to the frequency distribution assumed by the model of best fit . It was found that the amino acid models were more effective in modelling the composition of the datasets compared to the nucleotide models. In total it was found that 1223/1284 of the SGO datasets had $P > 0.05$ for all 14 taxa, while only 487/1284 of the nucleotide SGO datasets had $P > 0.50$ for all 14 taxa. The frequency distribution of taxa passing the χ^2 test for compositional homogeneity ($P > 0.05$) is shown in Figure 2.13(B).

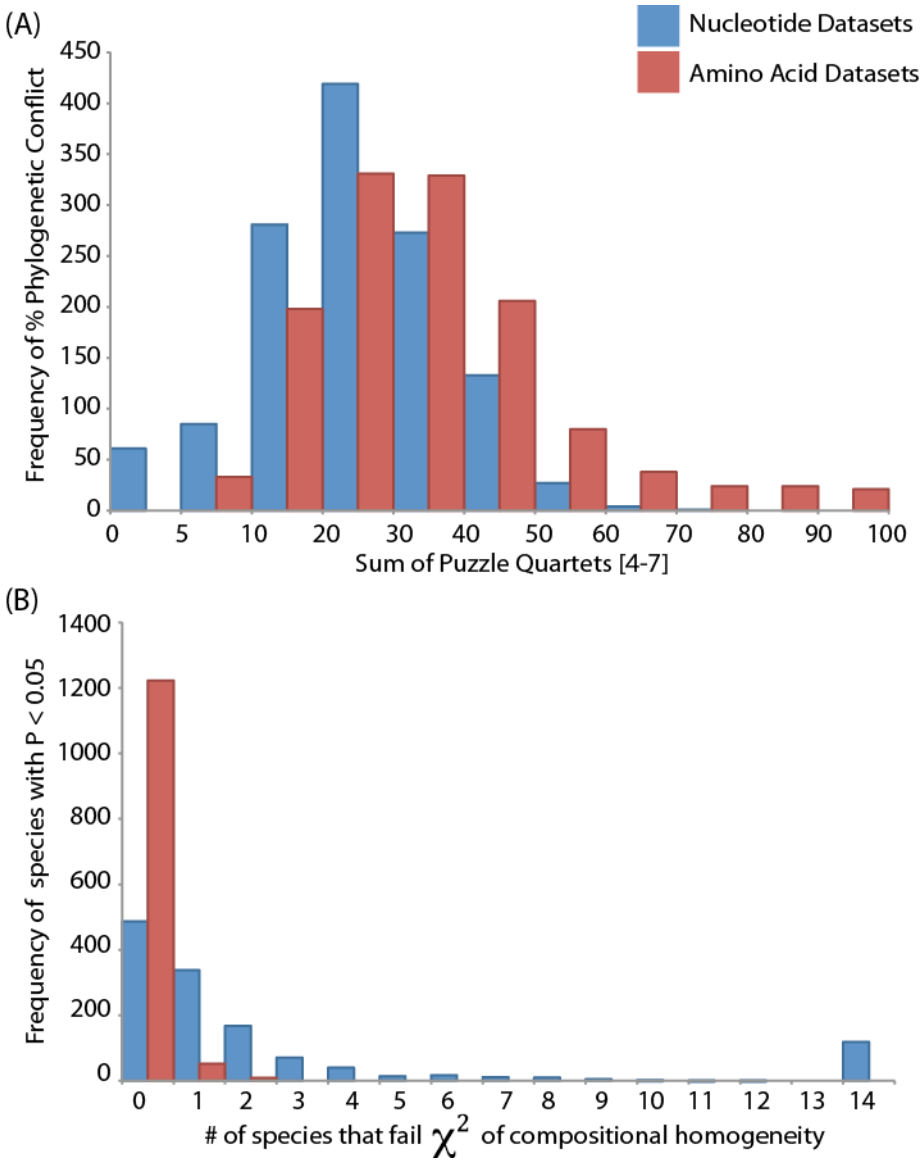


Figure 2.13 Phylogenetic Conflict and Composition Bias in amino acid and nucleotide Datasets

The frequency of (A) phylogenetic conflict (y-axis) is plotted against the sum of LM values for quartets 4 to 7 (x-axis) and (B) species specific compositional heterogeneity with $P < 0.05$ (y-axis) is plotted against the number of species that fail the χ^2 test for compositional homogeneity (x-axis) for nucleotide (red) and amino acid (blue) datasets.

2.3.4.1 Phylogenetic Results of 14TaxonSet

To ensure only high quality datasets were used to generate topologies; datasets with < 10% phylogenetic conflict and datasets whose model of best-fit do not adequately describe the composition of these data were discarded. These quality control cut-off criteria lead to 22 amino acid datasets and 8 nucleotide datasets, only one of which intersects both data types. Phylogenies were inferred using the associated models of best-fit in hybrid MrBayes v.3.1.2h (Huelsenbeck and Ronquist 2001). In cases where the model of best-fit was not available the next best fitting model was chosen (as indicated through BIC analysis in ModelGenerator v85 (Keane et al. 2006)). All resulting phylogenies have been made available in Appendix A.5.

Phylogenetic analysis of 14TaxonSet_nuc dataset resulted in 2/8 of the trees being fully bifurcating while 18/22 of 14TaxonSet_aa datasets were fully bifurcating. SH tests were applied to see if the data had the phylogenetic information necessary to reject any of the 10 Laurasiatherian hypotheses using the model of best-fit. The results of the SH tests are detailed in Table 2.15. High levels of support were observed for 14TaxonSet_nuc, where all competing hypotheses; 2/8 datasets were able to reject hypotheses A, C, D, E, and I, while only 1/8 datasets could reject hypotheses B, F, G, H and J. The 14TaxonSet_aa datasets performed marginally better, with 8/22 datasets able to reject hypothesis H and J, 3/22 can reject D and I, 2/22 can reject hypotheses A, C, E and G while only one dataset can reject hypothesis B and F.

In summary using these data and methods, there was not enough support for one hypothesis that describes the intra-order placement of the Laurasiatheria. No dataset was capable of supporting one hypothesis and fully rejecting all other competing hypotheses under its model of evolution. Finally, nucleotide datasets, which were shown to have stronger phylogenetic signal did not perform better than amino acid datasets when attempting to resolve this phylogenetic problem.

Table 2.15 Phylogenetic Results from 14TaxonSet

Ensembl Human ID's	Dataset Length	Model	Internal Nodes Gene Tree		Competing Laurasiatheria Hypotheses									
					A	B	C	D	E	F	G	H	I	J
14TaxonSet_nuc														
ENSG00000115414	7578 bp	GTR+I+4Γ	26	1	0.57	0.59	0.89	0.81	0.52	0.36	0.72	0.16	0.31	0.11
ENSG00000134222	1146 bp	K81uf+4Γ	24	1	0.03	0.15	0.02	0.03	0.02	0.1	0.05	0.06	0.04	0.13
ENSG00000120709	1179 bp	HKY+4Γ	25	1	0.56	0.14	0.17	0.17	0.17	0.54	0.14	0.15	0.55	0.37
ENSG00000106331	1104 bp	HKY+4Γ	23	1	0	0	0	0	0	0	0	0	0	0
ENSG00000001626	4539 bp	TVM+I+4Γ	26	0.84	0.84	0.25	0.45	0.3	0.44	0.7	0.25	0.19	1	0.28
ENSG00000139330	1068 bp	HKY+4Γ	25	1	0.89	0.8	0.79	0.69	0.73	0.92	0.58	0.18	0.81	0.8
ENSG00000188001	840 bp	K80+4Γ	22	0.52	1	0.51	0.85	0.51	0.85	0.6	0.54	0.32	0.69	0.27
ENSG00000205835	1011 bp	K80+4Γ	23	0.65	0.17	0.68	0.44	0.79	0.18	0.31	1	0.69	0.17	0.29
Total Statistical Deviation (P < 0.05)				0	2	1	2	2	2	1	1	1	2	1
14TaxonSet_aa														
ENSG00000115414	2526 aa	JTT+4Γ	26	1	0.16	0.08	0.16	0.13	0.09	0.15	0.09	0.04	0.14	0.11
ENSG00000155890	730 aa	JTT+4Γ	25	1	0.16	0.06	0.06	0.06	0.04	0.26	0.05	0.07	0.2	0.11
ENSG00000132464	1178 aa	JTT+4Γ	25	0.71	0.23	0.68	0.17	0.54	0.5	0.2	1	0.5	0.32	0.14
ENSG00000134061	661 aa	JTT+4Γ	26	1	0.15	0.12	0.12	0.42	0.08	0.17	0.46	0.48	0.54	0.04
ENSG00000182348	1361 aa	JTT+4Γ	24	0.19	1	0.62	0.75	0.42	0.86	0.71	0.59	0.06	0.67	0.08
ENSG00000110427	1945 aa	JTT+4Γ	26	1	0.41	0.13	0.11	0.52	0.16	0.42	0.54	0.56	0.67	0.13
ENSG00000132591	440 aa	JTT+4Γ	25	1	0.11	0.08	0.09	0.78	0.09	0.11	0.78	0.33	0.26	0.03
ENSG00000039537	938 aa	JTT+4Γ	26	1	0.23	0.49	0.26	0.38	0.39	0.25	0.61	0.15	0.25	0.21
ENSG00000165282	1157 aa	JTT+4Γ	26	1	0.15	0.15	0.15	0.15	0.15	0.16	0.14	0.07	0.15	0.05
ENSG00000148429	864 aa	JTT+4Γ	26	1	0.21	0.07	0.39	0.34	0.08	0.2	0.09	0.06	0.2	0.04
ENSG00000010818	2753 aa	JTT+4Γ	26	1	0.09	0.38	0.09	0.04	0.1	0.21	0.08	0.05	0.04	0.1
ENSG00000164691	745 aa	JTT+4Γ	26	1	0.43	0.41	0.54	0.5	0.46	0.41	0.41	0.02	0.37	0.05
ENSG00000171456	1555 aa	JTT+4Γ	26	1	0.96	0.47	0.5	0.42	0.59	0.9	0.42	0.08	0.86	0.07
ENSG00000182326	739 aa	JTT+4Γ	26	1	0.16	0.17	0.18	0.38	0.08	0.22	0.15	0.08	0.21	0.03

Ensembl Human ID's	Dataset Length	Model	Internal Nodes Gene Tree		Competing Laurasiatheria Hypotheses									
					A	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**	**J**
ENSG00000112818	767 aa	JTT+4Γ	26	1	0.04	0.05	0.03	0.02	0.02	0.08	0.02	0.01	0.03	0.14
ENSG00000152582	1898 aa	JTT+4Γ	26	1	0.35	0.41	0.62	0.45	0.5	0.3	0.35	0.02	0.19	0.03
ENSG00000072121	2571 aa	JTT+4Γ	26	1	0.07	0.17	0.08	0.08	0.21	0.17	0.11	0.02	0.08	0.11
ENSG00000104067	1796 aa	JTT+4Γ	26	1	0.03	0.03	0.03	0.04	0.39	0.03	0.04	0.04	0.03	0.02
ENSG00000110723	2023 aa	JTT+4Γ	26	1	0.4	0.69	0.26	0.28	0.4	0.39	0.76	0.01	0.44	0.01
ENSG00000173230	3289 aa	JTT+4Γ	26	0.96	0.92	0.85	0.81	0.85	0.81	1	0.86	0.07	0.73	0.13
ENSG00000169031	1779 aa	JTT+4Γ	26	1	0.05	0.2	0.11	0.2	0.05	0.1	0.23	0.15	0.05	0.55
ENSG00000164309	4234 aa	JTT+4Γ	26	0.8	0.65	0.18	1	0.67	0.8	0.32	0.11	0	0.22	0.03
Total Statistical Deviation ($P < 0.05$)				0	2	1	2	3	2	1	2	8	3	8

2.4 Discussion

The aim of this Chapter was to determine the importance of dataset suitability and model adequacy to address the phylogenetic question at hand. Three datasets were employed and each dataset was analysed with the objective of improving knowledge of the best available approach to undertake when (i) resolving the root of the placental mammal phylogeny and (ii) intra-order placements in the Laurasiatheria.

Re analysis of the 66TaxonSet Murphy *et al* (2001a) showed that heterogeneous models fitted the data better than homogeneous models. It was also shown that the homogeneous models employed in the original publication of 66TaxonSet_nuc (Murphy et al. 2001a) did not fit the data, while the more sophisticated models did fit the 66TaxonSet. Model adequacy is critical to ensure a robust phylogenetic analysis and several studies have shown that models that fit the data poorly consistently find the wrong phylogeny (Foster et al. 2009). The 66TaxonSet was found to be compositionally heterogeneous through the χ^2 test of compositional homogeneity and the model fit test (Foster 2004). Previous work by Foster (2009) showed that when compositionally heterogeneous data is not modelled adequately an incorrect topology is recovered (Foster et al. 2009). Whether the 66TaxonSet was analysed at the nucleotide, amino acid or the dayhoff recoded level it was found that models that accommodate exchange rate and compositional heterogeneity over the phylogeny or the data out performed models that did not. One of the primary issues with the 66TaxonSet was that it lacked the in formativeness to definitely resolve the root of the placental mammals and LM tests showed that there were high levels of phylogenetic conflict. These conflicts could be attributed to the use of mitochondrial and non coding data which is known to saturate faster than nuclear data (Brown et al. 1982, Burger et al. 2003) and has been known to cause homoplasy resulting from saturation (Sanderson 1989, Sanderson and Hufford 1996). Therefore the dataset employed by (Murphy et al. 2001a) did not contain enough phylogenetically informative characters to accommodate the parameter rich models and infer the phylogeny of mammals.

The 39TaxonSet overcomes the identified issues with the 66TaxonSet. The 39TaxonSet is superior to the 66TaxonSet in that it is sufficiently large to accommodate parameter rich models and has the phylogenetic informativness to reject alternative hypotheses for the root of the placental mammals. Phylogenetic results from models that allow for heterogeneity over the phylogeny and the dataset on both 39TaxonSet_aa and 39TaxonSet_day are congruent with one another; the only region where there is disagreement is in the placement of Orders within the Laurasiatheria. Previous studies have used network analyses; single gene and Supermatrix analyses to try resolving the intra-ordinal placement of the Laurasiatheria, however the focus has predominantly been on sequence length and taxon sampling (Hallstrom and Janke 2010, Hou et al. 2009). The 39TaxonSet did not contain sufficient signal to resolve the intra order placements of the Laurasiatheria and therefore a Laurasiatheria focused 14TaxonSet was created.

Both nucleotide and amino acid data were used to generate datasets for the Laurasiatheria as divergence estimates between Orders are short (between 1 and 4 Million Years) (Hallstrom and Janke 2008). It was computationally prohibitive to apply sophisticated heterogeneous models to all 2568 datasets (combined nucleotide and amino acid count). Instead only datasets that had low levels of phylogenetic conflict (< 10%) and whose model adequately described the composition of these data were used. This drastically decreased the usable datasets from 1284 to 8 nucleotide datasets and 1284 to 22 amino acid datasets. Regardless of criteria employed, no dataset contained enough phylogenetic signal to support one hypothesis over another or over all others. It is possible that neither nucleotide nor amino acid based datasets are suitable to place the Orders within the Laurasiatheria and that other data types such as rare genomic events should be explored. Hallstrom and Janke (2010) also employed nucleotide and amino acid data to infer the intra-order placements within the Laurastheria and attributed the lack of resolution on intra-ordinal placements to introgression of gene flow between Orders (Hallstrom and Janke 2010). There are 6 known Orders within the Laurasiatheria, but only 5 have had representative genomes sequenced. Until a representative from the Pholidota is sequenced and until there is more sampling and sequencing performed on taxa from other poorly sequenced Orders, resolving the intra order placements within the Laurasiatheira will be imperfect. In addition, it is clear that no single data type can be considered in isolation when attempting to resolve difficult phylogenetic problems, and that inferences based on the interpretation of patterns of congruence/incongruence

among alternative data types would be most desirable (Rota-Stabelli et al. 2011, Campbell et al. 2011, Branger et al. 2011).

The results of this Chapter show that having more sequence data does not eradicate all phylogenetic conflict. It is difficult to expand taxon sampling due to the limited number of fully sequenced nuclear genomes. Coding genes from mitochondrial genomes have been sequenced in over 2000 mammal species and the impact of taxon rich sampling using mitochondrial data has been explored in Chapter 3.

Heterogeneous models out perform homogeneous models, however only when the dataset is of high enough quality and large enough to accommodate the parameter rich models. The importance of incorporating compositional homogeneity tests to ensure that the model of choice is adequate in modelling the composition of the data is evident following these analyses. There is no suitable method to test the fit of the rate exchange matrix to these data and therefore there is no way of knowing whether this is causing ambiguity in the data.

To conclude, this Chapter indicates the importance of having data suitable to address the phylogenetic problem at hand and a model that is able to adequately describe the evolution of such data.

Chapter 3

3 An assessment of the suitability of Mitochondrial Data for Inferring the Placental Mammal Phylogeny.

3.1 Introduction

3.1.1 Previous applications of mitochondrial data to resolving the Mammal Phylogeny

Mitochondrial genes (mtGenes) have been used to resolve the phylogenetic relationships of deep divergences such as the placement of the Superorders in the mammal phylogeny (Gibson et al. 2005, Tobe et al. 2010, Milinkovitch et al. 1993), and also for the analysis of more shallow relationships, such as those amongst the Cetacea (Milinkovitch et al. 1993), the Caniforma (Arnason et al. 2007) and the Rodentia (Frye and Hedges 1995). The mitochondrial gene CYTB was once the primary locus involved in phylogenetic studies (Irwin et al. 1991), but recently the Bar code of Life Consortium has adopted the mitochondrial gene CO1 (Hebert et al. 2003) for the resolution of the relationships amongst eukaryotes and the more rapidly evolving plastid loci matK and rbcL are being used for phylogenetic studies of plants (Li et al. 2011). The concept of using bar codes as a method of identifying species was first proposed by (Hebert et al. 2003) and was used to construct a public reference library containing species identifiers which has enabled identification of unknown species. This project has since expanded into the international Bar code of Life (iBOL) project whose goals are to collect over half a million bar coded sequences and to use phylogeny in species conservation (Vernooy et al. 2010).

The most taxon rich study to date has used both CYTB and CO1 to reconstruct the phylogeny of mammals across 204 taxa (Tobe et al. 2010). This study revealed that while CYTB was a stronger candidate than CO1 for phylogeny reconstruction, neither gene was able to resolve the branching of the Superorders (Tobe et al. 2010). An analysis that used the entire mitochondrial genome of 78 Eutherian taxa found strong support for the four Superorders of placental mammals (Kjer and Honeycutt 2007). This study had a number of inconsistencies in terms of placement of the Scandentia at the basal position in the Euarchontoglires where previous nuclear based studies have placed it as a sister group to the Primates or Dermoptera (Murphy et al. 2001a, Murphy et al. 2001b, Novacek 1992, Springer et al. 2004). Traditionally, Primate Orders have been a monophyletic group in studies based on nuclear genes (Murphy et al. 2001a, Murphy et al. 2001b). When Kjer and Honeycutt (2007) reconstructed their phylogeny based on the entire mitochondrial genome, the primates were found to be paraphyletic with the

Dermoptera grouping with anthropoid Primates to the exclusion of lineages such as tarsiers and prosimians (Kjer and Honeycutt 2007). This indicates that there may be some problems with the application of mitochondrial data to the mammal phylogeny. In this chapter I apply a variety of data assessment tests to determine which genes and at what depth mitochondrial data can be applied to mammal phylogenetics.

3.1.2 Comparison of nuclear and mitochondrial genes as phylogenetic loci.

There are several differences between the nuclear and mitochondrial genome including but not restricted to: size of genome, mode of inheritance, levels and extent of recombination, number of introns and DNA repair mechanisms (Ballard and Whitlock 2004). Mitochondrial genes (mtGenes) undergo more mutations compared to nuclear genes (nucGenes) and are therefore more susceptible to saturation of base changes - a major challenge in phylogeny reconstruction (Brown et al. 1982). In contrast, the benefits of using mtGenes in phylogenetic studies are that mtGenes have very low rates of recombination (Lunt and Hyman 1997, Ladoukakis and Zouros 2001, Awadalla et al. 1999, Hoarau et al. 2002), mtGene order is relatively well conserved across vertebrates (Pereira 2000) and mtGene sequence data is available for over 1,000 mammals (UniProt 2012). The number of fully sequenced mammal nuclear genomes remains relatively low with only 40 mammal genomes available in the Ensembl database (Hubbard et al. 2007) out of 5,488 classified mammal species (ICUN 2012). This has placed restrictions on extensive taxon sampling within the Superorders. As mitochondrial sequences are readily available for so many taxa, the use of mitochondrial sequences could serve to counterbalance the taxon-sampling deficiency in nuclear sequences. Over the past number of years, studies have used both mtGenes and nucGenes to attempt to resolve the mammal phylogeny (Nishihara et al. 2006, Hallstrom and Janke 2008, Nikolaev et al. 2007, Tobe et al. 2010).

The preference for nucGenes over mtGenes has been explained in a different context in the past (e.g. spiny lizard phylogenies (Leache 2010)) while overall there were disagreements between phylogenies inferred using mtDNA and nucDNA, phylogenies inferred using 3 mtDNA loci were in more disagreement with one another compared to phylogenies inferred from 4 nucDNA loci. A study of *Plethodon* salamanders showed that while incongruence between inferred mtDNA phylogenies was higher than inferred

nucDNA phylogenies, the combined nuclear and mitochondrial data provided enough reliable phylogenetic signal that phylogenetic inconsistencies such as homoplasy and LBA present in the mitochondrial data were overcome (Fisher-Reid and Wiens 2011). An analysis of 66 Eutherian mammals using combined nuclear and mitochondrial data showed strong support for both Superorders and Orders (Murphy et al. 2001a) (see Chapter 2 for detailed re-analysis of these data). In summary, the debate on whether mitochondrial data is suitable for phylogeny reconstruction when combined with nuclear data remains unresolved (Wiens et al. 2010, Leache 2010, Fisher-Reid and Wiens 2011). Furthermore, many studies have published taxonomic relationships based solely on mitochondrial data (Hillis and Wilcox 2005, Hyman et al. 2007) even though publications have warned against this approach (Shaw 2002, Rubinoff and Holland 2005).

Springer (2001) carried out an investigation of the informativeness of mitochondrial versus nuclear gene sequences in deep-level mammal phylogeny reconstruction. He used the available data at the time, i.e. 32 taxa across 12 mitochondrial protein coding genes, together with a parsimony and minimum evolution approach (Springer et al. 2001). The conclusions were that concatenated nuclear genes were more effective at recovering benchmark clades compared with concatenated mitochondrial genes (Springer et al. 2001). Since this study, there has been a surge in sequencing and there are mitochondrial sequence data for over 1,000 placental mammals. In addition, there have been major improvements to ML methods and this can facilitate phylogenetic reconstruction of large datasets (Stamatakis 2006).

This chapter assesses the suitability of mitochondrial data as a phylogenetic marker in resolving the placental mammal phylogeny. I sought to test the phylogenetic in formativeness of each gene for the inference of phylogenetic relationships and ultimately sought to identify mtGenes that provide the greatest phylogenetic information. I assessed the phylogenetic congruence between individual mitochondrial gene phylogenies and compared these to a phylogeny resolved from a dataset of concatenated mitochondrial genes. Phylogenetic conflict can arise as a result of taxon sampling (Hedtke et al. 2006), lack of sufficient phylogenetic characters (Rosenberg and Kumar 2003) and saturation, resulting in homoplasy at deeper phylogenetic nodes (Caterino et al. 2001, Reed and Sperling 1999). Therefore I have addressed these phylogenetic conflicts within mitochondrial data by (i) sampling fewer taxa, (ii)

assessing the phylogenetic informativeness of gene coverage versus taxon sampling, (iii) removing rapidly evolving sites, (iv) removing taxa, and finally (v), sampling sequence data at different depths on the known phylogenetic tree to assess where the phylogenetic signal starts to break down.

3.2 Materials and Methods

3.2.1 Gene and Taxon Sampling

Mitochondrion-encoded protein coding genes were downloaded for taxa that spanned the four mammal Superorders (Euarchontoglires, Laurasiatheria, Xenarthra and Afrotheria) as well as non-placental mammal outgroup species (*Monodelphis domestica and Ornithorhynchus anatinus)* and Aves (*Gallus gallus*) from the UniProtKB database (UniProt 2012) resulting in a total of 1,556 taxa. Only taxa that were represented in at least 2 out of 13 mitochondrial genes (mtGenes) were used in this analysis, which resulted in 455 taxa. For details of data used in this analysis see Table 3.1 and for extended details on individual taxon coverage see Appendix B.1.

3.2.2 Multiple Sequence Alignment

Datasets were aligned using Muscle v3.7 (Edgar 2004) using the default parameters.

3.2.3 Model Choice and Phylogeny reconstruction

Model testing was performed using ModelGenerator v85 (Keane et al. 2006), which has been previously described in the materials and methods section of chapter 2 (section 2.2.4.1). Bayesian based phylogeny reconstruction algorithms were computationally prohibitive, therefore the ML program RAxML (Stamatakis 2006) was employed for phylogeny reconstruction of the 446 individual datasets. Using the rapid bootstrapping algorithm (Stamatakis et al. 2007), 1,000 bootstrap replicates were performed on each dataset using the best-fit model. A list of models used and lnL scores for un-treated mitochondrial data is given in Table 3.1 and all models, lnL scores and phylogenetic trees are available in Appendix B.2

Table 3.1 Details of untreated mitochondrial data and model choice.

Protein Name	mtGene Name	Taxa #	MSA Length (aa)	Model of Evolution	-lnL
ATP synthase subunit a	ATP6	253	228	MtMam+I+4Γ	-13653.63
ATP synthase protein 8	ATP8	281	71	MtMam+I+4Γ	-9145.23
Cytochrome c oxidase subunit 1	CO1	187	518	MtMam+I+4Γ	-7530.62
Cytochrome c oxidase subunit 2	CO2	217	237	MtMam+4Γ	-6430.97
Cytochrome c oxidase subunit 3	CO3	189	269	MtMam+I+4Γ	-7175.67
Cytochrome b	CYTB	267	383	MtMam+I+4Γ	-23093.23
NADH-ubiquinone oxidoreductase chain 1	ND1	129	326	MtMam+4Γ	-12503.65
NADH-ubiquinone oxidoreductase chain 2	ND2	152	350	MtMam+4Γ	-27716.40
NADH-ubiquinone oxidoreductase chain 3	ND3	141	119	MtMam+4Γ	-5619.87
NADH-ubiquinone oxidoreductase chain 4	ND4	163	486	MtMam+4Γ	-25191.86
NADH-ubiquinone oxidoreductase chain 4L	ND4L	246	98	MtMam+4Γ	-7264.63
NADH-ubiquinone oxidoreductase chain 5	ND5	149	626	MtMam+I+4Γ+F	-41499.46
NADH-ubiquinone oxidoreductase chain 6	ND6	94	200	JTT+4Γ+F	-10035.93
Supermatrix(concatenated alignment)	SM	455	3906	MTMam+G+F	-204073.11

The total number of taxa, sequence lenght are given for each dataset along with their associated models of evolution and lnL values for phylogeny generated through RAxML (Stamatakis 2006).

3.2.4 Likelihood mapping tests

Likelihood mapping (LM) was performed on all datasets using TreePuzzle v5.2 (Schmidt et al. 2002) and is described in materials and methods section 2.2.9. The mtMAM+4Γ model was not available in TreePuzzle v5.2 (Schmidt et al. 2002) so the next available model of best-fit defined through BIC analysis was chosen (usually mtREV+4Γ). LM analysis can only be performed on an MSA that contains between 4 and 257 taxa. This was a built-in limit in TreePuzzle v5.2 (Schmidt et al. 2002), that exists to avoid overflow of internal integer variables. Therefore, LM analysis was performed on 413 datasets (33 datasets had >257 taxa). The LM scores for the untreated mitochondrial data is given in Table 2.3 and a fully comprehensive list of LM scores are given in Appendix B.3.

3.2.5 Removal of Saturated Sites

The rates of change of characters were categorized using TIGER (Cummins 2010), a phylogeny independent method for classification of rates across sites. Twenty bin categories were generated; where bin 1 represents characters associated with slowly evolving sites and bin 20 represents characters that are rapidly evolving. The sites that were associated with bin 20, 19 or 18 were removed sequentially using PAUP* (Swofford 2002) and the MSA was realigned and tested for phylogenetic signal using LM (Schmidt et al. 2002). An example of removal of saturated sites is shown in Figure 3.1, for illustrative purposes only 5 categories (bins) are used.

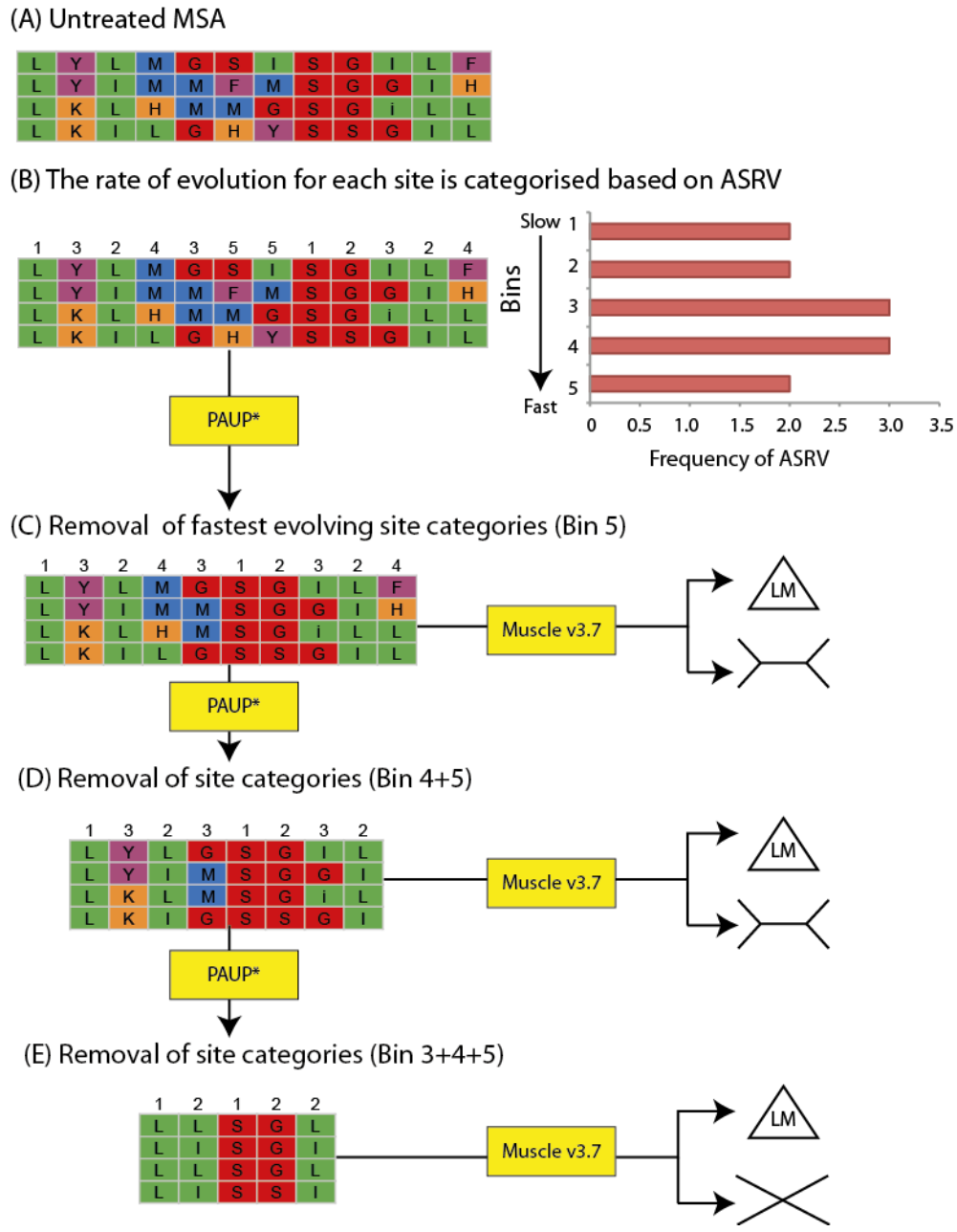


Figure 3.1 Pipeline showing the process of removal of "bins" of sites classified using TIGER.

The alignment in (A) has had the individual characters characterised based on their rate of evolution using TIGER and the output is shown in (B). The fastest evolving sites are removed (C), the alignment is re-generated as denoted by the presence of the Muscle v3.7 box, the phylogenetic signal is tested using LM and phylogenetic trees are estimated (these steps are denoted by the triangle for LM and the cartoon tree respectively). The process of removing sites and testing of remaining alignment is repeated until there is no longer sufficient phylogenetic signal.

3.2.6 Removal of "Rogue" Taxa

The mean pairwise percent identity (Mean %ID) was calculated for all sequences in a given alignment reciprocally (described as "FamID" in (Muller et al. 2005)) using the program "calc_meanid" (Muller et al. 2005). The program was run on FASTA formatted alignment files as follows:

```
calc_meanid Infile.fa > Outfile.results
```

There was a mean pairwise identify score for every pair of species in the "Outfile.results" file. The following is an example of UNIX command used to obtain the mean percent identity score for human:

```
gawk 'NR==FNR{ a[$1] += $3; b[$1]++} NR!=FNR{ for(key in a)
{if($1==key)print key,a[key]/b[key]}}' Outfile.results
human
```

This command was used to calculate the mean pairwise identities for every species. A list of all the percentage identity scores for the mtGene data is given in Appendix B.4. If the percentage identity score of a placental mammal sequence was less than that of the outgroups (chicken, platypus and opossum), then the taxon was identified as "rogue" and removed from the alignment. The adjusted datasets were realigned and tested for phylogenetic signal using LM.

3.2.7 Calculate distance between topologies

To assess congruence between topologies, a majority rule (MR) consensus tree was generated using RAxML (Stamatakis 2006) and the Robinson-Foulds (RF) distance was calculated between two phylogenetic trees using the "rfdists" command in Clann (Creevey and McInerney 2005). This RF distance metric in Clann (Creevey and McInerney 2005) estimates the number of shared splits between the shared taxon set of two unrooted trees. The numbers are reported as the ratio of the number of shared splits across the two trees, therefore a value of 0 indicates that both trees share all splits while a number of 1 is given when the pair of trees share no splits. Individual RF scores for all comparisons have not been detailed in the text, as they are so extensive. Instead, a comprehensive list of all RF scores between topologies is detailed in Appendix B.5.

3.2.8 Generation of a legible phylogeny for ease of interpretation

In this study, the number of taxa within each dataset ranged from 91 to 455, which was difficult to display on a single page and maintain a readable font size. To overcome this issue the following treatments were performed on the MR consensus phylogenies. The topological information was extracted from the MR consensus phylogenies using Newick Utilities (Junier and Zdobnov 2010) to apply the following command:

```
nw_topology RAxML_MajorityRuleConsensusTree.MR > MR_tree.tre
```

The program "NameChange.py" (Appendix B.6.1) was written to change the individual taxon names into their associated mammal Order names using the list "Orders_Names.txt" (Appendix B.6.2). The name change was performed on the MR consensus topology as follows:

```
python NameChange.py Orders_Names.txt MR_tree.tre MR_Orders.tre
```

Monophyletic groups that contained species from the same order were condensed into a single leaf and the trees were rooted at the chosen outgroup. In the following example the Aves clade is the outgroup and the Newick Utilities (Junier and Zdobnov 2010) command is:

```
nw_condense MR_Orders.tre | nw_reroot - Aves > Figure_ready.tre
```

The trees generated using this command were used to display in a more legible form the topological information obtained from the dataset.

3.3 Results

3.3.1 Are mitochondrial data suitable for the resolution of the mammal phylogeny ?

Thirteen mitochondrial protein-coding genes were downloaded from the UniProtKB database (UniProt 2012). A total of 455 taxa had at least 2 sequences out of 13 mtGenes in the dataset (full list of genes and taxa are available in Appendix B.1). Taxa were sampled across 19 Placental Orders, and their allocation is shown in Figure 3.2(A). The 13 genes ranged in length from 71 aa to 626 aa and in coverage from 94 to 281 taxa. The phylogenetic conflict in these datasets was assessed using Likelihood Mapping (LM) (Strimmer and von Haeseler 1997), which gives a prior indication of tree-likeness. Alignments with a low percentage of phylogenetic conflict (<10%) were expected to give reasonably well supported bifurcating trees while datasets with a high proportion of phylogenetic conflict (>10%) were expected to produce less well-resolved nodes.

The results for the LM analysis are detailed in Table 3.2. In total, there are 11/13 mtGenes with greater than 10% phylogenetic conflict. The 2 genes that had less than 10% conflict were ND4 (9.7% conflict) and ND5 (8.1% conflict). Genes were concatenated to form a Supermatrix (SM) consisting of 3,906 aa and 455 taxa. The SM had too many taxa to be tested with the LM approach (as mentioned in the materials and methods section 2.2.9). Phylogeny reconstruction was carried out across all 13 mtGenes and the SM dataset in a ML framework using RAxML (Stamatakis 2006). The mtMAM+4Γ model of evolution was applied to 12/13 mtGene datasets and the SM dataset. The additional parameter (+F) which refers to amino acid frequencies was used for the alignment of the ND5 gene and for the SM dataset (see section 2.2.4.1 for parameters in ModelGenerator v85 (Keane et al. 2006)). The JTT+4Γ+F model was used for the analysis of the ND6 alignment. LnL scores are listed in Table 3.1 and associated topologies are given in Appendix B.2.

Table 3.2 Levels of phylogenetic conflict in mitochondrial data.

Data Names	Quartets							Conflict [4-7]
	1	2	3	4	5	6	7	
ATP6	27.80	26.70	27.70	2.20	2.00	2.00	11.60	17.80
ATP8	20.80	21.60	21.20	3.20	3.10	3.40	26.90	36.60
CO1	27.30	27.10	27.20	1.90	1.80	2.00	12.80	18.50
CO2	27.90	26.60	26.10	1.90	1.80	1.70	14.10	19.50
CO3	28.80	28.30	28.60	2.10	2.00	2.00	8.20	14.30
CYTB	29.80	29.00	29.00	2.00	2.30	2.20	5.70	12.20
ND1	28.80	28.60	28.80	1.60	1.80	1.60	9.00	14.00
ND2	28.90	29.70	29.10	1.60	1.50	1.70	7.60	12.40
ND3	24.80	25.20	24.60	1.90	2.00	1.90	19.70	25.50
ND4	29.90	30.10	30.10	1.40	1.50	1.20	5.60	9.70
ND4L	24.90	25.40	24.60	2.00	2.00	2.00	19.20	25.20
ND5	30.20	31.10	30.60	1.30	1.30	1.30	4.20	8.10
ND6	27.10	26.60	28.00	1.80	2.00	1.90	12.70	18.40
SM	NA	NA	NA	NA	NA	NA	NA	NA

Each of the 13 mtGenes is listed in the left column and their LM scores for each of the 7 quartets. The column on the left is the sum of quartets 4 to 7. This numbering scheme is as per Figure 2.7 of materials and methods section 2.2.8.

The resultant phylogenies from both the individual gene analyses and SM dataset contained large numbers of weak and un-supported nodes. Congruence between majority rule consensus topologies was assessed using Robinson-Foulds (RF) distances as implemented in the Clann software (Creevey and McInerney 2005). All scores are detailed in Table 3.3. These results show that the topology obtained from the ND5 gene was the closest to the topology obtained using the SM dataset, with a RF distance of 0.1301. The two genes used in the Barcode of Life project (Hebert et al. 2003, Borisenko et al. 2008), CYTB and CO1, manifested RF distances to the SM dataset of 0.2140 and 0.2609 respectively and had an RF distance of 0.2021 to one another. In Chapter 2 and in previous independent studies, placental mammals have been placed

into four Superorders (Hallstrom and Janke 2010, Murphy et al. 2001a, Prasad et al. 2008), as shown in Figure 3.2(A). In this analysis it was observed that none of the datasets generated from mitochondrial data, including the SM dataset (see Figure 3.2(B), were able to resolve these four Superorders.

Table 3.3 Robinson-Foulds distances between Phylogenies generated using untreated mtGene data

mtGene and SM Topologies	ATP6	ATP8	CO1	CO2	CO3	CYTB	ND1	ND2	ND3	ND4	ND4L	ND5	ND6
ATP8	0.1978												
CO1	0.2191	0.1867											
CO2	0.1879	0.1568	0.1977										
CO3	0.2246	0.1953	0.2414	0.2254									
CYTB	0.2000	0.2219	0.2021	0.1942	0.2368								
ND1	0.1918	0.2000	0.1923	0.1792	0.1667	0.2438							
ND2	0.2658	0.3034	0.2845	0.2672	0.2679	0.2449	0.2161						
ND3	0.1594	0.1375	0.1522	0.1354	0.1333	0.1951	0.1651	0.2642					
ND4	0.2308	0.2644	0.2414	0.2750	0.2321	0.2241	0.2107	0.2047	0.2192				
ND4L	0.1711	0.1855	0.1591	0.1720	0.2045	0.1824	0.1694	0.2500	0.1360	0.2196			
ND5	0.2808	0.3101	0.2800	0.2843	0.2449	0.2593	0.2542	0.2094	0.2756	0.1884	0.2711		
ND6	0.1628	0.1765	0.1750	0.0897	0.1923	0.1518	0.1778	0.2647	0.1772	0.2191	0.1500	0.2472	
SM	0.2040	0.2392	0.2609	0.2570	0.2930	0.2140	0.2460	0.1611	0.2609	0.1750	0.2778	0.1301	0.2692

The RF distance were calculated as described in section 3.2.7 and are shown for all comparisons of mtGene and SM topologies.

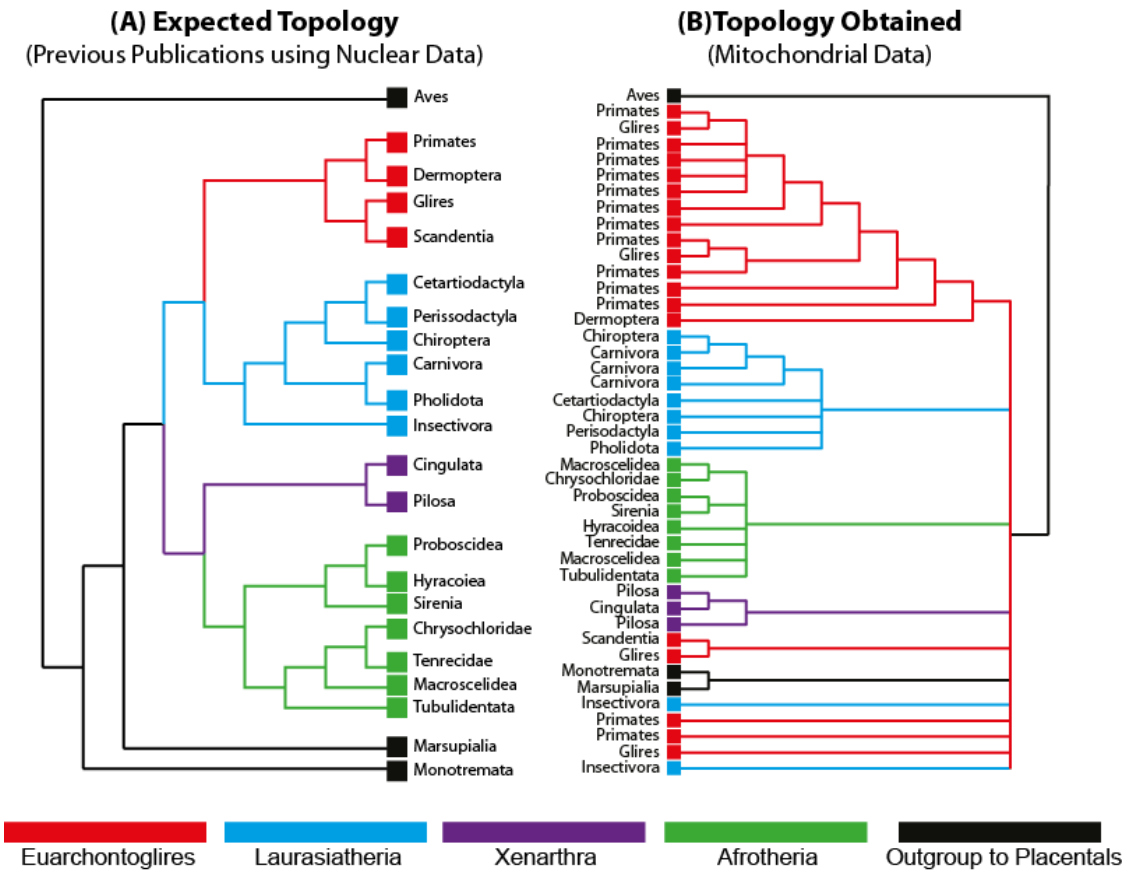


Figure 3.2 Phylogeny inferred from nuclear and mitochondrial data.

The phylogeny obtained using (A) nuclear data and (B) mitochondrial data is shown. The Superorders of the placental mammals are colour coded according to the following scheme: The Euarchontoglires are in red, the Laurasiatheria in blue, the Afrotheria in green and the Xenarthra in purple.

3.3.2 Treatment of the data to reduce phylogenetic conflict

MtDNA accumulates mutations more rapidly than nuclear data, and therefore is more likely to have both saturation and homoplasy (Brown et al. 1982, Rubinoff and Holland 2005), both of which contribute to phylogenetic conflict and have resulted in inconsistencies in phylogenies generated in the literature between different datasets (nuclear and mitochondrial) (Caterino et al. 2001, Reed and Sperling 1999, Rokas and Carroll 2008). In an effort to reduce phylogenetic conflict, improve node support and improve upon congruence between mtGene topologies a number of issues were addressed. First, the phylogenetic conflict was assessed to see if it decreased with a reduction in taxon numbers. Then tests were performed to assess whether phylogenetic signal is stronger when gene coverage across taxa is highee. The impact of the removal removal of saturated sites and whether the removal of rogue taxa reduced the phylogenetic conflict in the dataset was explored. Finally, and assessment of whether phylogenetic signal improves if shallower nodes are analysed was addressed. To answer each of these questions the original datasets listed in Table 3.2 were subjected to a series of treatments, which have been detailed in sections 3.3.2.1 to 3.3.2.5.

3.3.2.1 Does the phylogenetic conflict decrease with a reduction in the number of taxa?

It has been debated whether more sequence data or more thorough sampling improves phylogeny reconstruction (Hedtke et al. 2006, Rosenberg and Kumar 2003, Hillis et al. 2003, Pollock et al. 2002, Rosenberg and Kumar 2001). To test the impact that reduced taxon sampling has on phylogenetic signal, a subset of taxa were sampled (between 9 and 13 species) for each of the mtGenes. In each case a representative from each placental mammal Superorder was retained in the dataset. The reduced taxon datasets were re-aligned and re-tested for phylogenetic conflict using LM (Schmidt et al. 2002). From this analysis, it was observed that no individual gene dataset showed a significant reduction in phylogenetic conflict. More specifically, conflict increased in 12 out of 13 mtGenes, the exception was CO1 that manifested a small reduction from 18.5% to 17.3% conflict, the complete set of conflict scores are given in Appendix B.3. The SM dataset showed extremely low phylogenetic conflict at 3.4%, which was the lowest of all datasets tested. Phylogenetic reconstruction of the SM dataset was expected to result in four placental Superorders being resolved (as was found using nuclear data in Chapter 2) and that platypus would be positioned at the base of the mammal tree (van

Rheede et al. 2006). However, there were only low levels of support for the four placental Superorders and there was 97% bootstrap support for a relationship joining Opossum and Platypus as sister taxa to the exclusion of all other mammals. While LM statistics indicated low proportions of conflict, the test is strongly influenced by sequence length (Strimmer and von Haeseler 1997). Longer sequences (e.g. concatenated alignments) increase the number of usable characters and that has been shown to overcome the phylogenetic inconsistencies of individual gene data (Gadagkar et al. 2005). The RF distance between individual gene trees and the topology of the SM tree increased in 10/13 mtGenes, but decreased slightly in ND1, ND4 and ND6 to 0.1500, 0.1500 and 0.2143 respectively. The shared taxa between COX and CYTB were topologically identical as were the shared taxa between ND3 and ATP6, with RF distances of 0. Regardless of restricted sampling from the Superorders, the data were still unable to provide support for the placement of four placental mammal Superorders. Therefore, it was concluded that reduction in taxon sampling from the mtGene data did not remove phylogenetic conflict or improve phylogenetic resolution and the phylogenetic inconsistencies may have been a result of missing data that is discussed in the following section.

3.3.2.2 Is phylogenetic signal stronger when gene coverage across taxa is higher?

MtGenes have been sequenced to varying extents across placental mammals, and only 25 taxa have been sequenced for all 13 mtGenes. Congruence between phylogenies is an indication of how much error is contained in each phylogeny (Pisani et al. 2007). It is desirable to have a higher number of overlapping taxa between mtGenes to assess congruence between the phylogenies. Missing sequence data has been shown to cause problems in phylogeny reconstruction (Lemmon et al. 2009, Kearney 2002), however if enough phylogenetically informative characters are available then missing sequence data does not impact accurate phylogeny reconstruction (Wiens 2003, Philippe et al. 2004). Consequently, the next step in this analysis was to determine the impact of increasing gene coverage across the data by reducing the poorly represented taxa across the 13 mtGenes. In total there were 455 taxa with at least 2 out of the 13 mtGenes sequenced. The approach was to increase the gene coverage by increasing gradually from 2 to 13 genes, and generating datasets at each step of the process (consequently the taxon number will decrease at each step). The SM dataset and the individual mtGene

datasets were generated from each of these steps. Please refer to Figure 3.3 for an example of how this treatment was applied to the ATP6 mtGene.

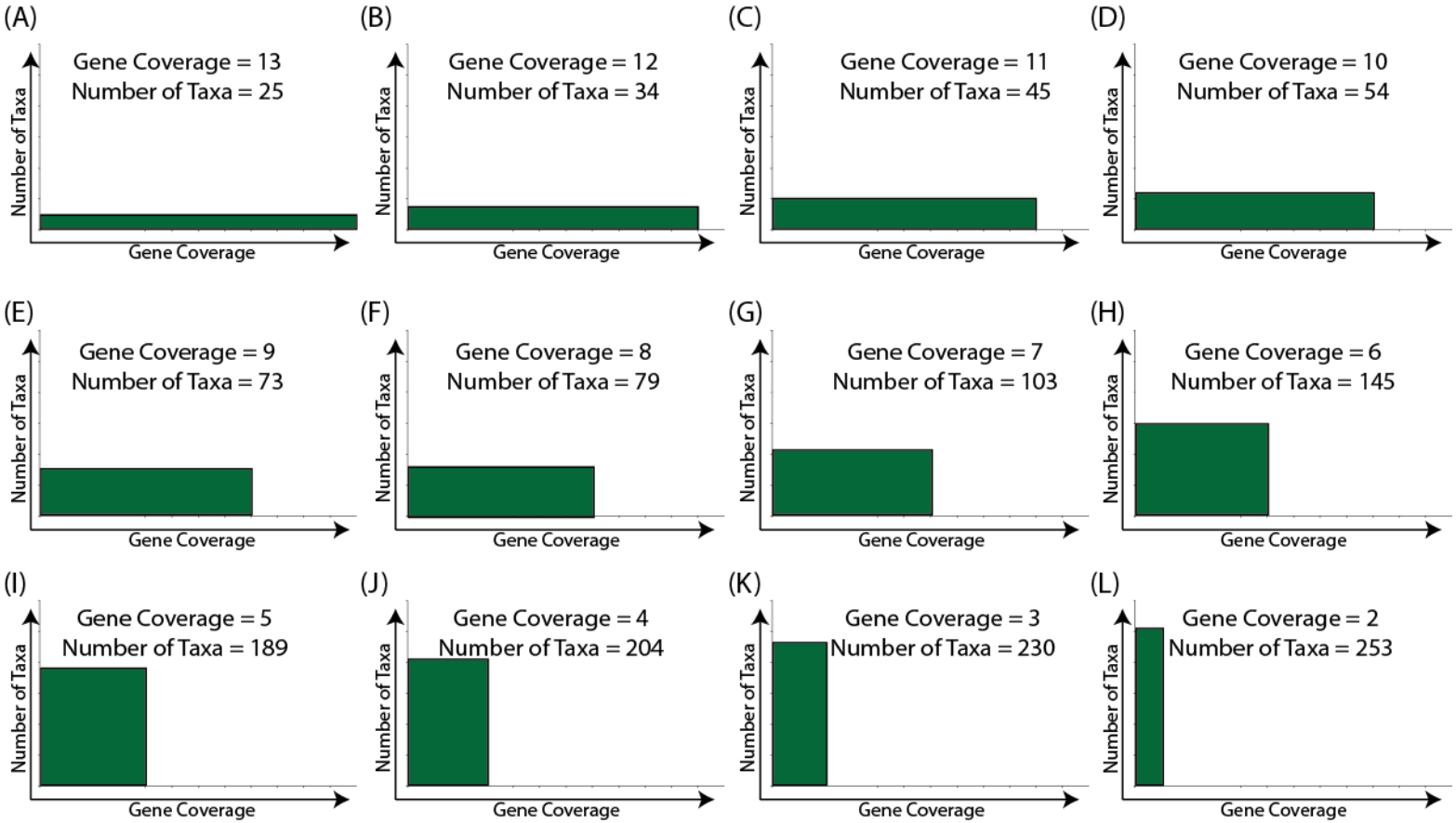


Figure 3.3 Relationship between gene coverage and number of taxa, illustrated here for the ATP6 mtGene.

Dataset treatments for stepwise decrease of gene coverage and increase of taxon number are shown (A to L). The gene coverage is given on the x- axis and the number taxa in each dataset on the y-axis. The green boxes represent the proportion of taxa number to gene coverage in each dataset.

LM (Schmidt et al. 2002) was employed to test the change in phylogenetic signal as gene coverage was increased, the results are shown in Figure 3.4. Phylogenetic conflict remained extremely high in ATP6, ATP8, CO1, CO2, CO3, ND3, ND4L and ND6 across all datasets regardless of gene coverage. ND1 showed variable phylogenetic conflict (12.2-14.9%) across the different levels of gene coverage but failed to reach our pre-defined cut-off value of <10% conflict. CYTB, ND2, and ND5 showed <10% phylogenetic conflict under the highest gene coverage and lowest taxon coverage conditions (see Figure 3.3). ND5 maintained reasonably low phylogenetic conflict across all gene coverage situations (5.8-8.6% conflict). The RF distance was calculated between ND5 gene topologies and topologies from other mtGene and the SM dataset to assess if congruence between gene trees improved at any coverage point (see Table 3.4). It was expected that if the datasets had more taxa in common (e.g. at higher gene coverage levels), then the topological distance between gene topologies would be smaller. The RF distances showed that when gene coverage was at its lowest (2 mtGenes) then the ND5 gene had the closest RF distance between 7 other mtGenes (CO1, CO2, CO3, ND1, ND4, ND6 and SM topologies). Therefore maximising the gene coverage across genes to improve congruence in these data does not have the expected effect. Only the Glires, Carnivora and Cetartiodactlya are represented in the 13 gene set, and so resolution for other clades is not possible. The RF distances were calculated between all pairs of resulting topologies at each coverage point and are available in Appendix B.5.

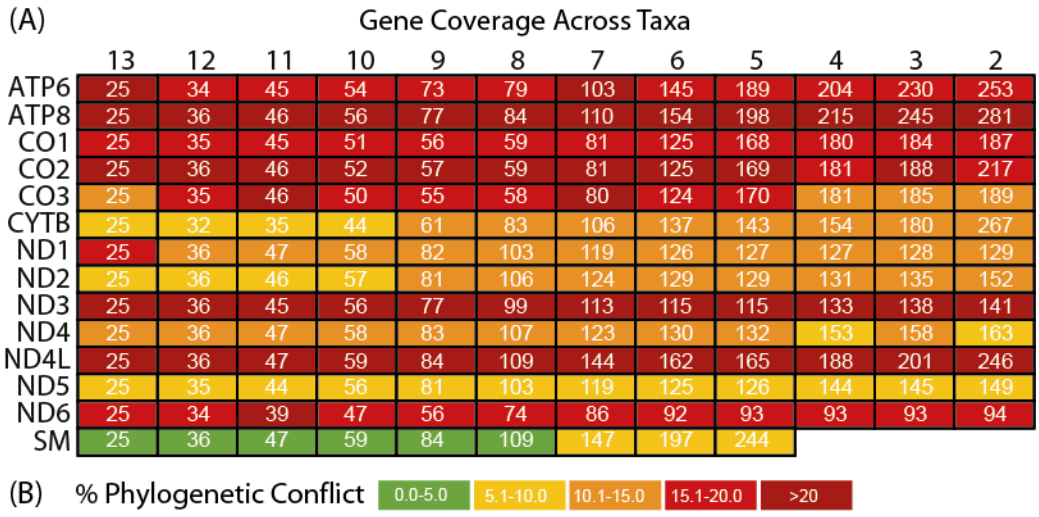


Figure 3.4 The impact of Gene Coverage versus Taxon Sampling on phylogenetic Signal

The rows represent datasets generated from the individual mtGenes and the SM dataset. The columns represent gene coverage across taxa from 13 to 2 genes for each dataset and the numbers in each cell represent the number of taxa in a given dataset. The percentage of phylogenetic conflict is colour coded as shown in (B) from acceptable levels (<10% conflict) represented by pale yellow and green, to unacceptable levels (≥10% conflict) represented by orange and red.

Table 3.4 Robinson-Foulds distances between ND5 phylogeny and all mtGene and SM phylogenys generated at each gene coverage point.

Gene Coverage	ATP6	ATP8	CO1	CO2	CO3	CYTB	ND1	ND2	ND3	ND4	ND4L	ND6	SM
2	0.2808	0.3101	**0.2800**	**0.2843**	**0.2449**	0.2593	**0.2542**	0.2094	0.2756	**0.1884**	**0.2711**	**0.2472**	**0.1301**
3	0.2808	0.2975	0.3000	0.3039	0.2551	0.2658	0.2625	0.2198	0.2874	0.1957	0.2887	0.3167	0.1585
4	0.2740	0.3038	0.3000	0.2941	0.2551	0.2532	0.2708	0.2198	0.2817	0.1957	0.2872	0.2841	0.1809
5	0.2778	**0.2911**	0.3000	0.2941	0.2653	**0.2436**	0.2667	0.2284	**0.2752**	0.2025	0.2846	0.2727	0.1707
6	0.2847	0.2975	0.3000	0.2941	0.2755	0.2500	0.2647	0.2241	**0.2752**	0.2167	0.2869	0.2759	0.1885
7	**0.2708**	0.2975	**0.2800**	0.3235	0.2755	0.2564	0.3009	0.2143	0.2944	0.2544	0.3233	0.3086	0.1897
8	0.3143	0.3267	0.3100	0.3529	0.3367	0.3224	0.3402	0.1907	0.3011	0.2323	0.3150	0.2826	0.1750
9	0.3060	0.3451	0.3200	0.3333	0.3469	0.3393	0.3421	**0.1867**	0.3239	0.2403	0.3205	0.3725	0.1795
10	0.3333	0.3400	0.3667	0.3478	0.3182	0.2692	0.3269	0.2059	0.3200	0.2788	0.2830	0.2976	0.1698
11	0.3590	0.3875	0.3333	0.3750	0.3250	0.3167	0.3659	0.2125	0.3077	0.2317	0.3049	0.3382	0.2317
12	0.3333	0.3594	0.3226	0.3750	0.3065	0.3214	0.3594	0.2500	0.2813	0.2813	0.2500	0.2833	0.2500
13	0.3636	0.3636	0.3636	0.3636	0.3409	0.2955	0.3409	0.2500	0.3182	0.3409	0.2727	0.3182	0.1364

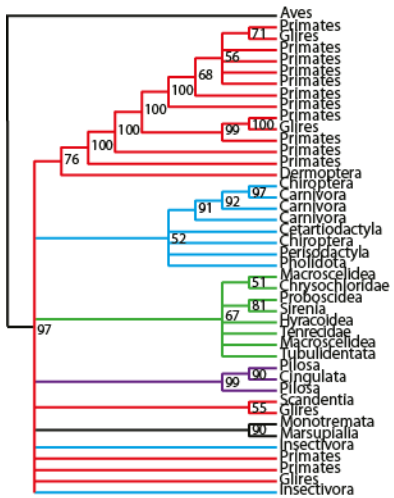
The gene coverage that gave the closest distance between the ND5 topologies and the mtGene or SM topologies are highlighted in red.

Upon examination of the SM dataset, there was a notable trend towards a decrease in phylogenetic conflict, from 7.2% to 1.1%, as gene coverage increased and taxa number decreased, see Figure 3.5. This suggests that there was more phylogenetic signal available when analysing all 13 mitochondrial genes together, rather than examining them individually (this point will be raised again in the discussion of this Chapter).

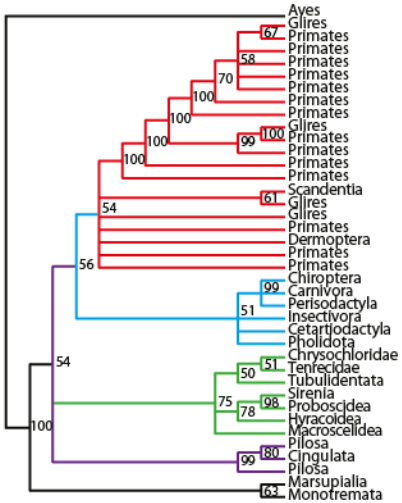
To test the quality of the phylogenetic signal, ML trees were drawn from the SM dataset across all gene coverage levels; detailed results are shown in Figure 3.5. The topologies do not reflect trends in LM tests, as improvement in node support is not observed with decrease in percentage of phylogenetic conflict. When gene coverage is between 2-4 genes, shown in Figure 3.5(A-C), there are multiple collapsed nodes with branch support of <50%, which is indicative of large proportions of phylogenetic conflict. Four clearly defined Superorders were observed when gene coverage was 4 and was between 6-9, with a range of 109 to 284 taxa, see Figure 3.5(C) and Figure 3.5(E-G).

The topology obtained from coverage across 10 genes and 59 taxa, Figure 3.5(I), has monophyletic groups for the Euarchontoglires and the Laurasiatheria, however the other placental orders are not resolved and Proboscidea node is also collapsed, showing clear incongruence with previous phylogenies (Meredith et al. 2011, Murphy et al. 2007). Topologies derived from coverage of 11 to 12 genes have placed the primates as a sister group to the Laurasiatheria with the exclusion of Glires a position that is incongruent with the majority of previous publications, see Figure 3.5(J-K). The topological distance between phylogenies for each mitochondrial gene and the SM dataset were calculated using RF distances at each level of gene coverage. It was found that there was no agreement between topologies (RF = 0.000) from individual mtGene or the SM datasets for the same gene coverage. While increase in gene coverage and decrease in missing data provided sufficient signal to resolve the 4 Superorders it was not possible to estimate phylogenetic trees with strong node support for intra-ordinal nodes using these data.

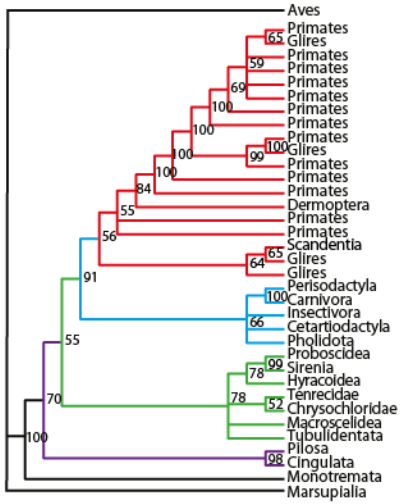
(A) 2 Genes, 455 Taxa



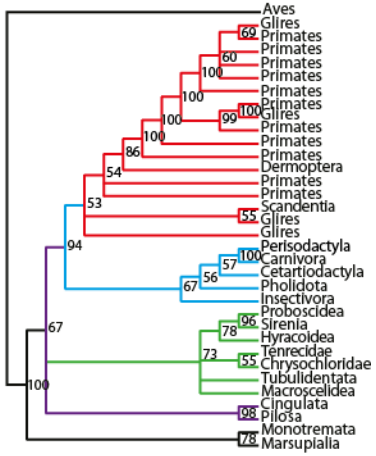
(B) 3 Genes, 326 Taxa



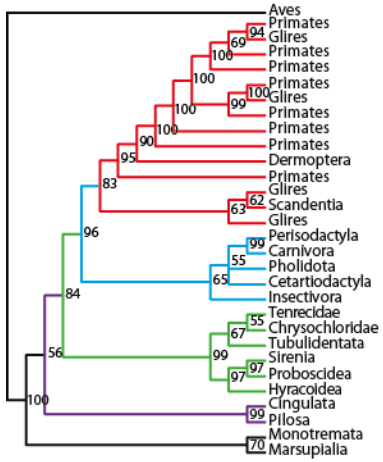
(C) 4 Genes, 284 Taxa



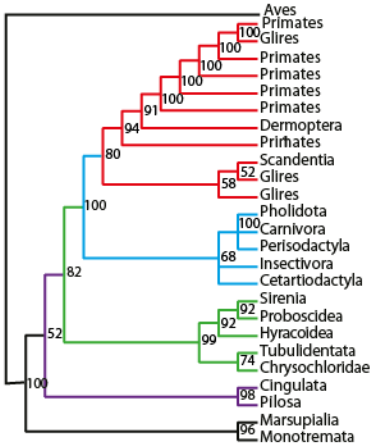
(D) 5 Genes, 244 Taxa



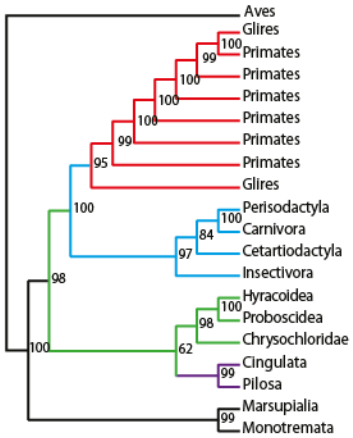
(E) 6 Genes, 197 Taxa



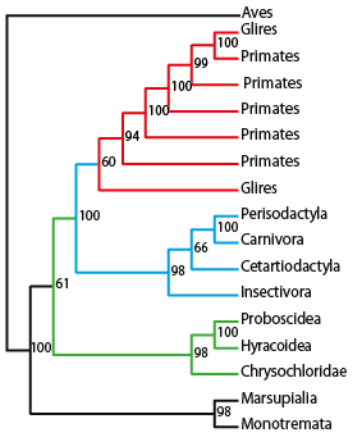
(F) 7 Genes,147 Taxa



(G) 8 Genes, 109 Taxa



(H) 9 Genes, 84 Taxa



Legend agrees with page 153

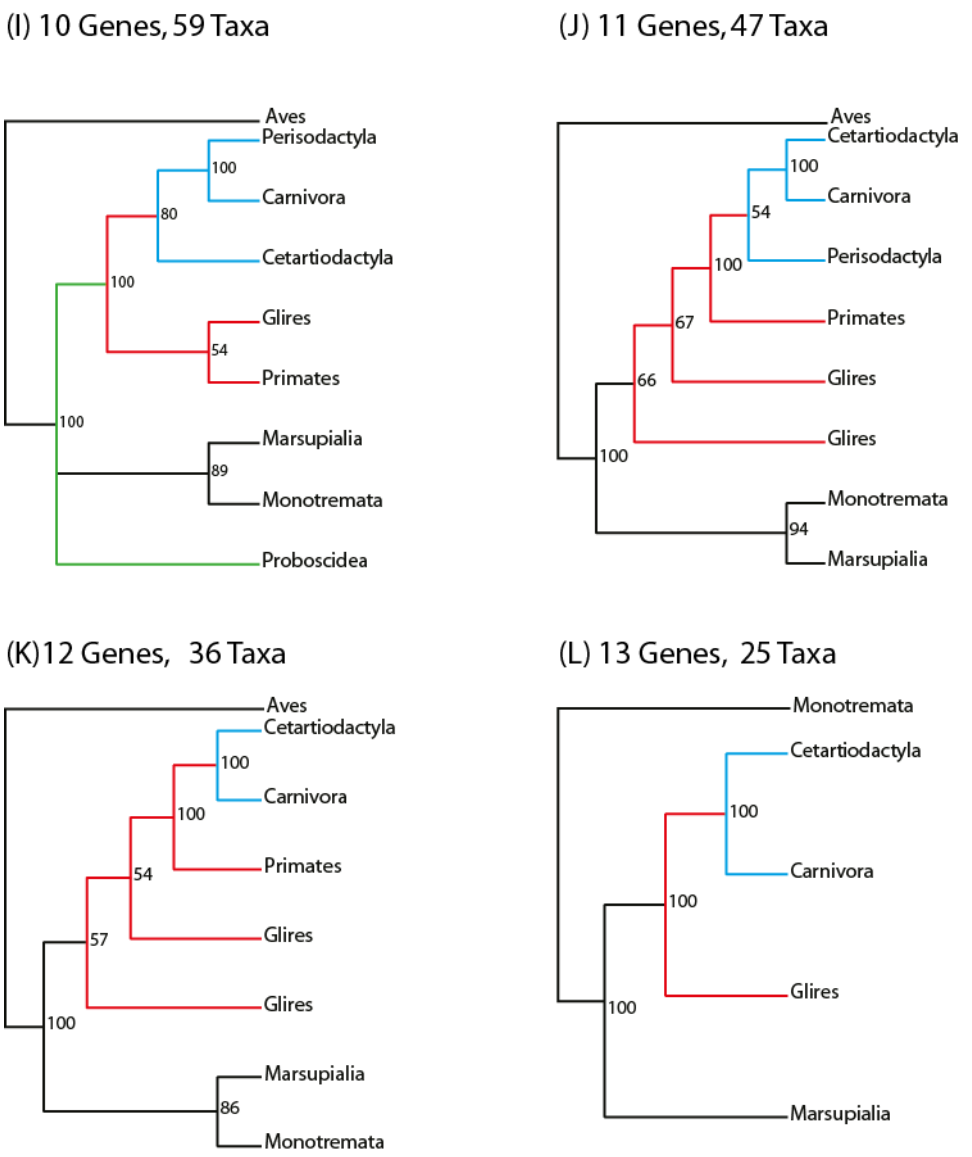


Figure 3.5 Phylogenetic Changes Observed in the analysis of mtGenes when gene coverage is increased and taxon coverage is decreased.

Topologies generated using RAxML and the SM dataset where gene coverage is increased incrementally from 2 genes in panel (A), to 13 genes in panel (L), while taxa number decreases from 455 taxa in panel (A), to 25 taxa in panel (L). Bootstrap values are shown for nodes with ≥ 50% support and nodes with bootstrap values <50% are collapsed.

3.3.2.3 Does the removal of saturated sites reduce the amount of conflict in the dataset?

Mitochondrial datasets tend to have more saturation compared to nuclear datasets (Brown et al. 1982), and so an assessment of whether their removal reduced the amount of phylogenetic conflict and improved overall phylogenetic resolution was carried out. In an effort to identify and remove rapidly evolving or saturated sites from the data, the sites in the data were categorised based on their rates of evolution (Cummins and McInerney 2011) as outlined in materials and methods section 3.2.5. LM was performed at each stepwise reduction in alignment length, and the change in levels of phylogenetic conflict at each step are shown in Figure 3.6.

When the fastest site categories were removed (site category 20), a slight reduction in phylogenetic conflict was observed for ATP8 (36.6% to 35.4%) and ND5 (8.1% to 8.0%), but there was no change in phylogenetic conflict observed in the ATP6 gene (17.8%) for the same manipulation. The removal of site category 20 resulted in an increase in phylogenetic conflict for the remaining 10 mtGenes, suggesting that removal of site category 20 could be removing necessary phylogenetic signal. Subsequent removals of site categories, e.g. site categories [20 and 19] and site categories [20, 19 and 18], resulted in an increase in the phylogenetic conflict in all 13 mtGenes. Removal of site category 20 from the SM dataset reduced the concatenated alignment from 4329 aa to 882 aa. Unfortunately, this reduction in sequence length left an unsuitable amount of overlapping characters per taxa for phylogeny reconstruction to be conducted.

Phylogenies were generated at each step for the individual mtGene datasets. However, as the fast evolving site categories were stepwise removed, the number of bifurcating nodes reduced and of the number of polytomies increased. For example, the ND5 gene (Figure 3.7), contained 149 taxa and had 257 nodes in its consensus tree from untreated data, this reduced to 237 nodes with the fastest site category (category [20]) removed, 213 nodes with site category [20 and 19] removed, and 199 nodes with site categories [20, 19 and 18] removed. The progressive breakdown of resolution, as illustrated in Figure 3.7, indicates that removal of fast evolving sites from the mtGene alignments does not improve the resolution for these data. For a complete list of all phylogenies, please refer to Appendix B.2.

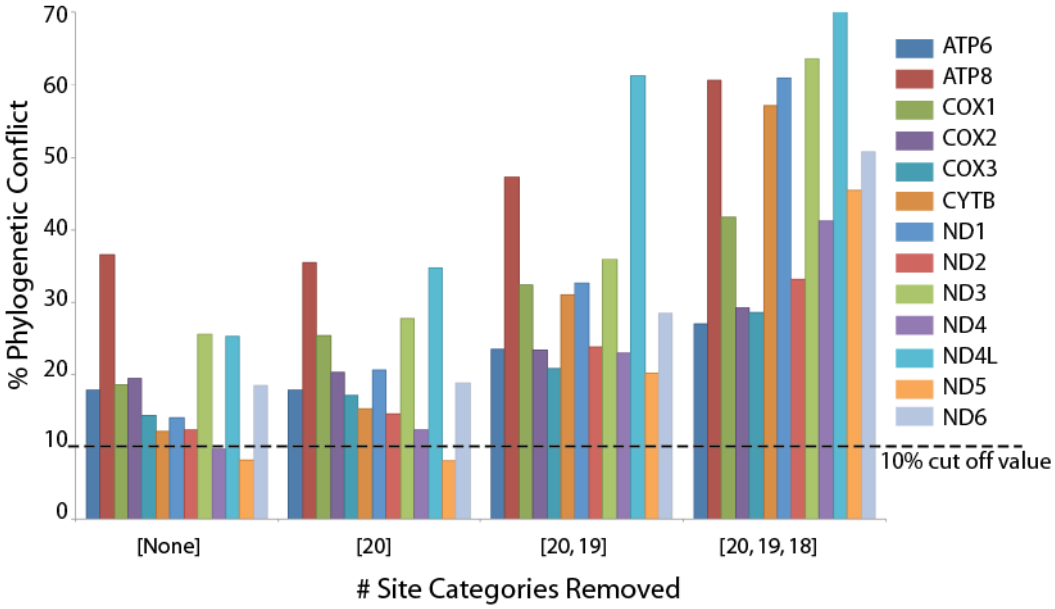


Figure 3.6 The impact of removal of fast evolving site categories on phylogenetic signal.

The mtGenes are colour coded as described in the key on the right hand side. The x-axis shows the mtGene datasets tested, where no site categories were removed [None], the fastest site categories were removed [20], site categories 20 and 19 removed [20,19] and site categories 20, 19 and 18 removed [20,19,18]. The y-axis shows the level of phylogenetic conflict which a black horizontal bar marking the 10% cut-off value.

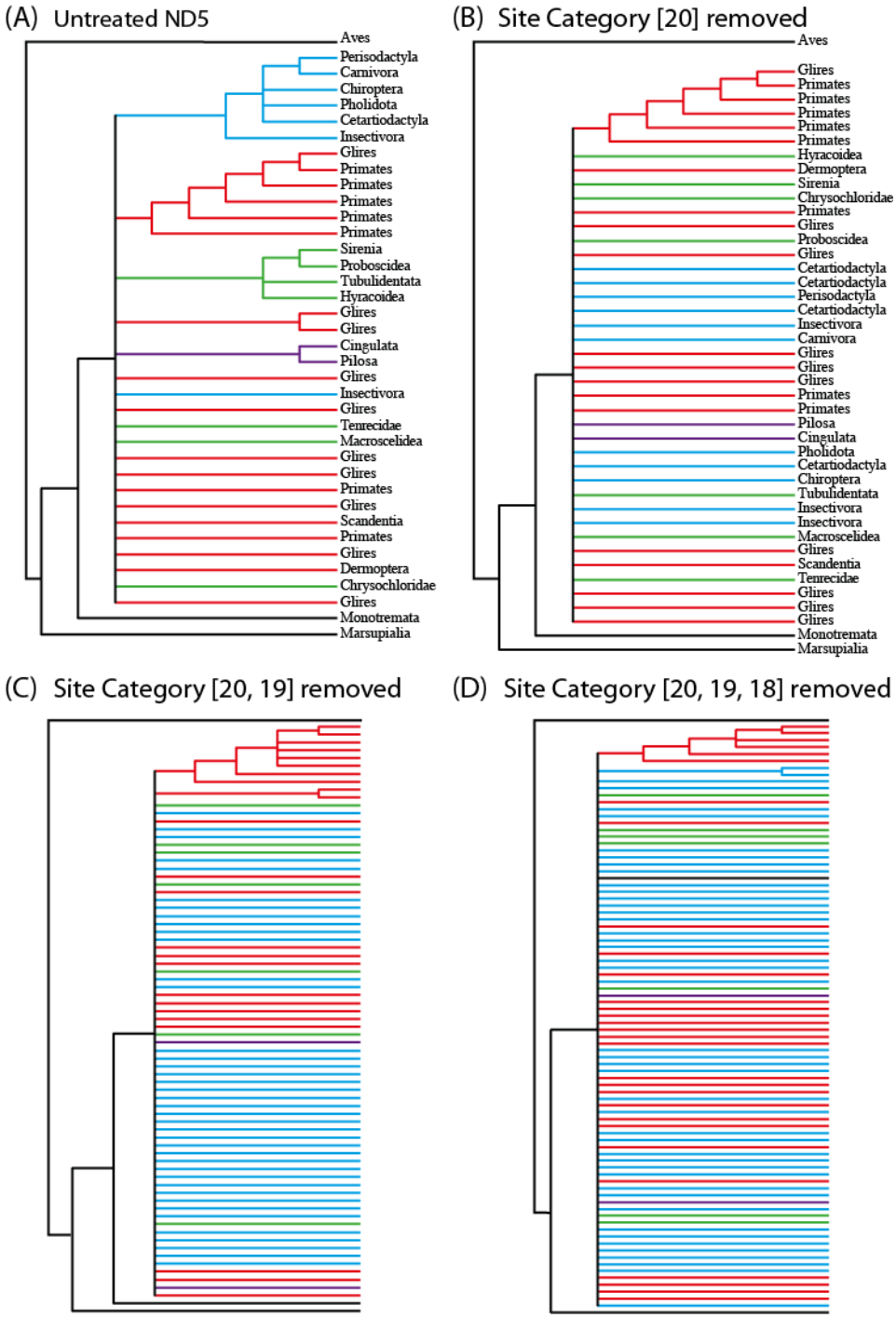


Figure 3.7 The topological effect of Site Stripping ND5 dataset

The majority rule consensus trees are shown for ND5 gene from (A) untreated data, (B) site categories [20] removed, (C) site categories [20 and 19] removed, and (D) site categories [20,19 and 18] removed. The Order names are given on the leaves of the phylogeny (A & B) along with colour codes for Superorders that were used in Figure 3.2. The Order names are not given for C and D as there were too many, therefore branches and colour codes serve only as an assessment of phylogenetic resolution.

In this part of the study the fast evolving sites in each mtGene have been identified and removed sequentially from the MSAs. This was done in an effort to remove saturated positions and improve overall phylogenetic signal. It was observed that removing these sites had a negative impact on the phylogenetic signal. Previous studies have indicated that mitochondrial genes saturate faster than nuclear genes (Brown et al. 1982), and it has been suggested that at deep nodes these fast evolving sites can contain large amounts of homoplasy (Caterino et al. 2001, Reed and Sperling 1999). Therefore, it is likely that these data and methods may not have been suitable to accurately identify the saturated sites and improve overall phylogenetic resolution.

3.3.2.4 Does the removal of "rogue taxa" reduce the phylogenetic conflict in the dataset?

"Rogue taxa" are those whose phylogenetic placement varies from dataset to dataset or from gene to gene. This type of conflict can arise due to missing data, high rates of mutation or indeed very low rates of mutation (Sanderson 2002). By removing these rogue taxa, it is expected that the conflict in a dataset would be reduced. The identification of rogue taxa involves removing placental mammal species from the alignment whose percentage identity score is less similar to the remainder of the MSA than it is to the outgroups. This is an *ad hoc* approach but with the large number of taxa present it is hoped that clades will retain representatives.

The largest number of rogue taxa identified were from the ATP6, ATP8, CO1, CO2, CO3 and CYTB datasets reporting between 22 and 61 rogue taxa. When removed from the MSA the result was a change in percentage identity score between 0.014 and 0.071, see Table 3.5. Far fewer rogue taxa were identified in the NADH genes (between 0 and 6 taxa), and, the change in percentage identity score for this gene was under 0.008, see Table 3.5.

Table 3.5 Identification of Rogue Taxa using Mean Identity Scores

Orders	ATP6	ATP8	CO1	CO2	CO3	CYTB	ND1	ND2	ND3	ND4	ND4L	ND5	ND6	SM
Carnivora	0	0	0	0	1	0	0	0	0	0	0	0	0	38
Cetartiodactyla	0	1	0	0	0	0	0	0	1	0	0	0	0	68
Chiroptera	0	0	0	0	0	1	0	0	0	0	0	0	0	27
Chrysochloridae	0	0	0	0	0	0	0	0	0	0	0	0	0	2
Cingulata	0	0	0	0	0	0	0	0	0	0	0	0	0	1
Dermoptera	0	0	0	0	0	1	0	0	0	0	0	0	0	2
Glires	1	3	0	2	2	5	3	3	3	0	0	3	2	58
Hyracoidea	0	0	0	1	0	0	0	0	0	0	0	0	0	2
Insectivora	0	0	1	1	0	0	0	0	0	0	1	1	0	20
Macroscelidea	0	0	0	1	0	0	0	0	0	0	0	0	0	2
Perisodactyla	0	0	0	0	0	0	0	0	0	0	0	0	0	2
Pholidota	0	0	0	0	0	0	0	0	0	0	0	0	0	1
Pilosa	0	0	0	0	0	1	0	0	0	0	0	0	0	3
Primates	19	39	21	52	15	17	3	0	1	1	0	0	0	97
Proboscidea	4	0	4	4	4	0	0	0	0	0	0	0	0	4
Scandentia	0	0	0	0	0	0	0	0	0	0	0	0	0	1
Sirenia	0	0	0	0	0	0	0	0	0	0	0	0	0	2
Tenrecidae	0	0	0	0	0	0	0	0	0	0	0	0	0	1
Tubulidentata	0	0	0	0	0	0	0	0	0	0	0	0	0	1
TOTAL	24/253	43/281	26/187	61/217	22/189	25/267	6/129	3/152	4/141	1/163	1/246	4/149	2/94	326/455
mean_id	0.797	0.592	0.938	0.845	0.881	0.841	0.824	0.667	0.760	0.777	0.736	0.719	0.657	0.647
mean_id (-rogue)	0.818	0.637	0.949	0.916	0.897	0.855	0.832	0.673	0.765	0.779	0.737	0.723	0.665	0.869

Each of the Orders listed in the first column were tested for the presence of rogue taxa across mtGenes and the SM dataset. The entries in each cell represent the number of rogue taxa identified, the total for each gene is also shown. The mean identification score for each gene before and after rogue taxa are removed is displayed.

The number of rogue taxa identified in the SM dataset was 326/455, which exceeds that identified on a gene-by-gene basis (221/455). These results show that the concatenation of all the genes and the use of a supermatrix approach could in fact introduce more conflict in the data.

The Orders where the majority of rogue taxa were identified were the Primates and the Proboscidea. There were 104 primates sampled in this study, in the SM dataset 94/104 primates were identified as rogue. Between 24.6% and 80.8% of the rogue taxa identified in ATP6, ATP8, CO1, CO2, CO3 and CYTB were from the Primates. There were only 4 species sampled from the Proboscidea Order and in ATP6, CO1, CO2, CO3 and the SM datasets, all 4 Proboscidea where identified as rogue. All identified rogue taxa were removed and the subsequent datasets were re-aligned and re-tested for phylogenetic signal using LM (Schmidt et al. 2002). A decrease in phylogenetic conflict upon removal of rogue taxa was observed in the following genes: ATP8 (Δ 1.0 %), COX2 (Δ 1.4 %), ND1 (Δ 0.2%), ND4L (Δ 0.4%) and ND5 (Δ 0.5%), where delta (Δ) indicates the change in phylogenetic conflict. Only ND5 passed the 10% phylogenetic conflict cut-off value. The SM dataset had 8.0% phylogenetic conflict. All resultant topologies were incongruent with the Superorder placement of mammals, as shown in Figure 3.2(A). RF distances showed that there was no agreement between any topology when compared.

Although there were several rogue taxa identified, only a marginal decrease in phylogenetic conflict was observed and there was no overall improvement in resolution of topology. These findings lead to the conclusion that rogue taxa are not the primary contributors to mixed phylogenetic signal in mtGenes.

3.3.2.5 Does the phylogenetic signal improve at more shallow phylogenetic depths?

Previous studies have shown that large levels of homoplasy are observed when sampling from deep nodes using mitochondrial data (Caterino et al. 2001, Reed and Sperling 1999). The aim of this segment of the analysis was to understand where the phylogenetic signal starts to degrade when using mtGenes. Groups of taxa were selected at different depths on the known species phylogeny (supported by the results from Chapter 2). Sampling taxa from various depths of divergence, the phylogenetic

signal was assessed to determine at what depth the phylogenetic signal from the mtGenes is lost. The various phylogenetic depths tested are shown in Figure 3.8(A) and were as follows: the Boreoeutheria and the Atlantogenata nodes, the four Superorder nodes (Euarchontoglires, Laurasiatheria, Afrotheria and Xenarthra), the two major subgroups within the Euarchontoglies (Glires and Primates), and the 5 major subgroups within the Laurasiatheria (Carnivora; Cetartiodactyla; Perissodactyla; Chiroptera and Insectivora). The node leading to the Cetacea and the Ruminata were also tested, as were the nodes from three Suborders (Caniforma, Feliformia, Tylopda). The closest available species were chosen as outgroups for each dataset. Phylogenetic conflict was estimated from each dataset using LM (Schmidt et al. 2002) and all topologies were generated using RAxML (Stamatakis 2006). The LM results for phylogenetic conflict across the 13 mitochondrial genes varied depending on node depths, for summary of datasets which passed the cut-off mark see Figure 3.8 (for detailed LM results see Appendix B.3).

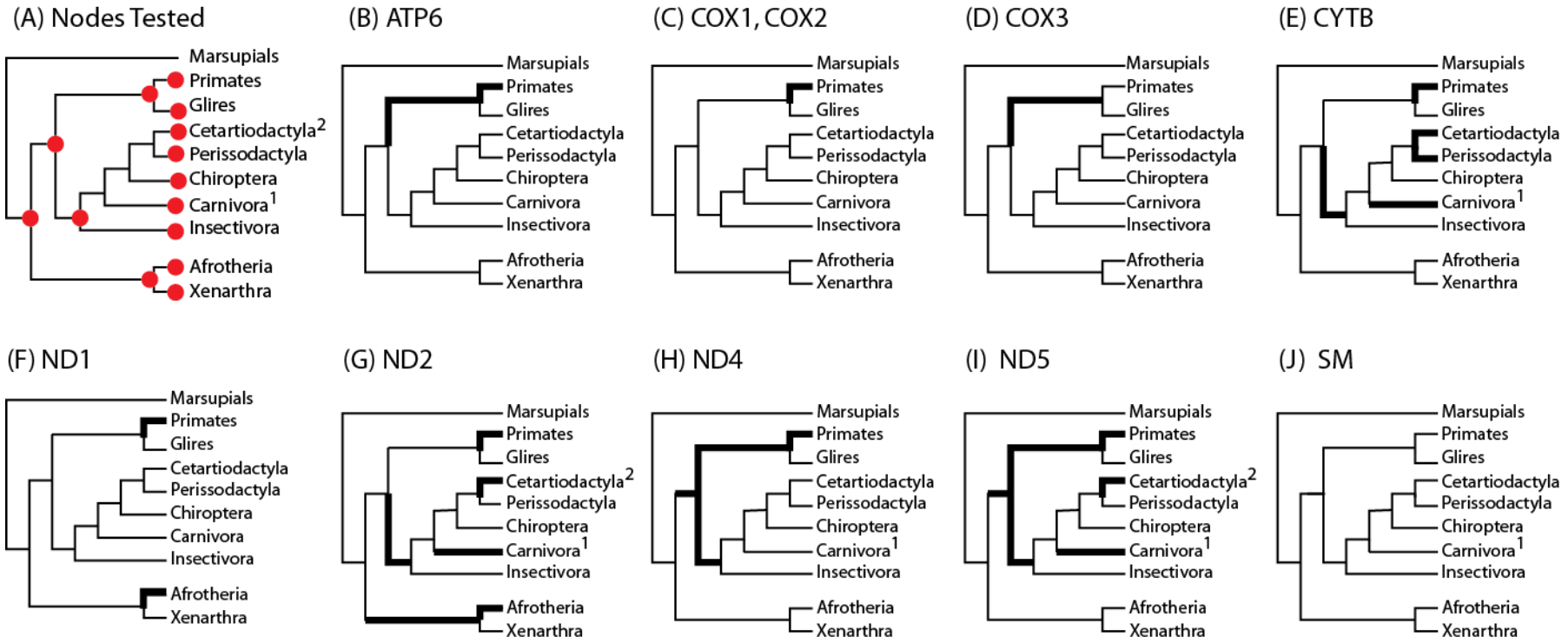


Figure 3.8 Assessing phylogenetic conflict from datasets sampled from different nodes on the known placental mammal phylogeny.

Panel (1) shows nodes that were tested in the analysis labelled with a red circle, each phylogenetic tree (B-I) represents the analysis of an mtGene as labelled, and (J) represents the Supermatrix (SM) dataset. The representative taxa used in each dataset (A-J) are the identical. The bolded lines represent either Superorders or Orders where the phylogenetic conflict was <10%. [1]Caniforma and [2]Cetacea denotes where these Orders within their Superorders also passed cut-off criteria of <10% phylogenetic conflict.

At the deepest diverging Boreoeutheria node, the mtGenes with <10% phylogenetic conflict were ND4 (8.2% conflict) and ND5 (6.10% conflict). Topologies obtained for ND4 and ND5 were more similar to each other (RF distance = 0.2042) than either was to the topologies of the remaining mtGenes. Furthermore, the topologies obtained for ND4 and ND5 were each more similar to the topology obtained for the SM dataset with RF distances of 0.1761 and 0.1445 respectively, than each other. This was an indication that ND4, ND5 and the SM dataset were starting to show more similarities than dissimilarities, and if this pattern continued it would indicate that at shallower nodes these genes may be suitable candidates for phylogeny reconstruction.

Moving to the next node on the tree, the ancestral Euarchontoglires node, there were 4 genes that had <10% phylogenetic conflict, these genes were ATP6 (8.5% conflict), COX3 (9.2% conflict), ND4 (8.9% conflict) and ND5 (3.5% conflict). The topologies resulting from these genes were incongruent with one another and the RF distances varied from 0.0056 to 0.3947.

There were 8 genes that had <10% phylogenetic conflict when sampled at the Primate node, these were as follows: ATP6 (8.3% conflict), COX1 (7.8% conflict), COX2 (7.4% conflict), CYTB (9.0% conflict), ND1 (9.6% conflict), ND2 (5.2% conflict), ND4 (5.8% conflict) and ND5 (1.9% conflict). The COX1 and ND3 topologies were completely congruent as were the COX2 and ND6 topologies, although these two sets of topologies were incongruent with one another due to a low level of common taxa between these two groups (only taxa in common are compared using the RF distance). The ND5 gene had the lowest proportion of conflict in its alignment and had the closest distance to the SM dataset (RF distance = 0.0882).

LM analysis on the Glires node showed large proportions of phylogenetic conflict ranging from 10.1% to 43.9%. There were 3 datasets where topologies produced RF distances of 0.0, these were: (i) COX2 and CYTB, (ii) COX2 and ND4L and (iii) ND2 and ND4L. It is important to note that in each of these gene trees there were several nodes with <50% bootstrap support therefore while there was agreement between bifurcating nodes, the overall phylogenetic signal for these genes is weak.

When sampling at the Laurasiatheria node, the four mtGenes that passed phylogenetic conflict were: CYTB (9.10% conflict), ND2 (9.5% conflict), ND4 (8.5% conflict) and

ND5 (4.3% conflict). There was no absolute agreement between the resultant topologies for any of the 4 mtGenes. Shallower nodes within the Laurasiatheria Superorder were also successful in achieving low levels of phylogenetic conflict such as the Cetartiodactyla node: CYTB (8.7% conflict), ND2 (9.8% conflict) and ND5 (5.6% conflict), the Perissodactyla node: CYTB (5.7% conflict) and the Carnivora node: CYTB (9.9% conflict), ND2 (7.9% conflict), and ND5 (4.1% conflict). The CYTB and ND2 gene sampled from the Chiroptera node were congruent while the topology for the CYTB and the topology for the SM datasets had a very low RF distance of 0.0652. When topologies were examined from the Insectivora node, the ATP8 and ND2 topology were in full topological agreement with that of the SM dataset. The shallow Cetacea node was tested for phylogenetic conflict across the mtGenes and it was found that ND2 and ND5 mtGenes had 9.10% and 6.3% phylogenetic conflict respectively. There were multiple topological similarities found when sampling was carried out within the *Cetacea* node. An RF distance of 0 was obtained when the COX1 topology was compared with ND1, ND2 and ND3, and RF distances of 0 was found for the cetacea node when ND6 was compared to ATP6 and COX2. MtGenes sampled from the *Caniforma* node showed that phylogenetic conflict was below the cut-off criterion of 10% specifically the conflict for CYTB = 7.7%, ND2 = 6.0%, ND4 = 9.0%, ND5 = 4.2% and the SM dataset = 9.7%. On comparison of the topologies for the mtGenes sampled from the *Caniforma* node, there was no agreement and RF distances ranged from 0.0870 to 0.4000.

The ND2 mtGene was the only gene that had sufficiently low phylogenetic conflict (7.8%) at the Atlantogenata node. This conflict decreased to 3.3% when the more recent daughter node for Afrotheria was sampled for the ND2 gene. Also at the Afrotheria node the ND1 gene contained 8.9% phylogenetic conflict. At the Atlantogenata node topological similarities were observed for the ND2 and COX1 datasets with RF distance = 0. This was seen again in the Afrotheria analysis where ND2 had RF distance = 0 when compared to both ATP6 and CO1. There were no datasets with <10% phylogenetic conflict when sampling at the Xenarthran node.

In summary, it was observed that there was less phylogenetic conflict when nodes were sampled from shallower depths on the known species tree. Phylogenetic conflict did not decrease uniformly from deep to shallow nodes. Taking the ND4 gene as an example, the phylogenetic conflict was as follows: Eutherian node (9.7% conflict), Boreoeutheria

node (8.2% conflict), Euarchontoglires node (8.9% conflict) and Primates (5.8% conflict). When the topology from the ND4 gene was sampled at the Eutherian node and compared with topologies generated from data sampled from shallower nodes, the distance between the trees varies as follows: Boreoeutheria node (RF distance = 0.0176), Euarchontoglires node (RF distance = 0.0294) and Primates (RF distance = 0.0405). Overall there are large amounts of variation in the topological findings and there is more discordance between the phylogenies from the mtGenes and the SuperMatrix datasets than there are topological agreements.

3.4 Discussion

Overall, a decrease in phylogenetic conflict is observed in mtGenes when sampling at shallower nodes. The reduction in phylogenetic conflict was not reflected in the resolution of congruent bifurcating phylogenies with high bootstrap support at nodes. Previous phylogenetic studies of mitochondrial data show that homoplasy is not as prevalent at shallower nodes (Caterino et al. 2001, Reed and Sperling 1999). Using these data and methods there was no conclusive evidence that the phylogenetic signal degrades in a clock like fashion towards deeper nodes. When mitochondrial data is employed phylogenetic conflict was present at both deep and shallow nodes. According to these analyses, none of the mtGenes were determined to be good candidates for phylogeny reconstruction. This includes CYTB and CO1 currently used in the bar code of life project (Hebert et al. 2003, Borisenko et al. 2008). The levels of homoplasy and saturation were too high in mtGenes from mammals to tease apart phylogenetic signal from phylogenetic noise.

The root of the placental mammal tree has been contested for the past number of years (Kriegs et al. 2006, Murphy et al. 2007, Murphy et al. 2001a) so it is unsurprising to see variations in the position of the Xenarthra and the Afrotheria at the base of the placental tree. The 4 Superorders of placental mammals are observed by multiple independent studies using nuclear data (Meredith et al. 2011, Murphy et al. 2001b, Hallstrom and Janke 2008), rare genomic change (Murphy et al. 2007), nuclear and mitochondrial data combined (Murphy et al. 2001a) and a study that used the entire mitochondria genome on 78 taxa (Kjer and Honeycutt 2007).

Initially the SM dataset appeared to have less phylogenetic conflict than the individual gene datasets, but the four well-defined Superorders were not supported. While longer sequence data has been shown to overcome phylogenetic inconsistencies of smaller datasets (Gadagkar et al. 2005, Gee 2003), this is not always the case. Previous large scales phylogenomic studies have still found phylogenetic inconsistencies regardless of implementation of a large Supermatrix dataset (Dunn et al. 2008, Philippe et al. 2009, Schierwater et al. 2009). Phylogenomic studies of mammals have attributed this to introgression of gene flow as a result of hybridization (Hallstrom and Janke 2008). The observations from Hallstrom and Janke (2008) were based on nuclear data. Introgression in mtGenes has been shown within species of mammals such as the *Canis* genus (Hailer and Leonard 2008) and full mitochondrial genome replacement has been shown within the Chiroptera Order (Berthier et al. 2006). It is possible that these evolutionary phenomena acting on mtGenes are negatively impacting the accurate resolution of the genealogical history of mammals.

There are conflicting opinions on the impact of missing data on phylogeny (Lemmon et al. 2009, Kearney 2002, Wiens 2003, Philippe et al. 2004). In this study small improvements were observed when increasing gene coverage across the SM dataset with regards to the placement of the Superorders but conflict was still observed at shallower nodes.

In this study, the removal of fast evolving sites from mtGene sequence data did not reduce the phylogenetic conflict and improve overall phylogeny resolution. When sequence samples were obtained from shallower nodes on the known mammal phylogeny, LM indicated that conflict in phylogenetic signal was within the criteria set (<10%), however the resulting topologies were incongruent with one another. Incongruence between mtGene phylogenies is an indicator of the level of error between two trees (Pisani et al. 2007) and as high levels of incongruence have been observed throughout this study whether the data as been treated or not, it does not increase our confidence in the application of mtGenes as a phylogenetic marker.

While congruence in phylogenies generated from mtGene data is important, so too is congruence between different data types such as nuclear sequences, morphological data and microRNAs (Pisani et al. 2007, Rota-Stabelli et al. 2011, Campbell et al. 2011, Branger et al. 2011). Once again, the mtGene data was unable to generate topologies

that agreed with previous studies (Meredith et al. 2011, Murphy et al. 2007, Shoshani et al. 1996), and differed in the resolution of the four Superorders and inter-ordinal placements.

Mitochondrial encoding genes have unusual properties compared to nuclear coding genes as they are predominantly membrane proteins and therefore require models to accommodate their unusual physiochemical properties (Lio and Goldman 1999). Mitochondrial DNA in vertebrates is tightly packed with no introns (Anderson et al. 1981), whereas nuclear genes contain large numbers of introns and intergenic regions (Lander et al. 2001, Venter et al. 2001) indicating that the mitochondrial genome is very highly conserved. These key differences between mitochondrial and nuclear genomes illustrate the importance of improved modeling to accommodate the unusual and distinct properties of mitochondrial genomes.

In this Chapter, I show that mitochondrial coding data is not suitable for resolving phylogenies at either deep or shallow nodes on the placental mammal phylogeny. While improvements are observed upon treating the data using various partitioning techniques, the resultant topologies are incongruent with the well-known Superorder groupings (see Figure 3.2). Using individual genes is not recommended for further topological evaluations of the placental mammals; this includes those genes used in the bar code of life project (CO1 and CYTB). With the 10,000 vertebrate genomes project underway, it is becoming less necessary to rely upon one/two genes to infer the genealogical history between species. Increased taxon sampling was proposed in the discussion of Chapter 2 to be necessary to resolve certain placental mammal nodes (e.g. Laurasiatheria), following the analysis in this Chapter it is evident that the inclusion of the current mtGene data available is not the solution.

Even though phylogenies generated using mtGene data were inconsistent, the resulting phylogenies do hold some information about the evolutionary history of a species with regards to potential recombination and hybridization events in the ancestral placental lineages.

Chapter 4

4 The landscape of molecular adaptation and non-adaptive processes on telomere regulating genes in mammals.

4.1 Introduction

4.1.1 Potential Pitfalls in the Detection of Positive Selection

A major application of phylogeny in molecular studies of genes and genomes is to provide directionality to analyses of multiple sequence alignments (MSAs). One such example is assessing selective pressure variation across clades or individual lineages and to tracing signatures of molecular adaptation as described in section 1.1.3.2 (Messier and Stewart 1997, Creevey and McInerney 2002, McDonald and Kreitman 1991, Goldman and Yang 1994, Nielsen and Yang 1998, Yang and Nielsen 2002). These analyses can provide powerful evidence for adaptation at the molecular level that can be analysed further at the phenotypic level (Loughran et al. 2012, Levasseur et al. 2006). In this Chapter I apply the phylogeny I have resolved for the placental mammals to the analysis of selective pressure variation across a small group of proteins involved in telomere maintenance.

The method of selective pressure analysis employed is the codon based maximum likelihood method implemented in CodeML from the PAML package (Yang 1997, Yang et al. 1998). The success and validity of selective pressure analyses using CodeML (Yang 1997) has been frequently disputed (Zhai et al. 2012, Hughes 2007). There are a number of factors, algorithmic and sequence based, that can result in false positives in these selective pressure analyses. It has been shown that different alignment methods give different aligned sequences, which can impact phylogenetic reconstruction and subsequent detecting of selective pressure analyses (Markova-Raina and Petrov 2011, Fletcher and Yang 2009, Schneider et al. 2009). Whelan and Blackburne demonstrated that alignment methods that have a higher tendancy to place non-homologous characters together are more prone to errors when testing for positive selection (Whelan and Blackburne 2012). There are restrictions when using CodeML (Yang 1997, Yang et al. 1998) in terms of the size of the dataset and the level of sequence similarity. Simulations have shown that prediction of positively selected sites is unreliable when less than 6 species are used, when sequences are very similar (tree length < 0.11) and if sequence length is low (< 50 codons) (Anisimova et al. 2002, Anisimova et al. 2001).

The phylogeny used, and specifically whether it is a star phylogeny or a random tree, also impacts on these analyses (Anisimova et al. 2003). CodeML (Yang 1997, Yang et al. 1998) takes as input a phylogenetic tree to reconstruct ancestral sequences at nodes and from that determines when in time selective pressures varied and on what lineage (see section 1.1.3.2). Application of a star phylogeny or random tree topologies in a CodeML study results in extremely high error rates for the LRTs (Anisimova et al. 2003). Gene trees do not necessarily agree with the genealogical history of the species, therefore in this Chapter I have examined the different selective pressure results obtained using the gene and species trees.

There are also a number of evolutionary and population level phenomena that impact upon, and directly influence, the rates of mutation and fixation of non-adaptive substitutions. Therefore, $\omega > 1$ may not always be indicative of protein functional shift (see section 1.1.5) and may reflect variations in effective population size (N_e) (Chen and Li 2001), recombination events (Anisimova et al. 2003), biased gene conversion (gBGC) (Galtier and Duret 2007, Galtier et al. 2001) or relaxation of functional constraint (Hughes and Friedman 2004, Wilkinson et al. 2005). Recombination events (Anisimova et al. 2003) and gBGC (Galtier and Duret 2007, Galtier et al. 2001) can cause elevated fixation of mutations which can be misinterpreted for signatures of positive selection.

There are a number of physical factors that are thought to influence the ω observed for a given protein including position in pathway (hub or peripheral), degree of connectivity, intracellular location and the complexity of the biological process (Aris-Brosou 2005, Fraser et al. 2002, Hahn and Kern 2005, Vitkup et al. 2006, Hudson and Conant 2011). Studies have shown that proteins that are highly connected are less likely to be under adaptive evolution than proteins that are on the network periphery (Aris-Brosou 2005, Fraser et al. 2002, Hahn and Kern 2005, Vitkup et al. 2006, Hudson and Conant 2011). This set of telomere related proteins were chosen as the test set as they represent a set of genes that are present across all vertebrates, are strongly linked in terms of their functions and interactions, and contain a small number of genes previously published as being under positive selection.

4.1.2 The Importance of Telomere Maintenance in Cancer Evasion

Telomeres are short tandem "TTAGGG" sequence repeats (~10-15kb long in humans) that cap the termini of chromosomes and telomerase is the enzyme that maintains telomere length in human germline or stem cells (Figure 4.1). In humans the life span of a somatic cell is controlled by the length of the telomere cap, for each round of cell division a "TTAGG" repeat is lost and the telomeric sequences are shortened (de Lange 2002). After a series of cell divisions, the telomeric sequences reach a point where the cell reaches "replicate senescence", this mechanism has evolved in certain placental mammal species and protects against cancer (Seluanov et al. 2007). Species that display high telomerase activity in somatic cells have evolved other lineage specific mechanisms for cancer avoidance such as the naked mole rat where cell proliferation is controlled based on cell contact inhibition (Seluanov et al. 2009). In this Chapter I sought to investigate the evolutionary pressures acting on genes that interact with telomeres (Blasco 2005). These genes have critical functions in DNA repair and maintenance of chromosome stability (Blasco 2005). By analysing the site-specific and lineage-site specific selective pressure variations, the objective was to determine if there were sections of the telomere maintenance network that have been under strong selective pressure to change across all placental mammals or indeed in specific lineages.

Errors can be introduced into selective pressure analyses through the use of ill-fitting or inappropriate alignment methods. To combat this, a statistical comparison of alignment methods was used to determine the most significant alignment for each gene (Thompson et al. 2001). To address the potential impact of phylogeny of the selective pressure analysis, a simple approach was taken whereby both the gene and species tree for each of the single gene orthologs in the dataset were assessed for their fit to the data and were simultaneously used in the CodeML analysis, and the results (parameter estimates and likelihood scores) were compared. The impact on the level of false positive detection of positive selection of non-adaptive mutations was also assessed, specifically gBGC as measured by the percentage GC3 (GC3%) and recombination. Identification of (i) higher than expected GC3%, and or (ii) recombination events, coinciding with putative positively selected regions is taken as an indication of the presence of false positives. With these data and approaches, the results of the selective pressure analyses were filtered for true positives (as estimated by my criteria set out above). An in-depth analysis is presented on the microbat as it has an extremely high metabolic rate along with extreme longevity with respect to its body size.

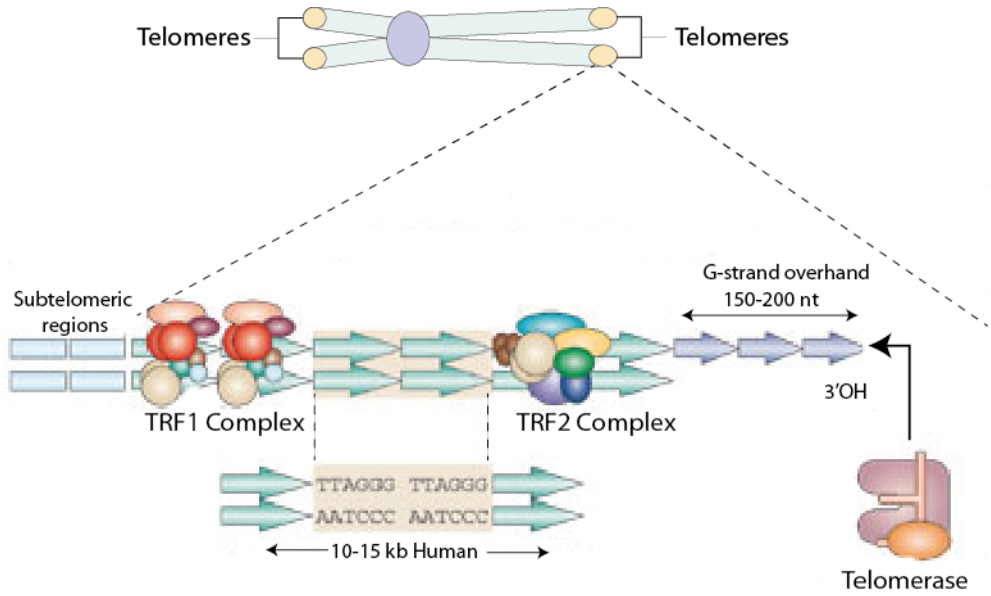


Figure 4.1 Structure of mammal telomeres.

The telomere end of the chromosome is shown along with three key structures necessary for its maintenance and control: (i) TRF1 complex, (ii) TRF2 complex and (iii) Telomerase. This image has been adapted from (Blasco 2005) with permission.

4.2 Materials and Methods

4.2.1 Data Assembly and Taxa Sampling

Coding DNA sequences (CDSs) were obtained for 29 vertebrate genomes through the Ensembl web server (Flicek et al. 2012) (http://www.ensembl.org/). Coding DNA for the Naked Mole Rat (NMR) genome was downloaded from the NMR database (http://mr.genomics.org.cn/page/species/index.jsp) (Kim et al. 2011). The details of all 30 species genomes used in this Chapter are in Table 4.1. The longest canonical transcript for each CDS was identified using the program "GetEnsemblCanonicalTranscripts.pl" (Appendix C.1) (Perl provided by Thomas Walsh, group-member). This program takes a FASTA formatted file containing all available transcripts, identifies the longest canonical transcript and outputs one transcript per gene. It works as follows:

```
perl GetEnsemblCanonicalTranscripts.pl input_fasta_file
output_fasta_file
```

The CDSs were translated to their amino acid counterparts using methods described in section 2.2.1.

Table 4.1 Taxon Sampling

Common Name	Latin Names	Genome Version	Species code
Alpaca	Vicugna pacos	vicPac1	Alp
Cat	*Felis catus*	CAT	Cat
Chicken (outgroup)	Gallus gallus	WASHUC2	Chi
Chimpanzee	*Pan troglodytes*	CHIMP2.1.4	Chk
Cow	*Bos taurus*	UMD3.1	Cow
Dog	*Canis lupus familiaris*	CanFam_2.0	Dog
Dolphin	*Tursiops truncatus*	turTru1	Dol
Elephant	*Loxodonta africana*	loxAfr3	Ele
Guinea_pig	*Cavia porcellus*	cavPor3	Gui
Hedgehog	*Erinaceus europaeus*	eriEur1	Hed
Horse	*Equus caballus*	EquCab2	Hor
Human	*Homo sapiens*	GRCh37.p7	Hum
Kangaroo_rat	*Dipodomys ordii*	dipOrd1	Kag
Macaque	*Macaca mulatta*	MMUL_1.0	Mac
Marmoset	*Callithrix jacchus*	calJac3	Mar
Megabat	Pteropus vampyrus	pteCVam1	Meg
Microbat	*Myotis Lucifugus*	myoLuc2	Mic
Mouse	*Mus musculus*	NCBIM37	Mou
Naked Mole Rat	Heterocephalus glaber	1	Nmr
Opossum (outgroup)	*Monodelphis domestica*	momDom5	Opo
Orangutan	*Pongo abelii*	PPYG2	Ora
Panda	*Ailuropoda melanoleuca*	ailMel1	Pan
Pig	*Sus scrofa*	Sscrofa10.2	Pig
Pika	*Ochotona princeps*	OchPri2.0	Pik
Platypus (outgroup)	Ornithorhynchu anatinus	OANA5	Pla
Rabbit	*Oryctolagus cuniculus*	oryCun2.0	Rab
Rat	*Rattus norvegicus*	RGSC3.4	Rat
Shrew	*Sorex araneus*	sorAra1	Shr
Squirrell	*Spermophilus tridecemlineatus*	spetri2	Squ
Zebra_Finch (outgroup)	*Taeniopygia guttata*	taeGut3.2.4	Zeb

Species used in this analysis are listed with Common names, Latin names, Genome versions and 3 letter short hand notation that was employed in this study.

4.2.2 Ortholog Identification and Telomere Gene Identification

Ortholog identification was carried out using orthoMCL (Li et al. 2003) see section 2.2.3.1, the pipeline and individual gene counts for each step of the process and identification of gene families are shown in Figure 4.2. The 56 Telomerase genes were taken from a review by Blasco (2005) and names are given in Table 4.4. The program "FindFile.py" (Appendix C.2) was written to extract the 56 candidate genes from the total number of gene families (23, 818), resulting in a total of 54 gene families.

4.2.3 Alignment Generation and Editing

Assembly of protein and nucleotide alignments along with alignment editing was performed using the methods described in section 2.2.3.2 and 2.2.3.3. Each MSA was manually edited using Se-AL (Rambaut 2001) to remove poorly aligned regions. Alignments have been provided in Appendix C.3. It has been recommended that at least 6 taxa are used in order to accurately identify positive selection using CodeML (Anisimova et al. 2002), therefore following editing of the MSA the number of useable datasets was reduced from 54 to 52. The impact of the alignment editing or filtering was not tested on these data, however, this methodological approach has been previously successful in similar peer reviewed work (Morgan et al. 2010, Morgan et al. 2012).

4.2.4 Phylogeny reconstruction

The identification of the protein evolutionary model and phylogenetic reconstruction were carried out using ModelGenerator v.85 (Keane et al. 2006) and hybrid MrBayes v.3.1.2h (Huelsenbeck and Ronquist 2001) respectively, using the processes described in section 2.2.4.1 and 2.2.6.1. Phylogenetic trees are available in Appendix C.4.

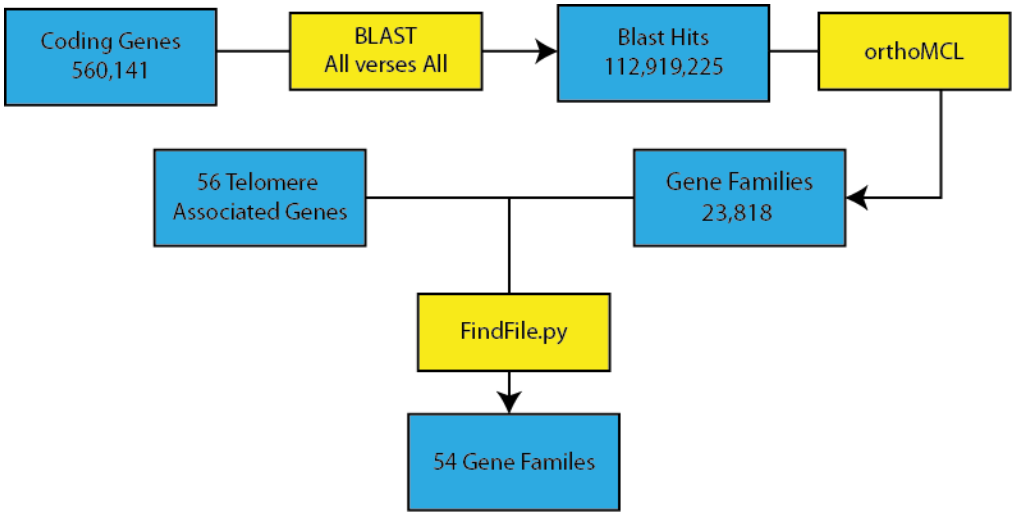


Figure 4.2 Pipeline of Ortholog Identification

This pipeline describes the process of identifying gene families across 30 species from an initial set of 560,141 using orthoMCL (Li et al. 2003).The Blue boxes represent datasets and yellow boxes represent methods applied.

Table 4.2 Telomere Associated Gene Set

Gene Name	Telomere Regulation Function	# of Seqs	MSA (bp)	MeanID	Phylogenetic Model of Evolution	Gene Family
ABL1, ABL2	DNA repair	48	3762	0.668	JTT+4Γ	MultiGene
ANKRD17, ANKHD1, ANKHD1-EIF4EBP3	DNA repair	39	8247	0.727	JTT+4Γ	MultiGene
ATM	DNA repair	21	9180	0.785	JTT+4Γ	SGO
ATRX	DNA repair	15	7485	0.794	JTT+4Γ	SGO
BLM	DNA repair	18	4263	0.773	JTT+4Γ	SGO
BRCA1	DNA repair	28	5622	0.575	JTT+4Γ	SGO
BRCA2	DNA repair	24	10425	0.577	JTT+4Γ	SGO
BRIP1	DNA repair	20	3756	0.694	JTT+4Γ	SGO
CBX1	DNA repair	29	555	0.973	JTT+I	Paralogs
CBX3	DNA repair	36	582	0.974	JTT+I	Paralogs
CBX5	DNA repair	23	573	0.973	JTT+4Γ	SGO
DKC1	Telomerase components	28	1545	0.800	JTT+4Γ	Paralogs
ERCC1	DNA repair	24	849	0.691	JTT+4Γ	SGO
ERCC4	DNA repair	26	2754	0.833	JTT+4Γ	Paralogs
EXO1	DNA repair	29	2529	0.665	JTT+4Γ	Paralogs
FANCA	DNA repair	23	4368	0.589	JTT+4Γ	SGO
FANCB	DNA repair	28	2550	0.590	JTT+4Γ	Paralogs
FANCC	DNA repair	24	1680	0.632	JTT+4Γ	SGO
FANCD2	DNA repair	25	4497	0.714	JTT+4Γ	SGO
FANCE	DNA repair	24	1617	0.615	JTT+4Γ	Paralogs
FANCF	DNA repair	19	1065	0.579	JTT+4Γ	SGO
FANCI	DNA repair	20	3972	0.773	JTT+4Γ	SGO
FANCL	DNA repair	28	1125	0.776	JTT+4Γ	SGO
MLH1	DNA repair	26	2280	0.813	JTT+4Γ	SGO
MLH3	DNA repair	27	4407	0.703	JTT+4Γ	SGO

Gene Name	Telomere Regulation Function	# of Seqs	MSA (bp)	MeanID	Phylogenetic Model of Evolution	Gene Family
MRE11A	DNA repair	25	2130	0.815	JTT+4Γ	Paralogs
MSH3	DNA repair	21	3369	0.756	JTT+4Γ	SGO
MUTYH	DNA repair	24	1560	0.714	JTT+4Γ	SGO
NBN	DNA repair	25	2262	0.638	JTT+4Γ	SGO
PALB2	DNA repair	23	3570	0.605	JTT+4Γ	SGO
PARP2	DNA repair	22	1710	0.837	JTT+4Γ	SGO
PCNA	DNA repair	23	786	0.969	JTT+4Γ	SGO
PMS1	DNA repair	26	2799	0.753	JTT+4Γ	SGO
PMS2	DNA repair	24	2610	0.719	JTT+4Γ	SGO
RAD50	DNA repair	27	3936	0.804	JTT+4Γ	SGO
RAD51D	DNA repair	24	987	0.736	JTT+4Γ	SGO
RB1	DNA repair	28	2793	0.817	JTT+4Γ	SGO
RBL2, RBL1	Chromatin regulators	44	3573	0.654	JTT+4Γ	MultiGene
SLX4	DNA repair	17	5514	0.582	JTT+4Γ	SGO
SUV39H1	Chromatin regulators	25	1236	0.800	JTT+4Γ	Paralogs
SUV39H2	Chromatin regulators	31	1230	0.826	JTT+4Γ	Paralogs
TDG	DNA repair	24	1236	0.794	JTT+4Γ	Paralogs
TERF1	Telomere-binding protein	26	1308	0.657	JTT+4Γ	Paralogs
TERF2	Telomere-binding protein	23	1635	0.750	JTT+4Γ	SGO
TERT	Telomerase components	22	3333	0.568	JTT+4Γ	SGO
TINF2	Telomere-binding protein	23	1359	0.707	JTT+4Γ	Paralogs
TNKS2, TNKS	Telomere-binding protein	50	4020	0.818	JTT+4Γ	MultiGene
TREX1	DNA repair	14	945	0.740	JTT+4Γ	SGO
WRN	DNA repair	26	4224	0.627	JTT+4Γ	SGO
XRCC3	DNA repair	23	1044	0.687	JTT+4Γ	SGO
XRCC5	DNA repair	26	2199	0.749	JTT+4Γ	SGO
XRCC6	DNA repair	28	1827	0.771	JTT+4Γ	Paralogs

Details on Components and Main Roles of Genes were taken from (Blasco 2005) with permission.

4.2.5 Phylogenetic Signal Tests

The level of phylogenetic conflict was assessed for each of the 56 genes using the likelihood mapping (LM) function in TreePuzzle v5.2 (Schmidt et al. 2002) (see section 2.2.8). The χ^2 test was also implemented in Tree-Puzzle (Schmidt et al. 2002) to test if the model of evolution accurately described the data (see section 2.2.5.3). A Shimodaira Hasegawa (SH) (Shimodaira and Hasegawa 2001) test was performed to assess whether the gene tree was statistically significantly better in describing the data under the evolutionary model than the species tree ($P < 0.05$).

4.2.6 GC3 Analysis-Evolutionary Analyses

The GC3 content of each sequence was calculated using codonW (Peden 1999) and these scores were assessed to determine if a species GC3 content was outside the standard deviation of the mean GC3 and therefore subject to gBGC. Along with this, a sliding window analysis of GC3 content was carried out using "GC_estimator" (Appendix C.5) to assess the GC3 variation across the gene for a given lineage. To run this program, gaps were removed from each MSA, and a non-overlapping sliding window of 90 bp moved across the length of each sequence calculating the percentage of GC3 (GC3%) frame by frame. It is run as follows:

```
python GC_estimator.py infile.fasta  > outfile.fasta
```

4.2.7 Recombination Detection- Evolutionary Analyses

Recombination breakpoints within sequences were calculated using RDP4 (Martin et al. 2010). RDP4 employs the following primary exploratory recombination signal detection methods: RDP (Martin and Rybicki 2000), GENECONV (Padidam et al. 1999), BOOTSCAN/RESCAN (Martin et al. 2005), MaxChi (Smith 1992), Chimaera (Posada and Crandall 2001), SiScan (Gibbs et al. 2000), 3Seq (Boni et al. 2007). A secondary method, PhylPro (Weiller 1998) was used to assess the recombination results of these primary methods. The recombination detection methods are divided into three groups based on their algorithmic approach: (i) substitution, (ii) distance, and (iii), phylogeny based. Each of these methods are detailed in Table 4.3. Seven of the methods mentioned below use a sliding-windows approach to analyse data, with the exception of

GENECONV that splits the alignment into every possible triplet pair. CHIMERA and 3SEQ require that both parental sequences are present in order to detect the recombinant sequences while other programs only require that one parental sequence is present. Recombination detection can be influenced by among site rate heterogeneity, sequence divergence and levels of recombination. As a result, the 8 methods described in Table 4.3 vary in their ability to detect recombination events and Posada and Crandall (2001) recommend that they are used in conjunction with one another to more accurately detect recombination events (Posada and Crandall 2001). CHIMERA and MAXCHI are identified as being the most accurate breakpoint detection methods implemented in RDP4 (Posada and Crandall 2001). As substitution methods were used to identify recombination events, validation of recombinant sequences should be carried out using a phylogenetic based method. In this analysis, the recombination detection methods have been weighted to increase accuracy of recombination detection. All recombination breakpoints detected must have statistical support from two separate programs. One of these programs must be from the substitution-based methods CHIMERA or MAXCHI and the other must be from the two other well performing phylogenetic based methods BOOTSCAN or SISCAN (Posada and Crandall 2001). Recombination detection is difficult, therefore detection of recombination events using additional methods increases overall confidence in accuracy. The results for recombination analysis are available in Appendix C.6.

Table 4.3 Recombination Detection Methods

Method	Category	Break Point Detection	Sliding Windows	Requires both parental sequences	Reference
RDP	Phylogenetic	+	+	-	(Martin and Rybicki 2000)
GENECONV	Substitution	+	-	-	(Padidam et al. 1999)
BOOTSCAN	Phylogenetic	+	+	-	(Salminen et al. 1995)
MAXCHI	Substitution	+	+	-	(Smith 1992)
CHIMAERA	Substitution	+	+	+	(Posada and Crandall 2001)
SISCAN	Phylogenetic	+	+	-	(Gibbs et al. 2000)
3SEQ	Substitution	+	+	+	(Boni et al. 2007)
PHYLPRO	Distance	+	+	-	(Weiller 1998)

Each of the methods used this study is tabulated along with the category the method belongs to. The "+" symbol signifies if the method is able to carry out break point detection, in the fourth column the "+" signifies that the method works using a sliding window approach. A method requiring both parently sequences to be present is denoted with a "+" in column five. The "-" symbol is given where a method does not have these features. This table is adapted from (Posada and Crandall 2001).

4.2.8 Testing for site and lineage site positive selection

Variation in selective pressures were analysed using site-specific and lineage-site specific models in CodeML from the PAML4.3 package (Yang 1997, Yang et al. 1998). Discussed in detail in section 1.1.3.2. All extant lineages were tested for signatures of adaptive evolution under these lineage specific CodeML models (Yang 1997, Yang et al. 1998). All results from CodeML analyses have been made available in Appendix C.7. The false discovery rate (FDR) was not calculated to correct for multiple tests.

4.2.9 Identification of Protein-Protein Interactions

STRING v0.9 (Jensen et al. 2009) is an online program available at http://string-db.org and was employed to show experimental interactions (with a high confidence value of 0.700) between 52 telomere regulating proteins used in this study. The network of telomere regulating gene connectivity was obtained through a list of Ensembl human IDs.

4.3 Results

In this analysis 56 telomere-associated genes were clustered into 54 gene families. MSAs were generated using both distance and evolutionary aware methods ensuring a comprehensive exploration of alignment space, as described in section 2.2.3.2. By applying methods with different approaches it was possible to minimise the effect of mis-aligned sequences on: (i) incorrect tree reconstruction, (ii) incorrect recombinant breakpoint detection and, (iii), false positives in selective pressure analyses. The norMD score was calculated for MAFFT, MAFFT+RASCAL, MUSCLE, MUSCLE+RASCAL and PRANK alignments. The MSA from the method with the highest norMD score was brought forward for further testing. In cases where multiple methods scored equally well, the MSA that was brought forward for analysis was chosen from the methods of choice in an arbitrary way (i.e. alphabetically). Sequences that had less than 60% coverage over the entire length of the MSA or did not have 0.6 minimum overlap of a position in the column (as described in section 2.2.3.3) were removed using trimAL (Capella-Gutierrez et al. 2009), giving a final number of 52 gene families that were suitable for selective pressure analysis. The 52 gene families were composed of 4 multigene families, 14 families that contained paralogous sequences and 34 SGO's.

There were 20/52 gene families that had equal norMD scores across the 5 separate alignment methods. MAFFT was the best alignment software in 7/52 cases, MAFFT and RASCAL in 2/52 cases, MUSCLE in 2/52 cases and PRANK in 10/52 cases. In no case did MUSCLE+RASCAL exclusively produce the best alignment. All alignments were checked by eye to ensure alignments were of good quality and poorly aligned regions were removed using Se-Al (Rambaut 2001).

4.3.1 Choosing the Best Phylogeny for Selective Pressure Analyses

LM was performed to assess the quantity of phylogenetic conflict in each dataset. There were high levels of phylogenetic conflict observed, with 33/52 datasets displaying > 10% phylogenetic conflict and only 19/52 datasets with sufficiently low levels of phylogenetic conflict (< 10%), see Table 4.5. Lineage-site specific rate variation has been shown to impact accurate phylogeny reconstruction (Yang 1994) and can be caused by lineage-specific directional selection (Dorus et al. 2004) (described in section 1.1.4), therefore the phylogeny employed to detect positive selection in CodeML may be incorrect as a result of strong selective pressures, and this could impact the false discovery rate (Anisimova et al. 2003). Ideally, to generate the phylogenies for further analysis one would apply heterogeneous models that accommodate heterogeneity over the phylogeny and the data, however as discussed in section 1.2.1, individual gene datasets tend to lack the phylogenetic information to accommodate sophisticated models. The fit of each of the sequences in the datasets to the model of evolution (JTT+4Γ or JTT+I) was assessed using the χ^2 test of ccompositional homogeneity in TreePuzzle v5.2 (Schmidt et al. 2002). A total of 40/52 datasets had 100% of sequences pass the χ^2 test of compositional homogeneity. Only 12/52 genes had between 1 and 4 sequences that did not fit the composition of the model, see Table 4.4. In summary, while high levels of phylogenetic conflict are observed across the data, the compositions of the data were modeled adequately by the associated model of evolution. Gene trees were applied to all paralogous sequences, however where LM and χ^2 test of compositional homogeneity showed systematic biases within the data, this reduced confidence in the phylogeny and results from these gene trees were treated with caution, these trees are highlighted in Table 4.4.

Phylogenetic trees obtained from SGO families were tested against the fit of the canonical species tree (Meredith et al. 2011) using an SH test (Shimodaira and Hasegawa 2001) under the protein model of evolution (JTT+4Γ). The canonical species tree used in this analysis was from Chapter 2. The results of the SH tests are shown in Table 4.4. In cases where systematic bias were indicated by LM and the χ^2 test of compositional homogeneity, and where the SH test indicated that the species tree was not statistically different from the gene tree (p-value > 0.05), then the species tree was employed in further analyses.

Table 4.4 Assessing the Best Phylogeny for Selection Analysis

Gene Name	Gene Family	Likelihood Mapping		χ^2 Compositional Homogeneity		Gene Tree Confidence	SH Tests		Phylogeny for Selection Analyses
		Quartets [4-7]	Pass/Fail	Sequences P < 0.05	Pass/Fail		Species Tree	Gene Tree	
ABL1_ABL2	MultiGene	9.2	PASS	0	PASS	YES	NA	NA	Gene Tree
ANKRD17, ANKHD1, ANKHD1-EIF4EBP3	MultiGene	5	PASS	0	PASS	YES	NA	NA	Gene Tree
RBL2_RBL1	MultiGene	8.8	PASS	0	PASS	YES	NA	NA	Gene Tree
TNKS2_TNKS	MultiGene	20.4	FAIL	0	PASS	NO	NA	NA	Gene Tree*
CBX1	Paralogs	90.3	FAIL	0	PASS	NO	NA	NA	Gene Tree*
CBX3	Paralogs	84.9	FAIL	0	PASS	NO	NA	NA	Gene Tree*
DKC1	Paralogs	29.7	FAIL	0	PASS	NO	NA	NA	Gene Tree*
ERCC4	Paralogs	16.1	FAIL	0	PASS	NO	NA	NA	Gene Tree*
EXO1	Paralogs	20.9	FAIL	1	FAIL	NO	NA	NA	Gene Tree*
FANCB	Paralogs	6.5	PASS	0	PASS	YES	NA	NA	Gene Tree
FANCE	Paralogs	16.3	FAIL	1	FAIL	NO	NA	NA	Gene Tree*
MRE11A	Paralogs	24.2	FAIL	0	PASS	NO	NA	NA	Gene Tree*
SUV39H1	Paralogs	36.7	FAIL	0	PASS	NO	NA	NA	Gene Tree*
SUV39H2	Paralogs	34.2	FAIL	0	PASS	NO	NA	NA	Gene Tree*
TDG	Paralogs	36.3	FAIL	0	PASS	NO	NA	NA	Gene Tree*
TERF1	Paralogs	19.3	FAIL	0	PASS	NO	NA	NA	Gene Tree*
TINF2	Paralogs	15.5	FAIL	0	PASS	NO	NA	NA	Gene Tree*
XRCC6	Paralogs	17.1	FAIL	1	FAIL	NO	NA	NA	Gene Tree*
ATM	SGO	5.3	PASS	1	FAIL	NO	0.085	1	Species Tree
ATRX	SGO	7	PASS	0	PASS	YES	0.258	1	Gene Tree
BLM	SGO	6.2	PASS	1	FAIL	NO	0.176	1	Species Tree
BRCA1	SGO	6.9	PASS	4	FAIL	NO	0.174	1	Species Tree
BRCA2	SGO	2.7	PASS	1	FAIL	NO	0.08	1	Species Tree
BRIP1	SGO	10	PASS	0	PASS	YES	0.232	1	Gene Tree
CBX5	SGO	69.5	FAIL	0	PASS	NO	1	0.083	Species Tree

Gene Name	Gene Family Family	Likelihood Mapping		χ^2 Compositional Homogeneity		Gene Tree Confidence	SH Tests		Phylogeny for Selection Analyses
		Quartets [4-7]	Pass/Fail	Sequences P < 0.05	Pass/Fail		Species Tree	Gene	
ERCC1	SGO	20.6	FAIL	0	PASS	NO	1	0.074	Species Tree
FANCA	SGO	2.9	FAIL	4	FAIL	NO	0.346	1	Species Tree
FANCC	SGO	13	FAIL	0	PASS	NO	0.022	1	Species Tree
FANCD2	SGO	5.4	PASS	0	PASS	YES	0.029	1	Gene Tree
FANCF	SGO	9.3	PASS	1	PASS	YES	0.024	1	Gene Tree
FANCI	SGO	8.2	PASS	0	PASS	YES	0.158	1	Gene Tree
FANCL	SGO	25	FAIL	0	PASS	NO	1	0.028	Species Tree
MLH1	SGO	17.3	FAIL	1	FAIL	NO	0.323	1	Species Tree
MLH3	SGO	6.5	PASS	1	FAIL	NO	0.069	1	Species Tree
MSH3	SGO	10	PASS	0	PASS	YES	0.051	1	Gene Tree
MUTYH	SGO	15.7	FAIL	0	PASS	NO	0.025	1	Species Tree
NBN	SGO	23.3	FAIL	1	FAIL	NO	0.104	1	Species Tree
PALB2	SGO	7.7	PASS	0	PASS	YES	0.395	1	Gene Tree
PARP2	SGO	23.2	FAIL	0	PASS	NO	0.185	1	Species Tree
PCNA	SGO	46.3	FAIL	0	PASS	NO	0.033	1	Species Tree
PMS1	SGO	12.4	FAIL	0	PASS	NO	0.066	1	Species Tree
PMS2	SGO	14.4	FAIL	0	PASS	NO	0.186	1	Species Tree
RAD50	SGO	20.5	FAIL	0	PASS	NO	1	0.342	Species Tree
RAD51D	SGO	23.6	FAIL	0	PASS	NO	0.076	1	Species Tree
RB1	SGO	23	FAIL	0	PASS	NO	1	0.292	Species Tree
SLX4	SGO	1.5	PASS	2	FAIL	YES	1	1	Species Tree
TERF2	SGO	22.9	FAIL	0	PASS	NO	0.069	1	Species Tree
TERT	SGO	3.9	PASS	3	PASS	YES	0.154	1	Gene Tree
TREX1	SGO	13.7	FAIL	0	PASS	NO	0.257	1	Species Tree
WRN	SGO	7	PASS	0	PASS	YES	0.041	1	Gene Tree
XRCC3	SGO	18.8	FAIL	0	PASS	NO	0.287	1	Species Tree
XRCC5	SGO	18.1	FAIL	0	PASS	NO	0.301	1	Species Tree

* is where gene tree fail LM test or data is compositionally heterogeneous.

4.3.2 Studying the effect of Species tree versus Gene Tree on CodeML parameter estimates for Single Gene Orthologs

Following LM analyses, $\chi 2$ compositional homogeneity tests and SH tests, it was determined that a species tree was a better description of the MSAs for 24/34 SGOs. Previous studies have compared the LRT results from data analyzed under a star phylogeny or random phylogenies against the result from the actual phylogeny and demonstrated that the incorrect phylogeny resulted in false positives (Anisimova et al. 2003). Here, I determine the variations in the LRT results when the species tree and gene tree are used in a CodeML analysis. In total 34 SGOs were analysed using both gene and species trees and the site-specific and lineage-site specific models outlined in Chapter 1 (section 1.1.3.2). The results from the M8 model and modelA were compared and are described below, see Table 4.5 for summary.

There were 21/34 SGOs where the difference in lnL values determined that the species tree was a better fit to the data than the gene tree using the M8 selection model in CodeML (Yang 1997). Five of the cases where the gene tree was deemed a suitable choice, the species tree was a better fit to the data under the M8 model of evolution. In contrast, there were 13 cases where the gene tree was a better fit to the data when compared to the species tree (Table 4.5).

Both gene tree and species tree analyses detected site-specific positive selection using M8, however they differed in the proportion of sites and the ω values (although these were slight fluctuations and did not shift the result from insignificant to significant or vice versa). The maximum difference in the estimation of the proportion of sites using the M8 model was 0.016, while the maximum difference of the estimation of ω using the M8 model was 0.24. In 17/34 cases both the species tree and gene tree identified an identical number of BEB sites with a PP > 0.50 using the M8 model. In 9 cases the species tree identified 1 or 2 more sites than the gene tree using BEB and in 8 cases the gene tree identified between 1 and 23 more positively selected sites using BEB compared with the species tree.

In 27 SGOs both gene and species tree identified an identical number of species under lineage-specific positive selection using modelA. There were 6 cases (ERCC1, FANCI, RB1, TERT, WRN and XRCC3) where the gene tree phylogeny identified either 1 or 2

more species as evolving under lineage-specific evolution that were not identified when the species phylogeny was applied. In each of these cases the lnL values indicated that the species phylogeny had a better fit to the data than the gene tree. When the species tree phylogeny was applied to the analysis of the BLM gene, an additional lineage was detected as being under positive selection. The gene tree phylogeny had an lnL value (-29678.97), lower than that of the species tree (-29690.61), indicating that it was also a better fit to the dataset.

The results of this study demonstrated that CodeML (Yang 1997, Yang et al. 1998) estimates vary depending on whether a gene tree or a species tree is used but that the effects of phylogeny are greater on the lineage-specific models. There are still some discrepancies within the placement of species within the mammal phylogeny (as discussed in Chapter 2), especially within the Laurasiatheria. While the species phylogeny appeared to out perform the gene phylogeny in some cases, this was not consistent across all genes tested. Therefore, the decision on whether to apply a gene or species phylogeny to data for a CodeML analysis (Yang 1997, Yang et al. 1998) was made based on confidence in the phylogeny as assessed in section 4.3.1. The option of applying a species tree phylogeny was not available when dealing with multigene families. Chapter 2 and 3 show how both compositional heterogeneity and phylogenetic conflict impact upon phylogenetic reconstruction, here it is shown how these features of the sequence data increase the probability of detecting false positives in both site and lineage-site selective pressure analyses.

Table 4.5 Species Tree versus Gene Tree M8 Selection Model Results4

Gene Name	lnL Scores (M8)		p1 and ω (M8)		BEB Sites (M8)		Lineage selection (modelA)	
	Species	Gene	Species	Gene	Species	Gene	Species	Gene
ATM	-68608.92	-68604.2	p1=0.04231 ω=1.08290	p1=0.04078 ω=1.09068	None	None	7	7
ATRX*	-42670.52	-42685.99	p1=0.02873 ω=1.84172	p1=0.03166 ω=1.73277	77	76	7	7
BLM	-29690.61	-29678.97	p1=0.01907 ω=2.41815	p1=0.01977 ω=2.37309	43	44	7	6
BRCA1	-70416.83	-70431.21	p1=0.05203 ω=1.90926	p1=0.05229 ω=1.89899	117	137	14	14
BRCA2	-113334.77	-113326.97	p1=0.04358 ω=1.96602	p1=0.04375 ω=1.95884	86	86	15	15
BRIP1*	-29718.51	-29730.61	p1=0.05544 ω=1.64715	p1=0.05580 ω=1.63258	45	47	9	9
CBX5	-2262.99	-2515.49	p1=0.00001 ω=1.00000	p1=0.00001 ω=1.00000	None	None	1	1
ERCC1	-7993.08	-8028.15	p1=0.03055 ω=1.31294	p1=0.02743 ω=1.38325	None	None	4	5
FANCA	-54188.44	-54184.42	p1=0.04216 ω=1.54261	p1=0.04073 ω=1.56497	51	50	10	10
FANCC	-17566.35	-17562.43	p1=0.11771 ω=1.26564	p1=0.10204 ω=1.29535	16	16	5	5
FANCD2*	-41964.31	-41963.1	p1=0.05662 ω=1.54966	p1=0.05936 ω=1.51734	52	54	7	7
FANCF*	-10919.71	-10899.64	p1=0.02252 ω=1.1742	p1=0.01684 ω=1.12264	None	None	4	4
FANCI*	-28086.14	-28076.61	p1=0.04859 ω=1.00000	p1=0.04439 ω=1.00000	None	None	4	5
FANCL	-9971.88	-10119.41	p1=0.06238 ω=1.00000	p1=0.06194 ω=1.00000	None	None	7	7
MLH1	-19978.59	-20020.54	p1=0.08220 ω=1.00000	p1=0.07920 ω=1.00000	None	None	4	4
MLH3	-46075.21	-46059.76	p1=0.07499 ω=1.42864	p1=0.07320 ω=1.42978	32	32	6	6
MSH3*	-26491.45	-26491.51	p1=0.05073 ω=1.37915	p1=0.04883 ω=1.39567	33	32	7	7
MUTYH	-14610.85	-14613.96	p1=0.03673 ω=2.20526	p1=0.04752 ω=1.96084	14	17	6	6
NBN	-24254.25	-24247.86	p1=0.05408 ω=1.73623	p1=0.05459 ω=1.73047	31	32	11	11

Gene Name	lnL Scores (M8)		p1 and ω (M8)		BEB Sites (M8)		Lineage selection (modelA)	
	Species	Gene	Species	Gene	Species	Gene	Species	Gene
PALB2*	-39785.84	-39798.18	p1=0.07360 ω=1.67105	p1=0.07623 ω=1.65863	75	98	10	10
PARP2	-10888.72	-10891.11	p1=0.09472 ω=1.29553	p1=0.09581 ω=1.28060	24	24	2	2
PCNA	-3975.76	-4109.79	p1=0.00001 ω=1.00000	p1=0.00001 ω=1.00000	None	None	0	0
PMS1	-26409.68	-26410	p1=0.04964 ω=1.37239	p1=0.04753 ω=1.38453	29	28	5	5
PMS2	-27893.84	-27876.69	p1=0.09490 ω=1.30701	p1=0.09123 ω=1.31912	24	24	2	2
RAD50	-28681.95	-28804.33	p1=0.00157 ω=2.40006	p1=0.00146 ω=2.44689	7	7	8	8
RAD51D	-9815.27	-9802.6	p1=0.02956 ω=1.73401	p1=0.02285 ω=1.82251	9	8	10	10
RB1	-19139.18	-19254.2	p1=0.01458 ω=1.68000	p1=0.01289 ω=1.79892	15	14	6	7
SLX4	-45924.23	-45924.23	p1=0.07380 ω=1.61905	p1=0.07380 ω=1.61905	74	74	9	9
TERF2	-11150.06	-11185.95	p1=0.04074 ω=1.96320	p1=0.03107 ω=2.05648	20	20	5	5
TERT*	-39181.45	-39183.74	p1=0.06761 ω=1.2743	p1=0.06816 ω=1.26374	42	43	9	10
TREX1	-5608.02	-5607.81	p1=0.11071 ω=1.62503	p1=0.11028 ω=1.62808	34	34	0	0
WRN*	-43958.42	-43977.16	p1=0.05703 ω=1.69283	p1=0.05831 ω=1.66250	42	41	9	10
XRCC3	-11769.45	-11795.73	p1=0.00877 ω=2.16549	p1=0.01046 ω=2.10012	8	6	2	4
XRCC5	-19090.57	-19120.69	p1=0.04717 ω=1.74596	p1=0.04392 ω=1.79767	32	30	4	4

The SGO associated gene name is listed along with species and gene phylogeny results comparisons under the M8 model and modelA.

* represents cases where there was greater confidence in the gene tree as assessed by LM and compositional fit of models.

4.3.3 Analysis of Site-specific Positive Selection results in the context of Protein-Protein interacting networks.

A protein-protein interaction network of human telomere regulating genes is shown in Figure 4.3(A). Only protein-protein interactions where strong experimental evidence exists were included. LRT calculations determined 36/52 of the telomere regulating genes were under directional selection ($\omega > 1$), and model M8 was a statistically better fit to the data than model M7 or M8a. The genes where site-specific positive selection was identified using model M8 were plotted against the amount of protein-protein interactions for each gene, see Figure 4.3(B). The most connected protein in the telomere network (RAD50) was identified to be evolving under positive selection ($\omega = 2.4$). Other proteins such as MUTYH and PALB2 were identified as having interactions with only 1 other protein and had similar levels of positive selection with ω values of 2.21 and 1.66 respectively, see Figure 4.3(B). Furthermore, other highly connected proteins (ATM, MLH1, ABL1 and PCNA) were found to be evolving neutrally. Using these data, it is observed that there is no correlation between level of connectivity of the proteins and selective pressures acting on a gene in this dataset. The results are more indicative of similar levels of positive selection acting across the network of protein-protein interactions. If however this dataset was expanded and in include all possible protein-protein interactions then a relationships between protein-protein connectivity and strength of positive selection may be observed.

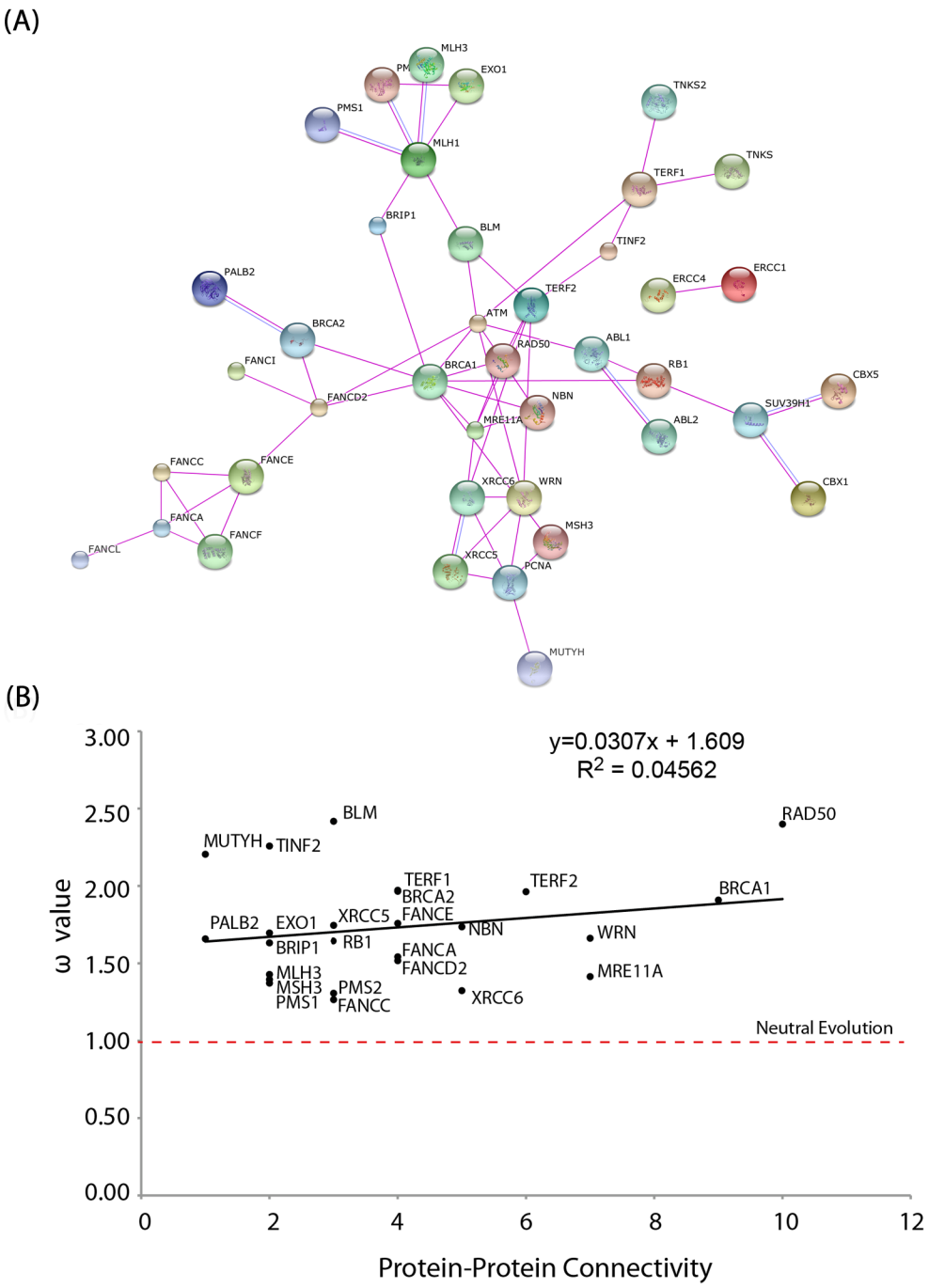


Figure 4.3 Degree of the protein in the network (protein-protein connectivity) compared to ω value estimated from model M8.

The protein-protein interaction network is shown for the telomere regulating proteins in the dataset is shown in (A). Where proteins are the nodes on the network, the pink lines are connections through experimental evidence and blue lines are where large portions of sequence similarity are observed between proteins. (B) The degree of each protein (a measurement of the number of interacting partners it has) is plotted against the ω value obtained using the M8 selection model. The red horizontal line is ω = 1 (neutral evolution).

4.3.4 Lineage-Specific Selective pressure analysis of Telomere Regulating Proteins

Lineage-specific positive selection acting on telomere regulating proteins occurred across all placental mammals studied regardless of life trait histories, see Figure 4.4. Four separate roles of telomere regulating proteins were analysed; (i) proteins involved in DNA repair in 43/52 genes, (ii) proteins that form the telomerase enzyme in 2/52 genes, (iii) proteins that directly bind with telomeres in 5/52 genes, and (iv), proteins that regulate the chromatin in 4/52 genes. The results of these four layers of activity are summarised here:

(i) DNA repair genes are critical in identification of mutational defects in sequences and initiating subsequent repair. The highest proportion of lineage-site specific selection was observed within the DNA repair genes and species such as human (3/52), orangutan (4/51), kangaroo rat (5/23), pika (2/26), horse (7/46), cat (2/19), shrew (8/26) and hedgehog (7/25) only showed lineage-specific evolution within this functional group. As the vast majority of genes tested in this chapter (43/52) are involved in DNA repair the large number of genes showing signs of positive selection in this functional category is more likely due to over representation of this functional category in the original sample rather than some significant functional bias.

(ii) TERC, TERT and DKC1 are components of the Telomerase enzyme that regulates telomere length, see Figure 4.1. There are 13 lineages that show evidence of positive selection acting on the components of this enzyme, see Figure 4.4. Marmoset, microbat and guinea pig species in total account for two thirds of the telomerase components under positive selection in this category while all other species have either none or 1 of their proteins in this category under positive selection.

(iii) Telomere binding proteins are directly involved in DNA repair and regulation. There were 15 lineages identified as under lineage specific positive selection.

(iv) Telomeres are bound by nucleosome arrays that, when modified by histone methytransferases (HMTases), methylate the histone in heterochromatic regions (Tommerup et al. 1994, Garcia-Cao et al. 2004). This epigenetic regulation of telomeric chromatin is as crucial step in telomere maintenance. In total 15 species had evidence for positive selection in their chromatin regulation proteins.

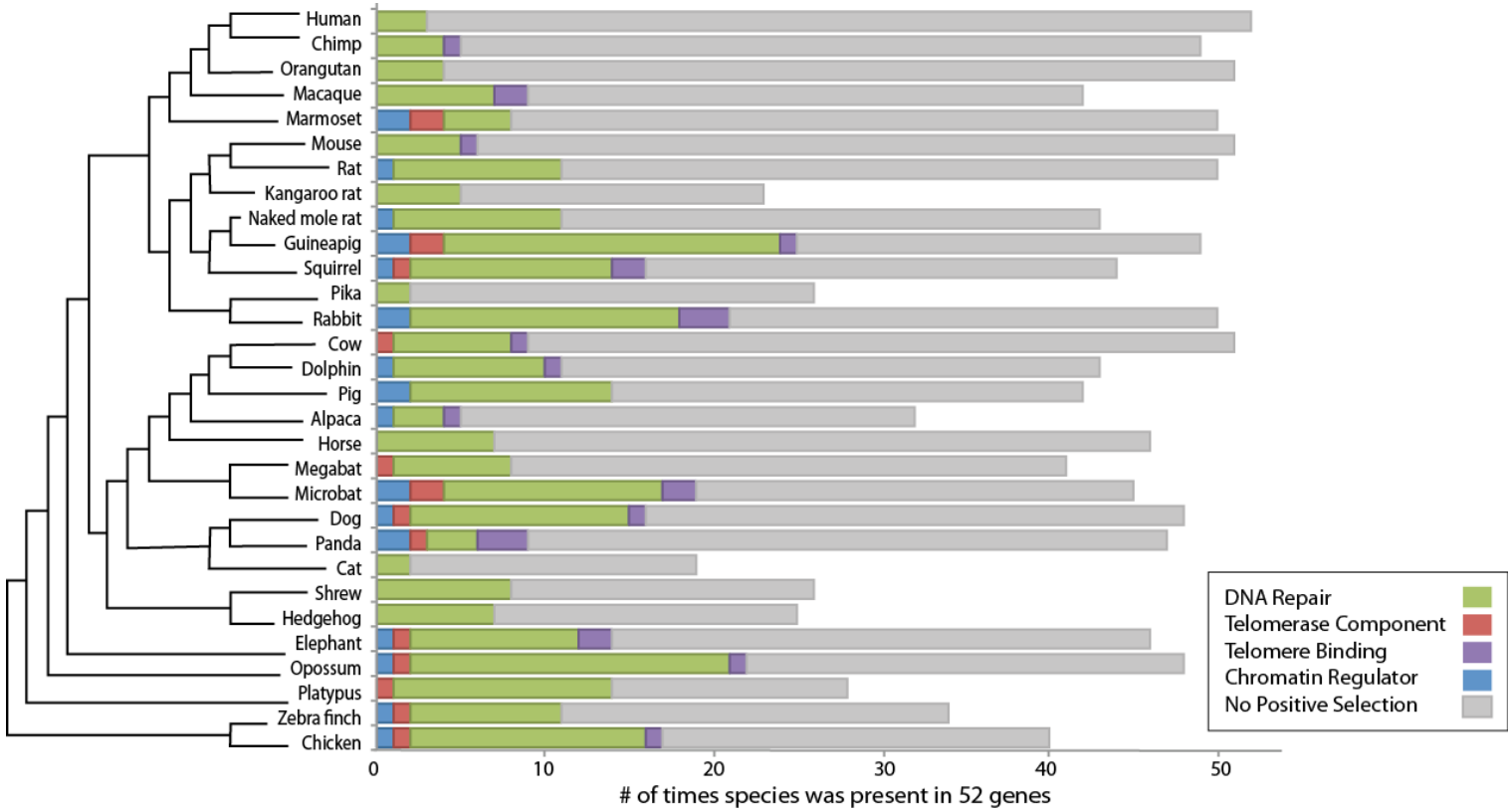


Figure 4.4 Lineage-site specific selective pressure analyses on different telomere functional groups

The length of each bar represents the number of times a species was represented in each MSA of the 52 telomere regulating genes. The absolute number of genes where no positive selection was identified are in grey and where positive selection was identified in DNA repair genes (green), telomeres components (red), telomere binding proteins (purple) and chromatin regulators (blue) are shown for each species. The species phylogeny used is shown on the left of the diagram.

4.3.5 Identification of non-adaptive substitution patterns in the data.

Recombination events (Anisimova et al. 2003) and gBGC (Galtier and Duret 2007, Galtier et al. 2001) are non-adaptive processes that can create signatures which mimic positive selection in LRT analyses. These data were tested for species-specific recombination events and results are shown alongside lineage-specific positive selection results in Figure 4.5. Recombination events were detected in 47/52 of genes in the dataset, and in all lineages at high levels. The GC3% was calculated for all sequences, and where the GC3 content for a species was above the standard deviation from the mean GC3, the species was highlighted (Figure 4.5). The species that showed the highest proportion of GC3 bias were the outgroups and the rodents, while the primates had no sequences that had elevated GC3 levels.

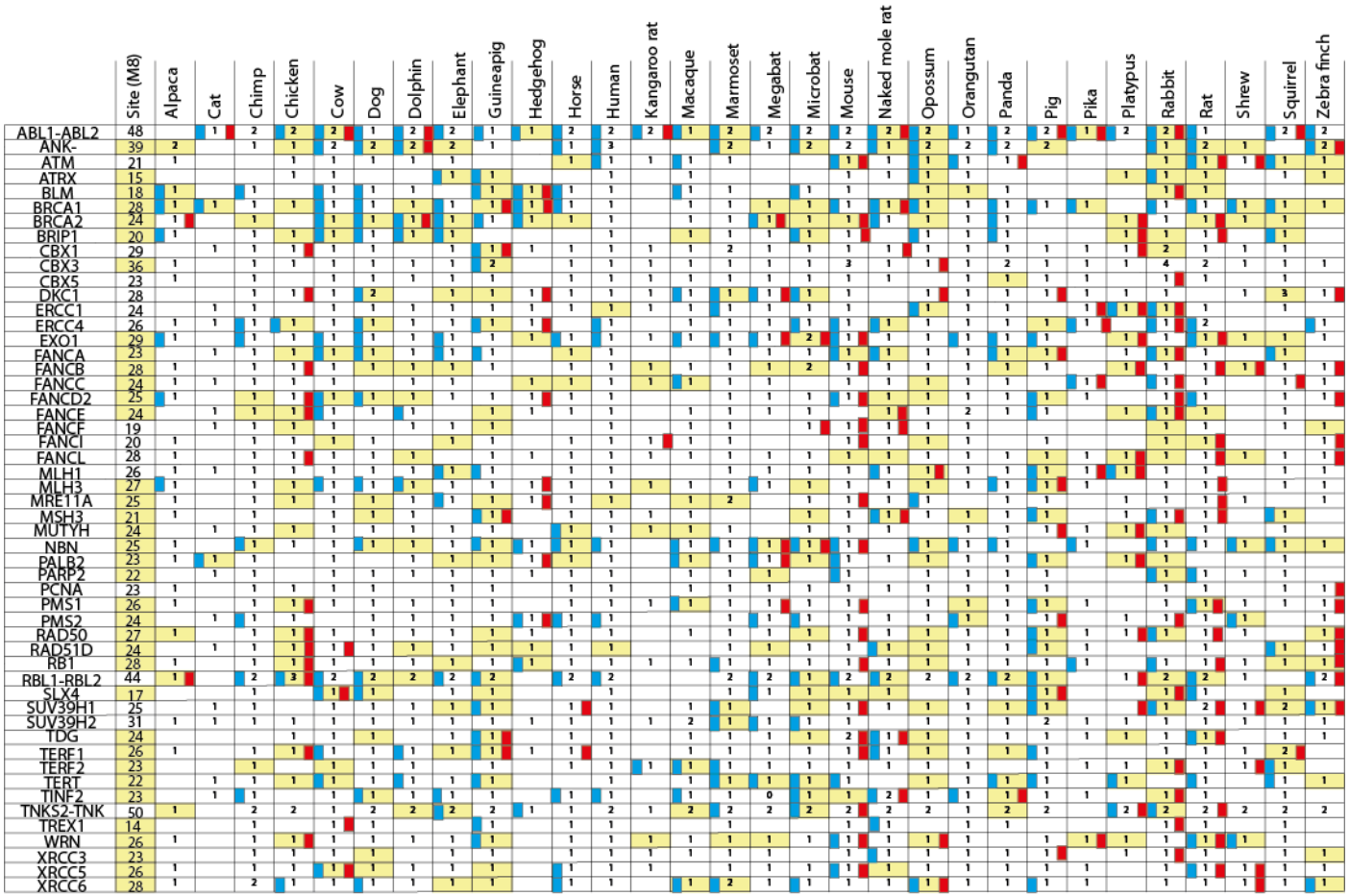


Figure 4.5 Positive selection, Recombination and GC3 deviation across Lineages.

Figure 4.5 legend: Telomere associated gene names are listed in left column, species are shown at the top. Inside each cell is the number of times a species is represented in a dataset. Yellow cells signify positive selection in sites for the second column (M8) and in lineages for all other columns. Blue boxes indicate species where recombination events have been detected and red boxes represent species who have possible gBGC within their alignment.

It is difficult to tease apart signatures for positive selection from recombination as accurate break point detection is difficult (Posada 2002), and positive selection can occur within recombinant regions (Orsi et al. 2007). The proportions of genes that have undergone positive selection and also show evidence of recombination are shown in Figure 4.6(A). No recombination events were detected in human or kangaroo rat, however the general level of recombination was extremely high in genes where positive selection was also identified. Over 50% of genes that showed signatures of positive selection also displayed signal for recombination: macaque (6/9), marmoset (7/8), rat (6/11), naked mole rat (8/11), guinea pig (13/25), squirrel (8/16), pika (2/2), cow (8/9), pig (11/14), dog (9/16), cat (2/2), shrew (4/8) and hedgehog (5/7).

The results presented here indicate that high levels of lineage-specific positively selected sites may not be under adaptive evolution and could be the product of non-adaptive processes. Elevated fixation of mutations as a result of gBGC can distrupt the underlying ω calculation and generate false positives (Ratnakumar et al. 2010). According to these data, none of the genes identified under positive selection in any of the 5 primates, kangaroo rat, dog and elephant had increased mutational fixation as a result of gBGC. The highest proportion (> 50% of genes) of elevated GC3 in genes identified under lineage-specific positive selection were observed in rat, pika and platypus.

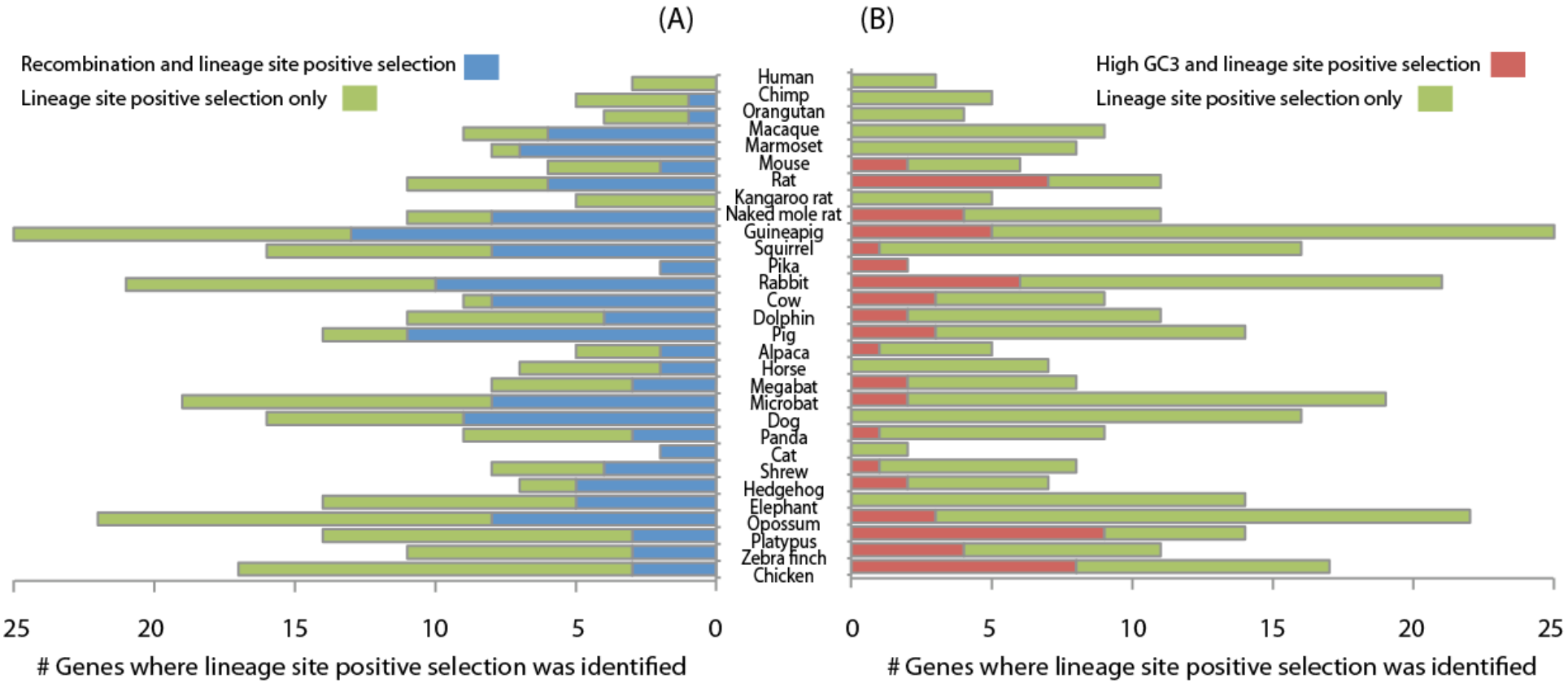


Figure 4.6 Proportion of lineages with ω > 1, recombination events and high GC3 content.

The list of species is shown in the centre of (A) and (B) .The proportion of ω > 1 (green), recombination events (blue) and GC3 content deviating above the mean (red) is shown.

4.3.6 Functional annotation on microbat candidate genes

Initially there were 19 genes identified as being under lineage-specific positive selection in the microbat lineage. In cases where the dataset fails the LM test, or the model of evolution does not adequately describe the composition of species, the phylogenetic confidence is reduced. Therefore, 5 genes were removed from further functional analysis based on low confidence in the phylogeny (EXO1, SUV39H1, TDG, TINF2 and TNKS2-TNK). There were 2 genes that displayed signatures of positive selection that may have been influenced by gBGC (EXO1 and NBN) and these were not analysed. High levels of recombination have been shown to impact LRT analysis and therefore genes where recombination was detected have been excluded from further functional analyses (ANK-Genes, BRIP1, DKC1, NBN, RBL1-RBL2, TERT, TINF2, TNKS2-TNK). A summary of the exclusion criteria is shown in Table 4.6.

In total, eight genes involved in telomere maintenance were identified as showing strong signs of lineage-site positive selection in the microbat. These are, in so far as is possible to detect, free from factors such as incorrect phylogeny, gBGC or recombination that could cause false positive results. For a summary of parameter estimation and positions of BEB positively selected sites - see Table 4.7.

Table 4.6 Identification of Microbat Candidate Genes

Positively Selected gene	Phylogeny Confidence	Possible gBGC	Recombination Detected	Candidate Gene
ANK-Genes	High	No	Yes	No
BRCA1	High	No	No	Yes
BRCA2	High	No	No	Yes
BRIP1	High	No	Yes	No
DKC1	High	No	Yes	No
EXO1	Low	Yes	No	No
FANCB	High	No	No	Yes
MLH3	High	No	No	Yes
MSH3	High	No	No	Yes
NBN	High	Yes	Yes	No
PALB2	High	No	No	Yes
RAD50	High	No	No	Yes
RBL1-RBL2	High	No	Yes	No
SLX4	High	No	No	Yes
SUV39H1	Low	No	No	No
TDG	Low	No	No	No
TERT	High	No	Yes	No
TINF2	Low	No	Yes	No
TNKS2-TNK	Low	No	Yes	No
Total Excluded	5/19	2/19	8/19	11/19

Table 4.7 Summary of Lineage-site results on Microbat Candidate Genes.

Gene Name	lnL	Parameter Estimation	BEB Sites	Sufficient Functional Information
BRCA1	-70598.50	p0=0.39612 p1=0.53457 p2=0.02950 p3=0.03981 ω0=0.21609 ω1=1.00000 ω2=3.52551	133, 347, 455, 1243, 1423, 1472	Yes
BRCA2	-113666.01	p0=0.47874 p1=0.51112 p2=0.00491 p3=0.00524 ω0=0.21191 ω1=1.00000 ω2=725.60726	54, 99, 253, 454, 470, 528, 602, 603, 604, 689, 1032, 1134, 1135, 1539, 1599, 1734, 1807, 1841I 1929, 1930, 1956, 2088, 2100, 2335, 2366, 2428, 3345, 3350	Yes
FANCB	-31847.86	p0=0.50293 p1=0.49198 p2=0.00257 p3=0.00252 ω0=0.25478 ω1=1.00000 ω2=373.09404	129, 414, 643, 781	No
MLH3	-46208.31	p0=0.52735 p1=0.46841 p2=0.00225 p3=0.00199 ω0=0.11786 ω1=1.00000 ω2=998.99970	160, 390	No
MSH3	-26594.47	p0=0.75650 p1=0.23603 p2=0.00570 p3=0.00178 ω0=0.10346 ω1=1.00000 ω2=48.48211	452	No
PALB2	-39874.60	p0=0.50554 p1=0.48795 p2=0.00331 p3=0.00320 ω0=0.20889 ω1=1.00000 ω2=73.35118	690, 1167	No
RAD50	-28862.10	p0=0.88511 p1=0.11318 p2=0.00152 p3=0.00019 ω0=0.07252 ω1=1.00000 ω2=84.92669	1094	No
SLX4	-46004.12	p0=0.49536 p1=0.49827 p2=0.00318 p3=0.00319 ω0=0.13148 ω1=1.00000 ω2=249.39697	817, 1149	No

The positively selected sites in the 8 candidate genes were compared against the human sequences in the Swiss-Prot (UniProt 2012) database to assign function through homology. There was only sufficient functional information for 2 of the 8 genes (BRCA1 and BRCA2) from which to do an in-depth study and these have been detailed in section 4.3.6.1.

4.3.6.1 BRCA1 and BRCA2 Microbat Gene Analysis

Fanconi anemia/BRCA pathway is a major DNA repair pathway (D'Andrea and Grompe 2003). FANCB is a subunit of a multi-protein core complex, which activates FANCD2 in response to DNA damaged (Soulier et al. 2005). FANCD2 then co-operates with BRCA1 and BRCA2 to repair damaged DNA (D'Andrea and Grompe 2003). All three of these genes have been identified, as evolving under positive selection within microbat however there is only functional information available for the discussion of BRCA1 and BRCA2.

Selective pressure analyses revealed 6.3% of sites in the microbat BRCA1 sequence evolving under strong positive selection with $\omega = 3.54$. The 6 positively selected sites identified using BEB were compared to homologous positions in human Swiss-Prot entry (P38398). Interaction regions, specific domains and mutagenesis sites are displayed on Figure 4.7(A). There were 64 naturally occurring cancer causing variants listed on Swiss-Prot (UniProt 2012), these were not included in the Figure 4.9(A) to avoid cluttering the results. Positively selected site 1423 is between mutagenesis sites 1423 and 1457 and has been shown *in vitro* to reduce phosphorylation by ATM related kinase when mutated, thus impacting the DNA repair pathway (Tibbetts et al. 2000). The 5 other positively selected sites were distributed throughout the BRCA1 gene but no further functional annotation could be gleaned from available data.

There were 28 sites or 1.01% of sites identified as positively selected in the microbat BRCA2 sequence, and they were compared to human Swiss-Prot entry (P51587) for further functional annotation. Four of these positively selected sites were found neighbouring and within BRCA2-NPM1 interacting regions (H. F. Wang et al. 2011). Three positively selected sites were located within the region functionally annotated to interact with FANCD2 a central protein in the Fanconi aneamia pathway, see Figure

4.7(B). Positively selected sites 3345 and 3350 were found in close proximity to human mutagenesis site T3387A (Bahassi et al. 2008). This site is critical for phosphorylation of interacting proteins CHK1 and CHK2 and subsequent recruitment of RAD51 to sites of DNA dammage (Bahassi et al. 2008).

The microbat lineage is undergoing strong positive selection at specific sites in the BRCA1 and BRCA2 genes. Functional annotation through homology indicates that positive selection is acting on regions involved in protein-protein interaction aswell as catalytically important sites.

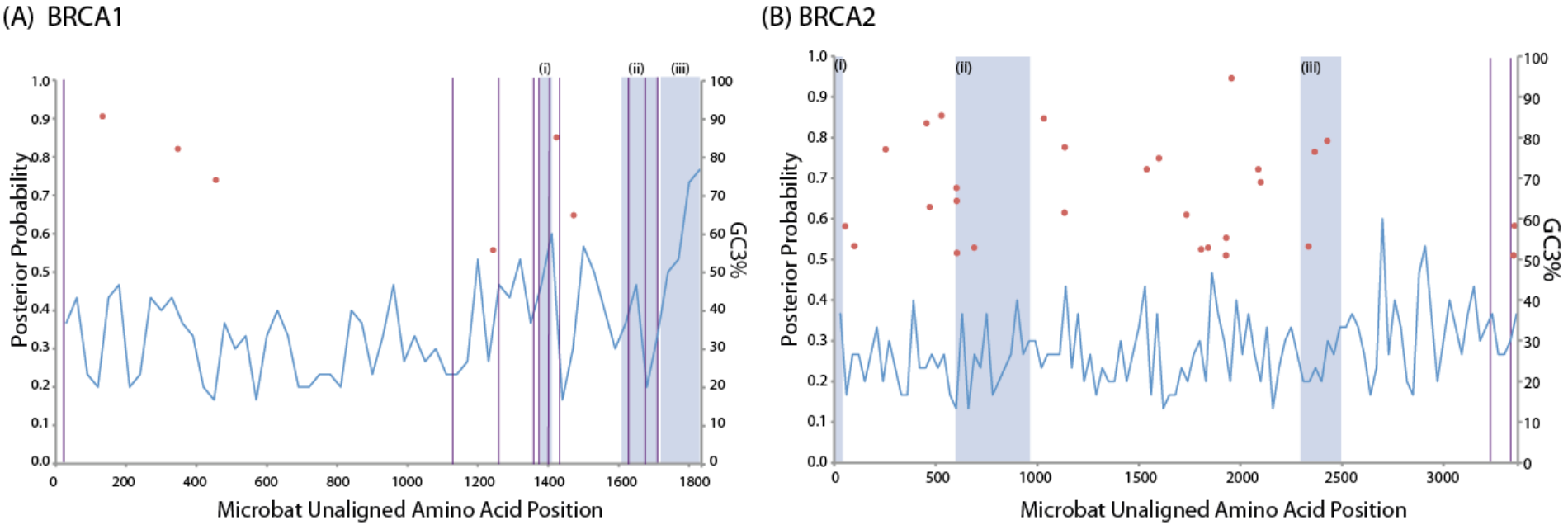


Figure 4.7 Lineage-specific position selection on BRCA1 and BRCA2 with Microbat treated as foreground

The amino acid position along (A) the BRCA1 protein and (B) the BRCA2 protein are shown on the x-axis. The position of the positively selected sites (red dots) and their corresponding posterior probability is given on the Y-axis to the left. The results of a sliding window analysis of the GC3% along the alignment is the blue line jagged line, values of GC3% are given on the right Y-axis. In panel (A) for BRCA1 the blue shaded regions (i), (ii) and (iii) correspond to the PALB2 interaction region, BRCT1 and BRCT2 Domains respectively, and the vertical Purple lines represent mutagenesis sites that impact BRCA1 function. In panel (B) the corresponding information for the BRCA2 protein is displayed, in this case the blue shaded regions (i), (ii) and (iii) are PALB2, NPM1 and FANCD2 interaction regions respectively.

4.4 Discussion

Recently, simulation studies have shown that alignment methods can give very dissimilar MSAs depending on the quantity of insertion or deletion events (Blackburne and Whelan 2012), which has been shown to impact phylogeny and detection of positive selection (Markova-Raina and Petrov 2011, Fletcher and Yang 2009, Schneider et al. 2009). Most recently it has been shown that the alignment method can also impact upon the level of false positive detection of positive selection using CodeML (Whelan and Blackburne 2012). In this Chapter, I have applied both sequence based and phylogeny aware alignment methods to the dataset in order to ensure appropriate alignment generation, accurate phylogenetic resolution, and to reduce of the detection of false positives.

Assessing the impact of phylogeny of the selective pressure analysis the simple approach taken in this chapter shows that in the application of the gene tree rather than the species tree impacts only subtly upon the estimates of parameters when using model M8. A greater impact is seen when applying lineage-specific models (modelA) with differences seen in the number of lineages detected as well as sites identified as positively selected. This is because the phylogeny gives directionality to detection of selective pressure variation and if species are incorrectly clustered together, the ancestral sites will be incorrectly reconstructed therefore impeding accurate detection of when and where adaptive evolutionary events have occurred. Therefore the assessment of phylogenetic conflict or compositional heterogeneity under the empirical protein models JTT+I and JTT+4Γ gave an indication of how accurate the phylogeny was reconstructed for each gene. Previous studies along with work carried out in this thesis address the impact of phylogeny on CodeML analyses (Anisimova et al. 2003), therefore in cases where there is reduced confidence in the phylogenetic output this in turn leads to reduced confidence in results from CodeML (Yang 1997, Yang et al. 1998) as the codon models applied require directionality from a phylogeny.

Previous studies have demonstrated that genes whose protein products are highly connected (have a high degree) and are central in a network evolve more slowly than genes that are less connected and at the network periphery (Aris-Brosou 2005, Fraser et al. 2002, Hahn and Kern 2005, Vitkup et al. 2006, Hudson and Conant 2011). In this Chapter I show that this is not the case for these telomere proteins, and that adaptive

evolutionary events are occurring across different functional categories of proteins involved in regulating the telomeres in vertebrates. The results suggest that these genes are potentially co-evolving, as work together to regulate telomeres, maintain chromosomal integrity and ensure that cancer causing mutations are removed, i.e. improving anti-cancer mechanisms within mammal species. There were 52 telomere-regulating genes tested for signatures of adaptive evolution and 43 of those are involved directly in the functional category of DNA repair. These DNA repair proteins are part of a much larger protein-protein interaction network (Begley and Samson 2004). It is possible that if additional protein-protein interacting DNA repair proteins that interact with these proteins for different functions were analysed a different pattern of positive selection versus connectivity may emerged.

The impact of non-adaptive mutations on the level of false positive detection of positive selection was also assessed by quantifying the amount of gBGC (as measured by the percentage GC3 – GC3%) and recombination present in these data. Calculating recombination breakpoints is not easy, and published methods have mixed levels of success in accurately identifying recombinant sequences and exact breakpoints positions (Posada and Crandall 2001), therefore it was difficult to interpret whether positively selected sites overlapping with recombination events were in fact false positives. Using the methods described in section 4.27 recombination was detected in 47/52 of genes tested and was present in 28/30 of the lineages were positive selection was also identified. The GC3% was elevated in 20 out of 30 lineages and possible gBGC was present in at least one lineage in 49 of the 52 telomere-associated genes.

To date, there have been no studies that indicate how to successfully separate gBGC from positive selection. The approach taken here directly estimates both gBGC and positive selection and overlays the results, in this way it is possible to identify where high GC3% may result in the fixation of mutations that could be incorrectly identified as positively selected. In summary, the level of non-adaptive evolutionary events and adaptive evolutionary events acting on these telomere-associated genes are high. Work by Lynch (2007) demonstrate that non-adaptive evolutionary forces such as recombination, and deleterious mutations are also major contributing factors in shaping network complexity (Lynch 2007), therefore both forces could be acting on the telomere interacting network of genes.

The rate of synonymous (Ds) substitutions within a coding gene, in theory should be free of selective pressures however, saturation of substitutions at the synonymous level has been previously reported (Gojobori 1983). To assess if the synonymous substitutions rates within the telomere associated genes tested are saturated it is possible to look at a pair wise comparison of Ds substitutions across all species using the Nei-Gojobori (NG) method (Nei and Gojobori 1986) in CodeML (Yang 1997, Yang et al. 1998). If a Ds pairwise score between species is > 2 it is an indication that saturation at the synonymous substitution level has taken place and therefore estimation of ω could be susceptible to false positives. However, previous studies have shown that saturation of synonymous substitutions does not tend to increase the proportion of false positive selection using LRT (Anisimova et al. 2001).

DNA repair mechanisms have been shown to vary across species (Britten 1986, Hart and Setlow 1974). Placental mammals have developed lineage specific mechanisms of regulating their cell proliferation and thus evading cancer (Seluanov et al. 2007, L. Wang et al. 2011). The results of this analysis indicate that adaptive evolutionary changes have occurred in all lineages, where species represented in 19 to 52 telomere maintenance genes had between 5.76% to 51.02% of the dataset showing signatures of positive selection. Non-adaptive evolutionary changes such as gBGC and recombination have been shown to occur in 28 out of the 30 species tested to the exclusion of human and kangaroo rat, however there are other factors such as effective population size (see section 1.1.5) that were not explored due to a lack of N_e information for many mammal populations, and that may cause fixation of mutations that are not the result of positive selection (Eyre-Walker 2002, Woolfit and Bromham 2003). In summary both adaptive and non-adaptive changes have played a role in the evolution of these genes in a lineage-specific manner.

A focused study on microbat BRCA1 and BRCA2 genes determined through homology that positive selection is acting on functionally important regions. The debate on the validity of inference of function through orthology is ongoing (Nehrt et al. 2011, Altenhoff et al. 2012, Dessimoz et al. 2012) and is discussed in section 1.1.7. Both BRCA1 and BRCA2 have been analysed by previous studies (O'Connell 2010) which found that M8 did not fit BRCA1 following LRT analysis, but that BRCA2 had 4.88% of sites evolving with ω = 2.33. In the reanalysis of these genes both BRCA1 and BRCA2 have been identified under the M8 model of evolution as having 5.2% and

4.36% with ω values of 1.91 and 1.97 respectively. The models applied in both studies are the same; the difference is the taxa sampled. The O'Connell (2010) study of the BRCA1 gene employed 10 species, 9 of which overlapped with the 28 species employed in the reanalysis of BRCA1 and all 8 species used in the O'Connell (2010) analysis of the BRCA2 gene overlap with the 24 species used in this study. Denser taxa sampling of clades allows more accurate reconstruction of phylogeny (Heath et al. 2008), which in turn impacts analysis of selective pressure variation using phylogeny dependent methods such as CodeML (Yang 1997, Yang et al. 1998).

In conclusion the application of an accurate phylogeny and densely sampled clades is fundamental in selective pressure analyses. Both adaptive and non-adaptive evolutionary events have played a role in shaping lineage-specific changes in genes involved in telomere maintenance, DNA repair, maintenance of chromosomal integrity and potentially lineage-specific cancer evasion.

Chapter 5

5 Discussion

5.1 Discussion

Since molecular sequence data was recognized to contain information pertaining to the genealogical history of species it has become increasing popular as a character choice in phylogeny reconstruction. The human genome was completed at the start of the century (Lander et al. 2001, Venter et al. 2001) and since then reduction in costs and improvements in sequencing techniques have heralded the age of phylogenomics giving rise to ambitious ventures such as the 10,000 vertebrate genomes project (http://genome10k.soe.ucsc.edu/). Using completed mammal genome data made available through Ensembl (http://www.ensembl.org/) and international sequencing consortia such as the BGI (B. Li et al. 2011), I sought to use a phylogenomics approach to resolve the root of the placental mammal phylogeny and then apply this robust phylogeny to determine the selective pressure variation acting on a network of telomere maintenance genes that play a dominant role in cancer evasion.

When datasets are of sufficient size (sequence length and taxon coverage), when they contain enough phylogenetically informative sites and the model of evolution is adequate in describing the composition and exchange rate of characters within the data then it is possible to address these complicated phylogenetic problems. Therefore, the constant theme of this body of work hinges on the importance of implementing these criteria to address the resolution of the placental mammal phylogeny.

In this thesis I have employed posterior predictive simulations to assess the fit of the model to the composition of the data, and while the phylogenetic models employed successfully model compositional heterogeneity there were other aspects of the model that were not assessed for goodness of fit, such as the rate matrix, invariant parameters and gamma parameters. Previous studies have employed posterior predictive simulations to assess the global fit of the model to the data (Bollback 2002), but this has not been performed in this thesis, therefore there maybe aspects of the model that are not fully adequate in describing the data.

Recent advances in models that allow for exchange rate heterogeneity across the phylogeny (Foster 2004) and the data (Lartillot and Philippe 2004) have made it possible to accommodate dataset heterogeneity and for the first time, these sophisticated

models have been applied to placental mammal data. In Chapter 2 I was able to show, in the re-analysis of a major publication on the placental mammal data (Murphy et al. 2001a), that not only did the models that accommodated heterogeneity over the phylogeny and data fit the data better than previously employed homogeneous models, but that these parameter rich models were able to adequately model the composition of the data while the homogeneous model of evolution (GTR+I+4Γ) was not.

The reanalysis of the Murphy et al (2001) dataset (66TaxonSet) unearthed the issue of having a dataset that was large enough to accommodate the parameter rich models and contained enough information to reject competing hypotheses. The requirement of dataset of suitable size (Quang et al. 2008) and phylogenetic informativeness (Simmons et al. 2004) has been demonstrated in previous studies. Therefore it was necessary to assemble of a novel dataset (39TaxonSet) using all available genomes and strict ortholog identification methods to effectively answer the problem of the mammal phylogeny. Using the dataset which I generated along with heterogeneous models that accommodate for heterogeneity across the phylogeny and the data I found strong statistical support for the Atlantogenata hypothesis (common ancestor of Xenarthra plus Afrotheria is the sister group to all other placental mammals), and through BF comparisons was able to fully rejected all other competing hypothesis for the root of the placental mammals.

While the 39TaxonSet dataset was capable in resolving the deep diverging nodal positions of the placental Superorders it was unable to address the issue of the intra-ordinal placements within the Laurasiatheria. Shallow relationships among mammals are frequently resolved using nucleotide data (Montgelard et al. 2008, Delsuc et al. 2002, Perelman et al. 2011), which accumulates mutations faster than amino acid sequences (Brown et al. 1982) therefore both nucleotide and amino acid datasets were assembled to assess whether one data type out performed the other in resolving the Laurasitheria intra-order placements. Hallstrom and Janke (2010) used both character types in their analysis of the Laurasiahteria, but focused on extensive sequence sampling and the generation of an enormous Supermatrix (2,863,797 bp) (Hallstrom and Janke 2010), while I focused removing data where large proportions of phylogenetic conflict lay and enforced criteria to ensure adequate fit of the composition of the data to the model, as models that do not fit the composition of the data have been shown to result in the wrong phylogeny (Felsenstein 1978, Sullivan and Swofford 1997).

Regardless of these approaches, the result from both Hallstrom and Jankes (2010) and the analysis presented in Chapter 2 of this thesis demonstrate that whether nucleotide or amino acid data are employed the problem of the intra-ordinal placement within the Laurasiatheria, neither are able to converge on one hypothesis and reject all the others. Hallstrom and Janke (2010) suggest that the problem of the Laurasiatheria phylogeny results from introgression of gene flow between Orders (Hallstrom and Janke 2010). While this is a probable cause, the fact remains that the Insectivora and Chiroptera Orders are sparsely sequenced and no species has yet been sequenced from the Pholidota order, which means stochastic errors could be a major contributing factor to the resolution of the Laurasiatheria intra-orders.

While the age of phylogenomics has helped alleviate the problem of stochastic errors from sequence sampling (Delsuc et al. 2005), stochastic errors rising from insufficient taxon sampling are still at large (Hedtke et al. 2006). Currently, there is less than 1% of extant mammals on our planet that have their nuclear genome fully sequenced. To overcome this taxon deficiency, multiple studies have incorporated mitochondrial sequence data to resolve phylogenies (Kjer and Honeycutt 2007, Tobe et al. 2010, Hillis and Wilcox 2005, Hyman et al. 2007). In Chapter 3 I explored the suitability of mitochondrial data as a phylogentic marker for placental mammals and attempted to separate the useful phylogenetic signal from phylogenetic conflict by partitioning the data in various ways addressing issues such as missing data (Lemmon et al. 2009, Kearney 2002), saturation of characters (Cummins and McInerney 2011), removal of rogue taxa (Sanderson 2002) and sampling at various phylogenetic depths to account for homoplasy at deeper nodes and assess where phylogenetic breakdown occurs. While sampling at different phylogenetic depths and analysing genes at higher coverage resulted in a decrease in phylogenetic conflict, there were still large numbers of unresolved nodes and more importantly incongruence observed between phylogenies inferred from different gene sets - indicating that these phylogenies contained large proportions of systematic error (Pisani et al. 2007). Therefore while increased taxon sampling has been shown to be successful in previous studies (Hedtke et al. 2006) and there is still considerable interest in using mitochondrial data for phylogenetic analyses such as the International Bar code of Life (iBOL) project (Vernooy et al. 2010) as well as many recent publications using mitochondrial data to resolve species phylogeny (Ozdil and Ilhan 2012, Waeschenbach et al. 2012, White et al. 2011). The results from Chapter 3 indicate that mitochondrial data is an unsuitable phylogenetic marker and

treatments of the data are not sufficient to remove the high levels of systematic errors. As the number of fully sequenced nuclear genomes becomes available as well as advancement of other methods using microRNAs, gene order and gene content, it will become less necessary to rely on mitochondrial as a phylogenetic marker.

Adaptation of mammals to a diverse variety of ecological niches, along with their observed variations in life traits, has lead to heterogeneous rates of change across genes, between genomes (nuclear and mitochondrial) and among lineages. Results from both Chapter 2 and 3 demonstrate that if the data being tested for selective pressure variation does not contain adequate phylogenetic signal or the model of evolution does not account for compositional heterogeneity then incorrect phylogeny reconstruction is probable. The downstream impact of this is assessed in Chapter 4 in the analyses of selective pressure variation across the mammal phylogeny. The results showed that when the M8 selection model was applied to both species and gene phylogeny, differences were observed in the lnL values, proportion of sites estimated under selection, ω values and the number of positively selected sites identified under the BEB model. When the lineage-site modelA was applied, differences were also observed between the numbers of lineages identified as evolving under positive selection along with differences in the estimates of proportions of sites and ω values. While the species tree performed marginally better than the gene tree in 21/34 of the Single gene orthologous families analysed, it is not correct to conclude that a species tree should be applied to all future selective pressure analyses of single gene orthologs, as the placement of Laurasiatheria Orders has not yet been accomplished (Chapter 2). Instead, the fundamental message of Chapter 4 is to ensure every precaution is taken to reduce the effect of alignment errors, systematic errors and stochastic errors in phylogenetic reconstruction, and to ensure the most accurate assessment of selective pressure variation using the ML method implemented in CodeML (Yang 1997, Yang et al. 1998).

While studies have shown that adaptive evolution has helped shaped gene networks, and the strength of selective pressure acting on these genes is correlated with protein-protein connectivity (Aris-Brosou 2005, Fraser et al. 2002, Hahn and Kern 2005, Vitkup et al. 2006, Hudson and Conant 2011), work by Lynch (2007) demonstrates a role for non-adaptive evolutionary forces in shaping gene networks. In Chapter 4, both lineage-site specific adaptive evolution and non-adaptive evolutionary events such as

recombination and possible gBGC were identified as possible forces in shaping these telomere maintenance genes. Extremely high levels of positive selection were observed (47/52 genes) when the M8 model was applied to the telomere maintenance proteins. However, the level of selection did not correlate with protein-protein connectivity as has been seen before in the analysis of metabolic networks (Vitkup et al. 2006). These results indicate that selective pressures are acting on these genes regardless of their degree or level of connectivity. It is of course a small sample size and if more genes were included in this study or these genes were analysed based on their hierarchical position within a wider functional network, it is possible that a different pattern could emerge.

Lineage-site specific analysis of selective pressure variation as well as analysis of lineage specific non-adaptive evolutionary processes demonstrated that extant lineages had varying levels of positive selection, recombination and gBGC across the network of telomere maintenance genes. A comparison of the results of a previous analysis of BRCA1 and BRCA2 (O'Connell 2010) which used a smaller number of taxa, demonstrating the impact of taxon sampling on detection of positive selection using phylogeny aware methods. From the results presented here in Chapter 4 it is evident that both adaptive and non-adaptive processes have contributed to the evolutionary history of telomere regulating proteins in mammals.

From the work presented here, it can be concluded that when addressing any phylogeny problem there are issues at every level of the analysis. To ensure accurate resolution to the phylogeny there are multiple considerations to take on board. The initial decisions regarding dataset assembly should consider sequence availability and whether there are sufficient taxa sampled, sequences covered and phylogenetically informative sites to address the phylogenetic question at hand. While it may be possible in certain cases to reduce or remove systematic errors by sampling at different nodes or by removing fast evolving sites, I have shown that this is not always possible, particularly when dealing with mitochondrial protein coding sequence data. This body of work demonstrates that heterogeneous models that account for variation in composition and rate exchange of characters over the phylogeny and the data perform better than homogeneous models in resolving difficult phylogenetic questions. It is important in phylogeny reconstruction to have phylogenetically informative data capable of rejecting all alternative hypotheses through careful and appropriate modeling. This goal was achieved here in the placement

of the placental mammal root. When the phylogeny has not been reconstructed under these specifications, it reduces confidence in the hypothesis causing increase likelihood of observing errors in selective pressure analyses. In conclusion, whether the question is directly related to phylogeny reconstruction or to the downstream analysis of selective pressure variation, dataset suitability, model selection and adequacy, along with taxa sampling are crucial.

Chapter 6

6 Bibliography

Adachi, J. and Hasegawa, M. (1996) 'Model of amino acid substitution in proteins encoded by mitochondrial DNA', *J Mol Evol,* 42(4), 459-68.

Adachi, J., Waddell, P. J., Martin, W. and Hasegawa, M. (2000) 'Plastid genome phylogeny and a model of amino acid substitution for proteins encoded by chloroplast DNA', *Journal of Molecular Evolution,* 50(4), 348-358.

Altenhoff, A. M., Studer, R. A., Robinson-Rechavi, M. and Dessimoz, C. (2012) 'Resolving the ortholog conjecture: orthologs tend to be weakly, but significantly, more similar in function than paralogs', *PLoS Comput Biol,* 8(5), e1002514.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. and Lipman, D. J. (1990) 'Basic local alignment search tool', *J Mol Biol,* 215(3), 403-10.

Amrine-Madsen, H., Koepfli, K. P., Wayne, R. K. and Springer, M. S. (2003) 'A new phylogenetic marker, apolipoprotein B, provides compelling evidence for eutherian relationships', *Mol Phylogenet Evol,* 28(2), 225-40.

Anderson, E. (1949) *Introgressive Hybridization,* New York: Wiley.

Anderson, J. B. and Kohn, L. M. (1998) 'Genotyping, gene genealogies and genomics bring fungal population genetics above ground', *Trends Ecol Evol,* 13(11), 444-9.

Anderson, S., Bankier, A. T., Barrell, B. G., de Bruijn, M. H., Coulson, A. R., Drouin, J., Eperon, I. C., Nierlich, D. P., Roe, B. A., Sanger, F., Schreier, P. H., Smith, A. J., Staden, R. and Young, I. G. (1981) 'Sequence and organization of the human mitochondrial genome', *Nature,* 290(5806), 457-65.

Ane, C., Burleigh, J. G., McMahon, M. M. and Sanderson, M. J. (2005) 'Covarion structure in plastid genome evolution: a new statistical test', *Mol Biol Evol,* 22(4), 914-24.

Anisimova, M., Bielawski, J. P. and Yang, Z. (2001) 'Accuracy and power of the likelihood ratio test in detecting adaptive molecular evolution', *Mol Biol Evol,* 18(8), 1585-92.

Anisimova, M., Bielawski, J. P. and Yang, Z. (2002) 'Accuracy and power of bayes prediction of amino acid sites under positive selection', *Mol Biol Evol,* 19(6), 950-8.

Anisimova, M., Nielsen, R. and Yang, Z. (2003) 'Effect of recombination on the accuracy of the likelihood method for detecting positive selection at amino acid sites', *Genetics,* 164(3), 1229-36.

Antolin, M. F., Jenkins, K. P., Bergstrom, C. T., Crespi, B. J., De, S., Hancock, A., Hanley, K. A., Meagher, T. R., Moreno-Estrada, A., Nesse, R. M., Omenn, G. S. and Stearns, S. C. (2012) 'Evolution and medicine in undergraduate education: a prescription for all biology students', *Evolution,* 66(6), 1991-2006.

Aris-Brosou, S. (2005) 'Determinants of adaptive evolution at the molecular level: the extended complexity hypothesis', *Mol Biol Evol,* 22(2), 200-9.

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

Arnason, U., Gullberg, A., Janke, A. and Kullberg, M. (2007) 'Mitogenomic analyses of caniform relationships', *Molecular Phylogenetics and Evolution,* 45(3), 863-874.

Aruna, S. and Ranganath, H. A. (2006) 'Introgressive hybridization and evolution of a novel protein phenotype: glue protein profiles in the nasuta-albomicans complex of Drosophila', *J Genet,* 85(1), 25-30.

Asher, R. J. (2007) 'A web-database of mammalian morphology and a reanalysis of placental phylogeny', *BMC Evol Biol,* 7, 108.

Asher, R. J. and Helgen, K. M. (2010) 'Nomenclature and placental mammal phylogeny', *BMC Evol Biol,* 10, 102.

Augee, M. L. (2007) *The Encyclopedia of Mammals,* Oxford University Press.

Awadalla, P., Eyre-Walker, A. and Smith, J. M. (1999) 'Linkage disequilibrium and recombination in hominid mitochondrial DNA', *Science,* 286(5449), 2524-5.

Ayala, F. J. (1997) 'Vagaries of the molecular clock', *Proc Natl Acad Sci U S A,* 94(15), 7776-83.

Bahassi, E. M., Ovesen, J. L., Riesenberg, A. L., Bernstein, W. Z., Hasty, P. E. and Stambrook, P. J. (2008) 'The checkpoint kinases Chk1 and Chk2 regulate the functional associations between hBRCA2 and Rad51 in response to DNA damage', *Oncogene,* 27(28), 3977-85.

Baldauf, S. L. (2003) 'Phylogeny for the faint of heart: a tutorial', *Trends Genet,* 19(6), 345-51.

Ballard, J. W. and Whitlock, M. C. (2004) 'The incomplete natural history of mitochondria', *Mol Ecol,* 13(4), 729-44.

Bapteste, E., Lopez, P., Bouchard, F., Baquero, F., McInerney, J. O. and Burian, R. M. (2012) 'Evolutionary analyses of non-genealogical bonds produced by introgressive descent', *Proc Natl Acad Sci U S A*.

Barja, G. (1999) 'Mitochondrial oxygen radical generation and leak: sites of production in states 4 and 3, organ specificity, and relation to aging and longevity', *J Bioenerg Biomembr,* 31(4), 347-66.

Baum, B. (2002) 'Combining trees as a way of combining data sets for phylogenetic inference, and the desirability of combining gene trees', *Taxon,* 41, 3-10.

Beall, C. M., Cavalleri, G. L., Deng, L., Elston, R. C., Gao, Y., Knight, J., Li, C., Li, J. C., Liang, Y., McCormack, M., Montgomery, H. E., Pan, H., Robbins, P. A., Shianna, K. V., Tam, S. C., Tsering, N., Veeramah, K. R., Wang, W., Wangdui, P., Weale, M. E., Xu, Y., Xu, Z., Yang, L., Zaman, M. J., Zeng, C., Zhang, L., Zhang, X., Zhaxi, P. and Zheng, Y. T. (2010) 'Natural selection on EPAS1 (HIF2alpha) associated with low hemoglobin concentration in Tibetan highlanders', *Proc Natl Acad Sci U S A,* 107(25), 11459-64.

Begley, T. J. and Samson, L. D. (2004) 'Network responses to DNA damaging agents', *DNA Repair (Amst),* 3(8-9), 1123-32.

Benton, M. J., Donoghue, M. J. and Asher, R. J. (2009) 'Calibrating and constrining molecular clocks' in Hedges, S. B. and Kumar, S., eds., *The TimeTree of Life*, Cambridge: Oxford Univeristy Press, 35-86.

Berglund, J., Pollard, K. S. and Webster, M. T. (2009) 'Hotspots of biased nucleotide substitutions in human genes', *PLoS Biol,* 7(1), e26.

Bergsten, A. (2005) 'A review of long-branch attraction', *Cladisticss,* 21(2), 163-193.

Bergthorsson, U., Adams, K. L., Thomason, B. and Palmer, J. D. (2003) 'Widespread horizontal transfer of mitochondrial genes in flowering plants', *Nature,* 424(6945), 197-201.

Berthier, P., Excoffier, L. and Ruedi, M. (2006) 'Recurrent replacement of mtDNA and cryptic hybridization between two sibling bat species Myotis myotis and Myotis blythii', *Proc Biol Sci,* 273(1605), 3101-9.

Betran, E., Wang, W., Jin, L. and Long, M. (2002) 'Evolution of the phosphoglycerate mutase processed gene in human and chimpanzee revealing the origin of a new primate gene', *Mol Biol Evol,* 19(5), 654-63.

Bininda-Emonds, O. R. (2004) 'Trees versus characters and the supertree/supermatrix "paradox"', *Syst Biol,* 53(2), 356-9.

Bininda-Emonds, O. R., Cardillo, M., Jones, K. E., MacPhee, R. D., Beck, R. M., Grenyer, R., Price, S. A., Vos, R. A., Gittleman, J. L. and Purvis, A. (2007) 'The delayed rise of present-day mammals', *Nature,* 446(7135), 507-12.

Bishop, M. J. and Friday, A. E. (1985) 'Evolutionary Trees from Nucleic-Acid and Protein Sequences', *Proceedings of the Royal Society of London Series B-Biological Sciences,* 226(1244), 271-302.

Blackburne, B. P. and Whelan, S. (2012) 'Measuring the distance between multiple sequence alignments', *Bioinformatics,* 28(4), 495-502.

Blanchette, M., Kunisawa, T. and Sankoff, D. (1999) 'Gene order breakpoint evidence in animal mitochondrial phylogeny', *J Mol Evol,* 49(2), 193-203.

Blanquart, S. and Lartillot, N. (2008) 'A site- and time-heterogeneous model of amino acid replacement', *Mol Biol Evol,* 25(5), 842-58.

Blasco, M. A. (2005) 'Telomeres and human disease: ageing, cancer and beyond', *Nat Rev Genet,* 6(8), 611-22.

Bleiweiss, R. (1998) 'Slow rate of molecular evolution in high-elevation hummingbirds', *Proc Natl Acad Sci U S A,* 95(2), 612-6.

Bollback, J. P. (2002) 'Bayesian model adequacy and choice in phylogenetics', *Mol Biol Evol,* 19(7), 1171-80.

Boni, M. F., Posada, D. and Feldman, M. W. (2007) 'An exact nonparametric method for inferring mosaic structure in sequence triplets', *Genetics,* 176(2), 1035-47.

Borisenko, A. V., Lim, B. K., Ivanova, N. V., Hanner, R. H. and Hebert, P. D. (2008) 'DNA barcoding in surveys of small mammal communities: a field study in Suriname', *Mol Ecol Resour,* 8(3), 471-9.

Bousquet, J., Strauss, S. H., Doerksen, A. H. and Price, R. A. (1992) 'Extensive variation in evolutionary rate of rbcL gene sequences among seed plants', *Proc Natl Acad Sci U S A,* 89(16), 7844-8.

Brandley, M. C., Schmitz, A. and Reeder, T. W. (2005) 'Partitioned Bayesian analyses, partition choice, and the phylogenetic relationships of scincid lizards', *Syst Biol,* 54(3), 373-90.

Branger, B., Gillard, P., Monrigal, C., Thelu, S., Robidas, E., Viot, S., Descamps, P., Philippe, H. J., Sentilhes, L. and Winer, N. (2011) '[Lessons and impact of two audits on postpartum hemorrhages in 24 maternity hospitals of the network "Securite Naissance - Naitre Ensemble" in "Pays-de-la-Loire" area]', *J Gynecol Obstet Biol Reprod (Paris),* 40(7), 657-67.

Britten, R. J. (1986) 'Rates of DNA sequence evolution differ between taxonomic groups', *Science,* 231(4744), 1393-8.

Brown, W. M., Prager, E. M., Wang, A. and Wilson, A. C. (1982) 'Mitochondrial DNA sequences of primates: tempo and mode of evolution', *J Mol Evol,* 18(4), 225-39.

Burger, G., Gray, M. W. and Lang, B. F. (2003) 'Mitochondrial genomes: anything goes', *Trends Genet,* 19(12), 709-16.

Butlin, R. K., Galindo, J. and Grahame, J. W. (2008) 'Review. Sympatric, parapatric or allopatric: the most important way to classify speciation?', *Philos Trans R Soc Lond B Biol Sci,* 363(1506), 2997-3007.

Campbell, L. I., Rota-Stabelli, O., Edgecombe, G. D., Marchioro, T., Longhorn, S. J., Telford, M. J., Philippe, H., Rebecchi, L., Peterson, K. J. and Pisani, D. (2011) 'MicroRNAs and phylogenomics resolve the relationships of Tardigrada and suggest that velvet worms are the sister group of Arthropoda', *Proc Natl Acad Sci U S A,* 108(38), 15920-4.

Cantrell, M. A., Filanoski, B. J., Ingermann, A. R., Olsson, K., DiLuglio, N., Lister, Z. and Wichman, H. A. (2001) 'An ancient retrovirus-like element contains hot spots for SINE insertion', *Genetics,* 158(2), 769-77.

Capella-Gutierrez, S., Silla-Martinez, J. M. and Gabaldon, T. (2009) 'trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses', *Bioinformatics,* 25(15), 1972-3.

Carulli, J. P., Krane, D. E., Hartl, D. L. and Ochman, H. (1993) 'Compositional heterogeneity and patterns of molecular evolution in the Drosophila genome', *Genetics,* 134(3), 837-45.

Castoe, T. A., de Koning, A. P., Kim, H. M., Gu, W., Noonan, B. P., Naylor, G., Jiang, Z. J., Parkinson, C. L. and Pollock, D. D. (2009) 'Evidence for an ancient adaptive episode of convergent molecular evolution', *Proc Natl Acad Sci U S A,* 106(22), 8986-91.

Caterino, M. S., Reed, R. D., Kuo, M. M. and Sperling, F. A. (2001) 'A partitioned likelihood analysis of swallowtail butterfly phylogeny (Lepidoptera:Papilionidae)', *Syst Biol,* 50(1), 106-27.

Caulin, A. F. and Maley, C. C. (2011) 'Peto's Paradox: evolution's prescription for cancer prevention', *Trends Ecol Evol,* 26(4), 175-82.

Chakrabarti, S., Lanczycki, C. J., Panchenko, A. R., Przytycka, T. M., Thiessen, P. A. and Bryant, S. H. (2006) 'Refining multiple sequence alignments with conserved core regions', *Nucleic Acids Research,* 34(9), 2598-606.

Chen, F. C. and Li, W. H. (2001) 'Genomic divergences between humans and other hominoids and the effective population size of the common ancestor of humans and chimpanzees', *Am J Hum Genet,* 68(2), 444-56.

Chen, R. and Jeong, S. S. (2000) 'Functional prediction: identification of protein orthologs and paralogs', *Protein Sci,* 9(12), 2344-53.

Churakov, G., Kriegs, J. O., Baertsch, R., Zemann, A., Brosius, J. and Schmitz, J. (2009) 'Mosaic retroposon insertion patterns in placental mammals', *Genome Research,* 19(5), 868-75.

Conant, G. C. and Wolfe, K. H. (2008) 'Turning a hobby into a job: how duplicated genes find new functions', *Nature Reviews Genetics,* 9(12), 938-50.

Cooke, M. S., Evans, M. D., Dizdaroglu, M. and Lunec, J. (2003) 'Oxidative DNA damage: mechanisms, mutation, and disease', *FASEB J,* 17(10), 1195-214.

Cotton, J. A. and Wilkinson, M. (2007) 'Majority-rule supertrees', *Syst Biol,* 56(3), 445-52.

Creevey, C. J. and McInerney, J. O. (2002) 'An algorithm for detecting directional and non-directional positive selection, neutrality and negative selection in protein coding DNA sequences', *Gene,* 300(1-2), 43-51.

Creevey, C. J. and McInerney, J. O. (2005) 'Clann: investigating phylogenetic information through supertree analyses', *Bioinformatics,* 21(3), 390-2.

Cummins, C. and 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*.

Cummins, C., McInerney, J.O. (2010) 'TIGER: Understanding evolutionary rate variation'.

Currat, M. and Excoffier, L. (2004) 'Modern humans did not admix with Neanderthals during their range expansion into Europe', *PLoS Biol,* 2(12), e421.

D'Andrea, A. D. and Grompe, M. (2003) 'The Fanconi anaemia/BRCA pathway', *Nat Rev Cancer,* 3(1), 23-34.

Darling, A., carey, L., Feng, W., (2003) 'The design, implementation, and evaluation of mpiBLAST.', in *4th International Conference on Linux Clusters: The HPC Revolution 2003 in conjunction with ClusterWorld Conference & Expo, June 2003.*, June 2003,

Darwin, C. (1859) *The Origin of Species,* London: John Murray.

Darwin, C. (1871) *The Desent of Man, and Selection in Relation to Sex,* London: John Murray.

Dayhoff, M. O., Schwartz,R.M. , Orcutt,B. (1978) 'A model of evolutionary change in proteins', *Atlas of Protein Sequence and Structure National Biomedical Research Foundation*.

de Lange, T. (2002) 'Protection of mammalian telomeres', *Oncogene,* 21(4), 532-40.

de Magalhaes, J. P. and Costa, J. (2009) 'A database of vertebrate longevity records and their relation to other life-history traits', *Journal of Evolutionary Biology,* 22(8), 1770-1774.

de Queiroz, A. and Gatesy, J. (2007) 'The supermatrix approach to systematics', *Trends Ecol Evol,* 22(1), 34-41.

Degnan, J. H. and Rosenberg, N. A. (2009) 'Gene tree discordance, phylogenetic inference and the multispecies coalescent', *Trends Ecol Evol,* 24(6), 332-40.

Delsuc, F., Brinkmann, H. and Philippe, H. (2005) 'Phylogenomics and the reconstruction of the tree of life', *Nat Rev Genet,* 6(5), 361-75.

Delsuc, F., Scally, M., Madsen, O., Stanhope, M. J., de Jong, W. W., Catzeflis, F. M., Springer, M. S. and Douzery, E. J. (2002) 'Molecular phylogeny of living xenarthrans and the impact of character and taxon sampling on the placental tree rooting', *Mol Biol Evol,* 19(10), 1656-71.

Dessimoz, C., Gabaldon, T., Roos, D. S., Sonnhammer, E. L. and Herrero, J. (2012) 'Toward community standards in the quest for orthologs', *Bioinformatics,* 28(6), 900-4.

Dimmic, M. W., Rest, J. S., Mindell, D. P. and Goldstein, R. A. (2002) 'rtREV: An amino acid substitution matrix for inference of retrovirus and reverse transcriptase phylogeny', *Journal of Molecular Evolution,* 55(1), 65-73.

Dolgin, E. (2012) 'Phylogeny: Rewriting evolution', *Nature,* 486(7404), 460-2.

Donoghue, M. J. and Sanderson, M. J. (1992) 'The suitability of molecular and morphological evidence in reconstructing plant phylogeny' in Soltis, P. S., Soltis, D. E. and Doyle, J. J., eds., *Molecular Systematics of Plants*, New York: Chapman & Hall, 340-368.

Dorus, S., Evans, P. D., Wyckoff, G. J., Choi, S. S. and Lahn, B. T. (2004) 'Rate of molecular evolution of the seminal protein gene SEMG2 correlates with levels of female promiscuity', *Nat Genet,* 36(12), 1326-9.

Dreszer, T. R., Wall, G. D., Haussler, D. and Pollard, K. S. (2007) 'Biased clustered substitutions in the human genome: the footprints of male-driven biased gene conversion', *Genome Res,* 17(10), 1420-30.

du Plessis, L., Skunca, N. and Dessimoz, C. (2011) 'The what, where, how and why of gene ontology--a primer for bioinformaticians', *Brief Bioinform,* 12(6), 723-35.

Dunn, C. W., Hejnol, A., Matus, D. Q., Pang, K., Browne, W. E., Smith, S. A., Seaver, E., Rouse, G. W., Obst, M., Edgecombe, G. D., Sorensen, M. V., Haddock, S. H., Schmidt-Rhaesa, A., Okusu, A., Kristensen, R. M., Wheeler, W. C., Martindale, M. Q. and Giribet, G. (2008) 'Broad phylogenomic sampling improves resolution of the animal tree of life', *Nature,* 452(7188), 745-9.

Duret, L. and Arndt, P. F. (2008) 'The impact of recombination on nucleotide substitutions in the human genome', *PLoS Genet,* 4(5), e1000071.

Edgar, R. C. (2004) 'MUSCLE: multiple sequence alignment with high accuracy and high throughput', *Nucleic Acids Research,* 32(5), 1792-7.

Egan, J. B., Shi, C. X., Tembe, W., Christoforides, A., Kurdoglu, A., Sinari, S., Middha, S., Asmann, Y., Schmidt, J., Braggio, E., Keats, J. J., Fonseca, R., Bergsagel, P. L., Craig, D. W., Carpten, J. D. and Stewart, A. K. (2012) 'Whole-genome sequencing of multiple myeloma from diagnosis to plasma cell leukemia reveals genomic initiating events, evolution, and clonal tides', *Blood,* 120(5), 1060-6.

Embley, T. M., Thomas, R. H. and Williams, R. A. D. (1993) 'Reduced Thermophilic Bias in the 16S rDNA Sequence from Thermus ruber Provides Further Support for a Relationship Between Thermus and Deinococcus', *Systematic and Applied Microbiology,* 16(1), 25-29.

Enright, A. J., Van Dongen, S. and Ouzounis, C. A. (2002) 'An efficient algorithm for large-scale detection of protein families', *Nucleic Acids Research,* 30(7), 1575-84.

Erixon, P., Svennblad, B., Britton, T. and Oxelman, B. (2003) 'Reliability of Bayesian posterior probabilities and bootstrap frequencies in phylogenetics', *Syst Biol,* 52(5), 665-73.

Escobar, J. S., Glemin, S. and Galtier, N. (2011) 'GC-biased gene conversion impacts ribosomal DNA evolution in vertebrates, angiosperms, and other eukaryotes', *Mol Biol Evol,* 28(9), 2561-75.

Eulenstein, O., Chen, D., Burleigh, J. G., Fernandez-Baca, D. and Sanderson, M. J. (2004) 'Performance of flip supertree construction with a heuristic algorithm', *Syst Biol,* 53(2), 299-308.

Eyre-Walker, A. (2002) 'Changing effective population size and the McDonald-Kreitman test', *Genetics,* 162(4), 2017-24.

Feil, E. J., Holmes, E. C., Bessen, D. E., Chan, M. S., Day, N. P., Enright, M. C., Goldstein, R., Hood, D. W., Kalia, A., Moore, C. E., Zhou, J. and Spratt, B. G. (2001) 'Recombination within natural populations of pathogenic bacteria: short-term empirical estimates and long-term phylogenetic consequences', *Proc Natl Acad Sci U S A,* 98(1), 182-7.

Felsenstein, J. (1978) 'Cases in which parsimony or compatibility methods will be positivly misleading', *Syst Biol,* 27, 401.

Felsenstein, J. (1981) 'Evolutionary trees from DNA sequences: a maximum likelihood approach', *J Mol Evol,* 17(6), 368-76.

Felsenstein, J. (1988) 'Phylogenies from molecular sequences: inference and reliability', *Annu Rev Genet,* 22, 521-65.

Felsenstein, J. (2004) *Inferring Phylogenies,* Massachusetts.

Fisher, R. A. (1958) *The Genetical Theory of Natural Selection,* 2nd ed., New York: Oxford University Press.

Fisher-Reid, M. C. and Wiens, J. J. (2011) 'What are the consequences of combining nuclear and mitochondrial data for phylogenetic analysis? Lessons from Plethodon salamanders and 13 other vertebrate clades', *BMC Evol Biol,* 11, 300.

Fitch, W. M. (1970) 'Distinguishing homologous from analogous proteins', *Systematic Zoology,* 19(2), 99-113.

Fitch, W. M. and Margoliash, E. (1967) 'Construction of phylogenetic trees', *Science,* 155(3760), 279-84.

Fletcher, W. and Yang, Z. (2009) 'INDELible: a flexible simulator of biological sequence evolution', *Mol Biol Evol,* 26(8), 1879-88.

Flicek, P., Amode, M. R., Barrell, D., Beal, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fairley, S., Fitzgerald, S., Gil, L., Gordon, L., Hendrix, M., Hourlier, T., Johnson, N., Kahari, A. K., Keefe, D., Keenan, S., Kinsella, R.,

Komorowska, M., Koscielny, G., Kulesha, E., Larsson, P., Longden, I., McLaren, W., Muffato, M., Overduin, B., Pignatelli, M., Pritchard, B., Riat, H. S., Ritchie, G. R., Ruffier, M., Schuster, M., Sobral, D., Tang, Y. A., Taylor, K., Trevanion, S., Vandrovcova, J., White, S., Wilson, M., Wilder, S. P., Aken, B. L., Birney, E., Cunningham, F., Dunham, I., Durbin, R., Fernandez-Suarez, X. M., Harrow, J., Herrero, J., Hubbard, T. J., Parker, A., Proctor, G., Spudich, G., Vogel, J., Yates, A., Zadissa, A. and Searle, S. M. (2012) 'Ensembl 2012', *Nucleic Acids Research,* 40(Database issue), D84-90.

Force, A., Lynch, M., Pickett, F. B., Amores, A., Yan, Y. L. and Postlethwait, J. (1999) 'Preservation of duplicate genes by complementary, degenerative mutations', *Genetics,* 151(4), 1531-45.

Forslund, K., Pekkari, I. and Sonnhammer, E. L. (2011) 'Domain architecture conservation in orthologs', *BMC Bioinformatics,* 12, 326.

Foster, P. G. (2001) 'The Idiot's Guide to the Zen of Likelihood in a Nutshell in Seven Days for Dummies, Unleashed', [online], available: [accessed

Foster, P. G. (2004) 'Modeling compositional heterogeneity', *Syst Biol,* 53(3), 485-95.

Foster, P. G., Cox, C. J. and Embley, T. M. (2009) 'The primary divisions of life: a phylogenomic approach employing composition-heterogeneous methods', *Philos Trans R Soc Lond B Biol Sci,* 364(1527), 2197-207.

Foster, P. G. and Hickey, D. A. (1999) 'Compositional bias may affect both DNA-based and protein-based phylogenetic reconstructions', *J Mol Evol,* 48(3), 284-90.

Foster, P. G., Jermiin, L. S. and Hickey, D. A. (1997) 'Nucleotide composition bias affects amino acid content in proteins coded by animal mitochondria', *J Mol Evol,* 44(3), 282-8.

Fraser, H. B., Hirsh, A. E., Steinmetz, L. M., Scharfe, C. and Feldman, M. W. (2002) 'Evolutionary rate in the protein interaction network', *Science,* 296(5568), 750-2.

Frye, M. S. and Hedges, S. B. (1995) 'Monophyly of the Order Rodentia Inferred from Mitochondrial-DNA Sequences of the Genes for 12s Ribosomal-Rna, 16s Ribosomal-Rna, and Transfer-Rna-Valine', *Molecular Biology and Evolution,* 12(1), 168-176.

Fumagalli, M., Sironi, M., Pozzoli, U., Ferrer-Admetlla, A., Pattini, L. and Nielsen, R. (2011) 'Signatures of environmental genetic adaptation pinpoint pathogens as the main selective pressure through human evolution', *PLoS Genet,* 7(11), e1002355.

Gabaldon, T., Dessimoz, C., Huxley-Jones, J., Vilella, A. J., Sonnhammer, E. L. and Lewis, S. (2009) 'Joining forces in the quest for orthologs', *Genome Biol,* 10(9), 403.

Gadagkar, S. R. and Kumar, S. (2005) 'Maximum likelihood outperforms maximum parsimony even when evolutionary rates are heterotachous', *Mol Biol Evol,* 22(11), 2139-41.

Gadagkar, S. R., Rosenberg, M. S. and Kumar, S. (2005) 'Inferring species phylogenies from multiple genes: concatenated sequence tree versus consensus gene tree', *J Exp Zool B Mol Dev Evol,* 304(1), 64-74.

Galtier, N. (2003) 'Gene conversion drives GC content evolution in mammalian histones', *Trends Genet,* 19(2), 65-8.

Galtier, N. and Duret, L. (2007) 'Adaptation or biased gene conversion? Extending the null hypothesis of molecular evolution', *Trends Genet,* 23(6), 273-7.

Galtier, N., Duret, L., Glemin, S. and Ranwez, V. (2009) 'GC-biased gene conversion promotes the fixation of deleterious amino acid changes in primates', *Trends Genet,* 25(1), 1-5.

Galtier, N., Piganeau, G., Mouchiroud, D. and Duret, L. (2001) 'GC-content evolution in mammalian genomes: the biased gene conversion hypothesis', *Genetics,* 159(2), 907-11.

Gao, F., Bailes, E., Robertson, D. L., Chen, Y., Rodenburg, C. M., Michael, S. F., Cummins, L. B., Arthur, L. O., Peeters, M., Shaw, G. M., Sharp, P. M. and Hahn, B. H. (1999) 'Origin of HIV-1 in the chimpanzee Pan troglodytes troglodytes', *Nature,* 397(6718), 436-41.

Gao, F., Yue, L., Robertson, D. L., Hill, S. C., Hui, H., Biggar, R. J., Neequaye, A. E., Whelan, T. M., Ho, D. D., Shaw, G. M. and et al. (1994) 'Genetic diversity of human immunodeficiency virus type 2: evidence for distinct sequence subtypes with differences in virus biology', *J Virol,* 68(11), 7433-47.

Gao, F., Yue, L., White, A. T., Pappas, P. G., Barchue, J., Hanson, A. P., Greene, B. M., Sharp, P. M., Shaw, G. M. and Hahn, B. H. (1992) 'Human infection by genetically diverse SIVSM-related HIV-2 in west Africa', *Nature,* 358(6386), 495-9.

Garcia-Cao, M., O'Sullivan, R., Peters, A. H., Jenuwein, T. and Blasco, M. A. (2004) 'Epigenetic regulation of telomere length in mammalian cells by the Suv39h1 and Suv39h2 histone methyltransferases', *Nat Genet,* 36(1), 94-9.

Gee, H. (2003) 'Evolution: ending incongruence', *Nature,* 425(6960), 782.

Gelman, A. B., Carlin, J. S., Stern, H. S. and Rubin, D. B. (1995) *Bayesian data analysis.,* London: Chapman & Hall.

Geyer, C. J. (1992) 'Practical Markov Chain Monte Carlo', *Statistical Science,* 7(4), 473-483.

Gibbs, M. J., Armstrong, J. S. and Gibbs, A. J. (2000) 'Sister-scanning: a Monte Carlo procedure for assessing signals in recombinant sequences', *Bioinformatics,* 16(7), 573-82.

Gibson, A., Gowri-Shankar, V., Higgs, P. G. and Rattray, M. (2005) 'A comprehensive analysis of mammalian mitochondrial genome base composition and improved phylogenetic methods', *Mol Biol Evol,* 22(2), 251-64.

Gillespie, J. H. (1998) *Population genetics: a concise guide,* London: The John Hopkins University Press.

Glazier, D. S. (2005) 'Beyond the '3/4-power law': variation in the intra- and interspecific scaling of metabolic rate in animals', *Biol Rev Camb Philos Soc,* 80(4), 611-62.

Gojobori, T. (1983) 'Codon substitution in evolution and the "saturation" of synonymous changes', *Genetics,* 105(4), 1011-27.

Goldman, N. and Yang, Z. (1994) 'A codon-based model of nucleotide substitution for protein-coding DNA sequences', *Mol Biol Evol,* 11(5), 725-36.

Goloboff, P. A. and Pol, D. (2002) 'Semi-strict supertrees', *Cladistics,* 18, 514-525.

Gow, J. L., Rogers, S. M., Jackson, M. and Schluter, D. (2008) 'Ecological predictions lead to the discovery of a benthic-limnetic sympatric species pair of threespine stickleback in Little Quarry Lake, British Columbia', *Canadian Journal of Zoology-Revue Canadienne De Zoologie,* 86(6), 564-571.

Graham, J. (1983) 'Cancer and evolution: synthesis', *J Theor Biol,* 101(4), 657-9.

Graham, J. (1992) *Cancer Selection: The new theory of evolution,* Lexington: Aculeus Press Inc.

Grant, T. and Kluge, A. G. (2003) 'Data exploration in phylogenetic inference: scientific, heuristic, or neither', *Cladisticss,* 19, 379-418.

Greaves, M. and Maley, C. C. (2012) 'Clonal evolution in cancer', *Nature,* 481(7381), 306-13.

Green, R. E., Krause, J., Briggs, A. W., Maricic, T., Stenzel, U., Kircher, M., Patterson, N., Li, H., Zhai, W., Fritz, M. H., Hansen, N. F., Durand, E. Y., Malaspinas, A. S., Jensen, J. D., Marques-Bonet, T., Alkan, C., Prufer, K., Meyer, M., Burbano, H. A., Good, J. M., Schultz, R., Aximu-Petri, A., Butthof, A., Hober, B., Hoffner, B., Siegemund, M., Weihmann, A., Nusbaum, C., Lander, E. S., Russ, C., Novod, N., Affourtit, J., Egholm, M., Verna, C., Rudan, P., Brajkovic, D., Kucan, Z., Gusic, I., Doronichev, V. B., Golovanova, L. V., Lalueza-Fox, C., de la Rasilla, M., Fortea, J., Rosas, A., Schmitz, R. W., Johnson, P. L., Eichler, E. E., Falush, D., Birney, E., Mullikin, J. C., Slatkin, M., Nielsen, R., Kelso, J., Lachmann, M., Reich, D. and Paabo, S. (2010) 'A draft sequence of the Neandertal genome', *Science,* 328(5979), 710-22.

Gu, X. and Li, W. H. (1992) 'Higher rates of amino acid substitution in rodents than in humans', *Mol Phylogenet Evol,* 1(3), 211-4.

Guerzoni, D. and McLysaght, A. (2011) 'De novo origins of human genes', *PLoS Genet,* 7(11), e1002381.

Hahn, M. W. and Kern, A. D. (2005) 'Comparative genomics of centrality and essentiality in three eukaryotic protein-interaction networks', *Mol Biol Evol,* 22(4), 803-6.

Hailer, F. and Leonard, J. A. (2008) 'Hybridization among three native North American Canis species in a region of natural sympatry', *Plos One,* 3(10), e3333.

Hall, B. G. (2005) 'Comparison of the accuracies of several phylogenetic methods using protein and DNA sequences', *Mol Biol Evol,* 22(3), 792-802.

Hall, B. G. (2007) *Phylogenetic Trees Made Easy: A How-To Manual,* Massachusetts: Sinauer Associates, Inc.

Hall, B. G. and Zuzel, T. (1980) 'Evolution of a new enzymatic function by recombination within a gene', *Proc Natl Acad Sci U S A,* 77(6), 3529-33.

Hallstrom, B. M. and Janke, A. (2008) 'Resolution among major placental mammal interordinal relationships with genome data imply that speciation influenced their earliest radiations', *BMC Evol Biol,* 8, 162.

Hallstrom, B. M. and Janke, A. (2010) 'Mammalian Evolution may not be strictly bifurcating', *Molecular Biology and Evolution,* 27(12), 2804-2816.

Hallstrom, B. M., Kullberg, M., Nilsson, M. A. and Janke, A. (2007) 'Phylogenomic data analyses provide evidence that Xenarthra and Afrotheria are sister groups', *Mol Biol Evol,* 24(9), 2059-68.

Hamilton, W. D. (1964) 'The genetical evolution of social behaviour. I', *J Theor Biol,* 7(1), 1-16.

Hart, R. W. and Setlow, R. B. (1974) 'Correlation between deoxyribonucleic acid excision-repair and life-span in a number of mammalian species', *Proc Natl Acad Sci U S A,* 71(6), 2169-73.

Hasegawa, M. and Hashimoto, T. (1993) 'Ribosomal RNA trees misleading?', *Nature,* 361(6407), 23.

Hasegawa, M., Kishino, H. and Yano, T. (1985) 'Dating of the human-ape splitting by a molecular clock of mitochondrial DNA', *J Mol Evol,* 22(2), 160-74.

Hashimoto, T., Nakamura, Y., Kamaishi, T., Nakamura, F., Adachi, J., Okamoto, K. and Hasegawa, M. (1995) 'Phylogenetic place of mitochondrion-lacking protozoan, Giardia lamblia, inferred from amino acid sequences of elongation factor 2', *Mol Biol Evol,* 12(5), 782-93.

Hastings, W. K. (179) 'Monte Carlo sampling methods using Markov chains and their applications', *Biometrika,* 57(97-109).

Hawkes, K. (2003) 'Grandmothers and the evolution of human longevity', *Am J Hum Biol,* 15(3), 380-400.

Heard, E. and Disteche, C. M. (2006) 'Dosage compensation in mammals: fine-tuning the expression of the X chromosome', *Genes Dev,* 20(14), 1848-67.

Heath, T. A., Zwickl, D. J., Kim, J. and Hillis, D. M. (2008) 'Taxon sampling affects inferences of macroevolutionary processes from phylogenetic trees', *Syst Biol,* 57(1), 160-6.

Hebert, P. D., Cywinska, A., Ball, S. L. and deWaard, J. R. (2003) 'Biological identifications through DNA barcodes', *Proc Biol Sci,* 270(1512), 313-21.

Hedges, S. B. and Maxson, L. R. (1996) 'Re: Molecules and morphology in amniote phylogeny', *Mol Phylogenet Evol,* 6(2), 312-4.

Hedges, S. B., Parker, P. H., Sibley, C. G. and Kumar, S. (1996) 'Continental breakup and the ordinal diversification of birds and mammals', *Nature,* 381(6579), 226-9.

Hedtke, S. M., Townsend, T. M. and Hillis, D. M. (2006) 'Resolution of phylogenetic conflict in large data sets by increased taxon sampling', *Systematic Biology,* 55(3), 522-529.

Hendy, M. D. and Penny, D. (1989) 'A framework for the quantitative study of evolutionary trees', *Syst Zool,* 38, 279-309.

Henikoff, S. and Henikoff, J. G. (1992) 'Amino-Acid Substitution Matrices from Protein Blocks', *Proceedings of the National Academy of Sciences of the United States of America,* 89(22), 10915-10919.

Hennig, W. (1966) *Phylogenetic systematics,* Urbana: University of Illinois Press.

Henricson, A., Forslund, K. and Sonnhammer, E. L. (2010) 'Orthology confers intron position conservation', *BMC Genomics,* 11, 412.

Hillis, D. M. (1987) 'Molecular versus Morphological Approaches to Systematics', *Annual Review of Ecology and Systematics,* 18, 13-42.

Hillis, D. M. (1996) 'Inferring complex phylogenies', *Nature,* 383(6596), 130-1.

Hillis, D. M. (1998) 'Taxonomic sampling, phylogenetic accuracy, and investigator bias', *Syst Biol,* 47(1), 3-8.

Hillis, D. M., Pollock, D. D., McGuire, J. A. and Zwickl, D. J. (2003) 'Is sparse taxon sampling a problem for phylogenetic inference?', *Syst Biol,* 52(1), 124-6.

Hillis, D. M. and Wilcox, T. P. (2005) 'Phylogeny of the New World true frogs (Rana)', *Molecular Phylogenetics and Evolution,* 34(2), 299-314.

Hindorff, L. A., Sethupathy, P., Junkins, H. A., Ramos, E. M., Mehta, J. P., Collins, F. S. and Manolio, T. A. (2009) 'Potential etiologic and functional implications of genome-wide association loci for human diseases and traits', *Proc Natl Acad Sci U S A,* 106(23), 9362-7.

Hoarau, G., Holla, S., Lescasse, R., Stam, W. T. and Olsen, J. L. (2002) 'Heteroplasmy and evidence for recombination in the mitochondrial control region of the flatfish Platichthys flesus', *Mol Biol Evol,* 19(12), 2261-4.

Hou, Z. C., Romero, R. and Wildman, D. E. (2009) 'Phylogeny of the Ferungulata (Mammalia: Laurasiatheria) as determined from phylogenomic data', *Mol Phylogenet Evol,* 52(3), 660-4.

Huang, R., Hippauf, F., Rohrbeck, D., Haustein, M., Wenke, K., Feike, J., Sorrelle, N., Piechulla, B. and Barkman, T. J. (2012) 'Enzyme functional evolution through improved catalysis of ancestrally nonpreferred substrates', *Proc Natl Acad Sci U S A,* 109(8), 2966-71.

Hubbard, T. J., Aken, B. L., Beal, K., Ballester, B., Caccamo, M., Chen, Y., Clarke, L., Coates, G., Cunningham, F., Cutts, T., Down, T., Dyer, S. C., Fitzgerald, S., Fernandez-Banet, J., Graf, S., Haider, S., Hammond, M., Herrero, J., Holland, R., Howe, K., Johnson, N., Kahari, A., Keefe, D., Kokocinski, F., Kulesha, E., Lawson, D., Longden, I., Melsopp, C., Megy, K., Meidl, P., Ouverdin, B., Parker, A., Prlic, A., Rice, S., Rios, D., Schuster, M., Sealy, I., Severin, J., Slater, G., Smedley, D., Spudich, G., Trevanion, S., Vilella, A., Vogel, J., White, S., Wood, M., Cox, T., Curwen, V., Durbin, R., Fernandez-Suarez, X. M., Flicek, P., Kasprzyk, A., Proctor, G., Searle, S., Smith, J., Ureta-Vidal, A. and Birney, E. (2007) 'Ensembl 2007', *Nucleic Acids Research,* 35(Database issue), D610-7.

Huber, P. J. (1964) 'Robust Estimation of a Location Parameter', *Annals of Mathematical Statistics,* 35, 73-101.

Hudelot, C., Gowri-Shankar, V., Jow, H., Rattray, M. and Higgs, P. G. (2003) 'RNA-based phylogenetic methods: application to mammalian mitochondrial RNA sequences', *Mol Phylogenet Evol,* 28(2), 241-52.

Hudson, C. M. and Conant, G. C. (2011) 'Expression level, cellular compartment and metabolic network position all influence the average selective constraint on mammalian enzymes', *BMC Evol Biol,* 11, 89.

Huelsenbeck, J. and Rannala, B. (2004) 'Frequentist properties of Bayesian posterior probabilities of phylogenetic trees under simple and complex substitution models', *Syst Biol,* 53(6), 904-13.

Huelsenbeck, J. P. (1995) 'The robustness of two phylogenetic methods: four-taxon simulations reveal a slight superiority of maximum likelihood over neighbor joining', *Mol Biol Evol,* 12(5), 843-9.

Huelsenbeck, J. P. and Hillis, D. M. (1993) 'Success of Phylogenetic Methods in the Four-Taxon Case', *Systematic Biology,* 42(3), 247-264.

Huelsenbeck, J. P., Larget, B., Miller, R. E. and Ronquist, F. (2002) 'Potential applications and pitfalls of Bayesian inference of phylogeny', *Syst Biol,* 51(5), 673-88.

Huelsenbeck, J. P. and Ronquist, F. (2001) 'MRBAYES: Bayesian inference of phylogenetic trees', *Bioinformatics,* 17(8), 754-5.

Huelsenbeck, J. P., Ronquist, F., Nielsen, R. and Bollback, J. P. (2001) 'Bayesian inference of phylogeny and its impact on evolutionary biology', *Science,* 294(5550), 2310-4.

Hughes, A. L. (1999) *Adaptive Evolution of Genes and Genomes,* New York: Oxford University Press.

Hughes, A. L. (2007) 'Looking for Darwin in all the wrong places: the misguided quest for positive selection at the nucleotide sequence level', *Heredity (Edinb),* 99(4), 364-73.

Hughes, A. L. and Friedman, R. (2004) 'Recent mammalian gene duplications: robust search for functionally divergent gene pairs', *J Mol Evol,* 59(1), 114-20.

Hughes, A. L. and Nei, M. (1989) 'Evolution of the major histocompatibility complex: independent origin of nonclassical class I genes in different groups of mammals', *Mol Biol Evol,* 6(6), 559-79.

Hughes, A. L. and Piontkivska, H. (2008) 'Nucleotide sequence polymorphism in circoviruses', *Infect Genet Evol,* 8(2), 130-8.

Hyman, I. T., Ho, S. Y. and Jermiin, L. S. (2007) 'Molecular phylogeny of Australian Helicarionidae, Euconulidae and related groups (Gastropoda: Pulmonata: Stylommatophora) based on mitochondrial DNA', *Molecular Phylogenetics and Evolution,* 45(3), 792-812.

ICUN (2012) 'IUCN Red List of Threatened Species. Version 2012.2. <www.iucnredlist.org>', [online], available: [accessed

Ina, Y. (1995) 'New methods for estimating the numbers of synonymous and nonsynonymous substitutions', *J Mol Evol,* 40(2), 190-226.

Irwin, D. M., Kocher, T. D. and Wilson, A. C. (1991) 'Evolution of the cytochrome b gene of mammals', *J Mol Evol,* 32(2), 128-44.

Janecka, J. E., Miller, W., Pringle, T. H., Wiens, F., Zitzmann, A., Helgen, K. M., Springer, M. S. and Murphy, W. J. (2007) 'Molecular and genomic data identify the closest living relative of primates', *Science,* 318(5851), 792-4.

Javaud, C., Dupuy, F., Maftah, A., Julien, R. and Petit, J. M. (2003) 'The fucosyltransferase gene family: an amazing summary of the underlying mechanisms of gene evolution', *Genetica,* 118(2-3), 157-70.

Jenner, R. A. (2004) 'When molecules and morphology clash: reconciling conflicting phylogenies of the Metazoa by considering secondary character loss', *Evol Dev,* 6(5), 372-8.

Jennings, W. B. and Edwards, S. V. (2005) 'Speciational history of Australian grass finches (Poephila) inferred from thirty gene trees', *Evolution,* 59(9), 2033-47.

Jensen, L. J., Kuhn, M., Stark, M., Chaffron, S., Creevey, C., Muller, J., Doerks, T., Julien, P., Roth, A., Simonovic, M., Bork, P. and von Mering, C. (2009) 'STRING 8--a global view on proteins and their functional interactions in 630 organisms', *Nucleic Acids Res,* 37(Database issue), D412-6.

Jensen, R. A. (2001) 'Orthologs and paralogs - we need to get it right', *Genome Biol,* 2(8), INTERACTIONS1002.

Jones, D. T., Taylor, W. R. and Thornton, J. M. (1992) 'The rapid generation of mutation data matrices from protein sequences', *Comput Appl Biosci,* 8(3), 275-82.

Jukes, T. C., C. (1969) 'Evolution of protein molecules.', *In Mammalian protein metabolism (ed. H. Munro),*, 111.

Junier, T. and Zdobnov, E. M. (2010) 'The Newick utilities: high-throughput phylogenetic tree processing in the UNIX shell', *Bioinformatics,* 26(13), 1669-70.

Kass, R. E. and Raftery, A. E. (1995) 'Bayes Factors', *Journal of the American Statistical Association,* 90(430), 773-795.

Katoh, K. and Toh, H. (2008) 'Recent developments in the MAFFT multiple sequence alignment program', *Briefings in Bioinformatics,* 9(4), 286-98.

Katzman, S., Capra, J. A., Haussler, D. and Pollard, K. S. (2011) 'Ongoing GC-biased evolution is widespread in the human genome and enriched near recombination hot spots', *Genome Biol Evol,* 3, 614-26.

Keane, T. M., Creevey, C. J., Pentony, M. M., Naughton, T. J. and 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 Evol Biol,* 6, 29.

Kearney, M. (2002) 'Fragmentary taxa, missing data, and ambiguity: mistaken assumptions and conclusions', *Syst Biol,* 51(2), 369-81.

Kim, E. B., Fang, X., Fushan, A. A., Huang, Z., Lobanov, A. V., Han, L., Marino, S. M., Sun, X., Turanov, A. A., Yang, P., Yim, S. H., Zhao, X., Kasaikina, M. V., Stoletzki, N., Peng, C., Polak, P., Xiong, Z., Kiezun, A., Zhu, Y., Chen, Y., Kryukov, G. V., Zhang, Q., Peshkin, L., Yang, L., Bronson, R. T., Buffenstein, R., Wang, B., Han, C., Li, Q., Chen, L., Zhao, W., Sunyaev, S. R., Park, T. J., Zhang, G., Wang, J. and Gladyshev, V. N. (2011) 'Genome sequencing reveals insights into physiology and longevity of the naked mole rat', *Nature,* 479(7372), 223-7.

Kimura, M. (1957) 'Some problems of stochastic processes in genetics', *Annals of Mathematical Statistics,* 28, 882-901.

Kimura, M. (1968) 'Evolutionary rate at the molecular level', *Nature,* 217(5129), 624-6.

Kimura, M. (1979) 'The neutral theory of molecular evolution', *Sci Am,* 241(5), 98-100, 102, 108 passim.

Kimura, M. (1980a) 'Average time until fixation of a mutant allele in a finite population under continued mutation pressure: Studies by analytical, numerical, and pseudo-sampling methods', *Proc Natl Acad Sci U S A,* 77(1), 522-6.

Kimura, M. (1980b) 'A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences', *J Mol Evol,* 16(2), 111-20.

Kimura, M. (1981) 'Estimation of evolutionary distances between homologous nucleotide sequences', *Proc Natl Acad Sci U S A,* 78(1), 454-8.

Kjer, K. M. and Honeycutt, R. L. (2007) 'Site specific rates of mitochondrial genomes and the phylogeny of eutheria', *BMC Evol Biol,* 7, 8.

Kluge, A. G. (1989) 'A concern for evidence and a phylogenetic hypothesis of relationships among Epicrates (Boidae, Serpentes)', *Syst Zool,* 38, 7-25.

Kolaczkowski, B. and Thornton, J. W. (2004) 'Performance of maximum parsimony and likelihood phylogenetics when evolution is heterogeneous', *Nature,* 431(7011), 980-4.

Kondrashov, F. A. and Koonin, E. V. (2004) 'A common framework for understanding the origin of genetic dominance and evolutionary fates of gene duplications', *Trends Genet,* 20(7), 287-90.

Koonin, E. V. (2001) 'An apology for orthologs - or brave new memes', *Genome Biol,* 2(4), COMMENT1005.

Korbel, J. O., Snel, B., Huynen, M. A. and Bork, P. (2002) 'SHOT: a web server for the construction of genome phylogenies', *Trends Genet,* 18(3), 158-62.

Kornegay, J. R., Schilling, J. W. and Wilson, A. C. (1994) 'Molecular adaptation of a leaf-eating bird: stomach lysozyme of the hoatzin', *Mol Biol Evol,* 11(6), 921-8.

Kosiol, C., Holmes, I. and Goldman, N. (2007) 'An empirical codon model for protein sequence evolution', *Mol Biol Evol,* 24(7), 1464-79.

Kosiol, C., Vinar, T., da Fonseca, R. R., Hubisz, M. J., Bustamante, C. D., Nielsen, R. and Siepel, A. (2008) 'Patterns of positive selection in six Mammalian genomes', *PLoS Genet,* 4(8), e1000144.

Kriegs, J. O., Churakov, G., Kiefmann, M., Jordan, U., Brosius, J. and Schmitz, J. (2006) 'Retroposed elements as archives for the evolutionary history of placental mammals', *Plos Biology,* 4(4), e91.

Kuhner, M. K. and Felsenstein, J. (1994) 'A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates', *Mol Biol Evol,* 11(3), 459-68.

Kumar, S. (1996) 'A stepwise algorithm for finding minimum evolution trees', *Mol Biol Evol,* 13(4), 584-93.

Kumar, S., Dudley, J. T., Filipski, A. and Liu, L. (2011) 'Phylomedicine: an evolutionary telescope to explore and diagnose the universe of disease mutations', *Trends Genet,* 27(9), 377-86.

Ladoukakis, E. D. and Zouros, E. (2001) 'Recombination in animal mitochondrial DNA: evidence from published sequences', *Mol Biol Evol,* 18(11), 2127-31.

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., McKernan, K., Meldrim, J., Mesirov, J. P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, N., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., McMurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J. C., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R. H., Wilson, R. K., Hillier, L. W., McPherson, J. D., Marra, M. A., Mardis, E. R., Fulton, L. A., Chinwalla, A. T., Pepin, K. H., Gish, W. R., Chissoe, S. L., Wendl, M. C., Delehaunty, K. D., Miner, T. L., Delehaunty, A., Kramer, J. B., Cook, L. L., Fulton, R. S., Johnson, D. L., Minx, P. J., Clifton, S. W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J. F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., et al. (2001) 'Initial sequencing and analysis of the human genome', *Nature,* 409(6822), 860-921.

Lartillot, N. (2012) 'Phylogenetic patterns of GC-biased gene conversion in placental mammals, and the evolutionary dynamics of recombination landscapes', *Mol Biol Evol*.

Lartillot, N., Brinkmann, H. and Philippe, H. (2007) 'Suppression of long-branch attraction artefacts in the animal phylogeny using a site-heterogeneous model', *BMC Evol Biol,* 7 Suppl 1, S4.

Lartillot, N., Lepage, T. and Blanquart, S. (2009) 'PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating', *Bioinformatics,* 25(17), 2286-2288.

Lartillot, N. and Philippe, H. (2004) 'A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process', *Mol Biol Evol,* 21(6), 1095-109.

Lartillot, N. and Poujol, R. (2011) 'A phylogenetic model for investigating correlated evolution of substitution rates and continuous phenotypic characters', *Mol Biol Evol,* 28(1), 729-44.

Le, S. Q. and Gascuel, O. (2008) 'An improved general amino acid replacement matrix', *Mol Biol Evol,* 25(7), 1307-20.

Leache, A. D. (2010) 'Species trees for spiny lizards (genus Sceloporus): identifying points of concordance and conflict between nuclear and mitochondrial data', *Molecular Phylogenetics and Evolution,* 54(1), 162-71.

Lefevre, C. M., Sharp, J. A. and Nicholas, K. R. (2010) 'Evolution of lactation: ancient origin and extreme adaptations of the lactation system', *Annu Rev Genomics Hum Genet,* 11, 219-38.

Lemmon, A. R., Brown, J. M., Stanger-Hall, K. and Lemmon, E. M. (2009) 'The effect of ambiguous data on phylogenetic estimates obtained by maximum likelihood and Bayesian inference', *Syst Biol,* 58(1), 130-45.

Leroi, A. M., Koufopanou, V. and Burt, A. (2003) 'Cancer selection', *Nat Rev Cancer,* 3(3), 226-31.

Levasseur, A., Gouret, P., Lesage-Meessen, L., Asther, M., Record, E. and Pontarotti, P. (2006) 'Tracking the connection between evolutionary and functional shifts using the fungal lipase/feruloyl esterase A family', *BMC Evol Biol,* 6, 92.

Levine, M. T., Jones, C. D., Kern, A. D., Lindfors, H. A. and Begun, D. J. (2006) 'Novel genes derived from noncoding DNA in Drosophila melanogaster are frequently X-linked and exhibit testis-biased expression', *Proc Natl Acad Sci U S A,* 103(26), 9935-9.

Li, B., Zhang, G., Willersleve, E., Wang, J. and Wang, J. (2011) 'Genomic data from the Polar Bear (Ursus maritimus)', [online], available: http://dx.doi.org/10.5524/100008 [accessed

Li, F. W., Kuo, L. Y., Rothfels, C. J., Ebihara, A., Chiou, W. L., Windham, M. D. and Pryer, K. M. (2011) 'rbcL and matK earn two thumbs up as the core DNA barcode for ferns', *PLoS One,* 6(10), e26597.

Li, K. S., Guan, Y., Wang, J., Smith, G. J., Xu, K. M., Duan, L., Rahardjo, A. P., Puthavathana, P., Buranathai, C., Nguyen, T. D., Estoepangestie, A. T., Chaisingh, A., Auewarakul, P., Long, H. T., Hanh, N. T., Webby, R. J., Poon, L. L., Chen, H., Shortridge, K. F., Yuen, K. Y., Webster, R. G. and Peiris, J. S. (2004) 'Genesis of a highly pathogenic and potentially pandemic H5N1 influenza virus in eastern Asia', *Nature,* 430(6996), 209-13.

Li, L., Stoeckert, C. J., Jr. and Roos, D. S. (2003) 'OrthoMCL: identification of ortholog groups for eukaryotic genomes', *Genome Research,* 13(9), 2178-89.

Li, S. (1996) *Phylogenetic tree construction using Markov chain Monte Carlo*, unpublished thesis Ohio State Univ.

Li, W. H. (1993) 'Unbiased estimation of the rates of synonymous and nonsynonymous substitution', *J Mol Evol,* 36(1), 96-9.

Li, W. H., Ellsworth, D. L., Krushkal, J., Chang, B. H. and Hewett-Emmett, D. (1996) 'Rates of nucleotide substitution in primates and rodents and the generation-time effect hypothesis', *Mol Phylogenet Evol,* 5(1), 182-7.

Li, W. H., Tanimura, M. and Sharp, P. M. (1987) 'An evaluation of the molecular clock hypothesis using mammalian DNA sequences', *J Mol Evol,* 25(4), 330-42.

Li, W. H., Wu, C. I. and Luo, C. C. (1985) 'A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes', *Mol Biol Evol,* 2(2), 150-74.

Li, Y. and Agarwal, P. (2009) 'A pathway-based view of human diseases and disease relationships', *PLoS One,* 4(2), e4346.

Lin, J. and Gerstein, M. (2000) 'Whole-genome trees based on the occurrence of folds and orthologs: implications for comparing genomes on different levels', *Genome Res,* 10(6), 808-18.

Lio, P. and Goldman, N. (1999) 'Using protein structural information in evolutionary inference: transmembrane proteins', *Mol Biol Evol,* 16(12), 1696-710.

Lloyd, G. T., Davis, K. E., Pisani, D., Tarver, J. E., Ruta, M., Sakamoto, M., Hone, D. W., Jennings, R. and Benton, M. J. (2008) 'Dinosaurs and the Cretaceous Terrestrial Revolution', *Proc Biol Sci,* 275(1650), 2483-90.

Lockhart, P. J., Howe, C. J., Bryant, D. A., Beanland, T. J. and Larkum, A. W. (1992a) 'Substitutional bias confounds inference of cyanelle origins from sequence data', *J Mol Evol,* 34(2), 153-62.

Lockhart, P. J., Penny, D., Hendy, M. D., Howe, C. J., Beanland, T. J. and Larkum, A. W. (1992b) 'Controversy on chloroplast origins', *FEBS Lett,* 301(2), 127-31.

Long, M. and Langley, C. H. (1993) 'Natural selection and the origin of jingwei, a chimeric processed functional gene in Drosophila', *Science,* 260(5104), 91-5.

Loomis, W. F. and Smith, D. W. (1990) 'Molecular phylogeny of Dictyostelium discoideum by protein sequence comparison', *Proc Natl Acad Sci U S A,* 87(23), 9093-7.

Lopez, P., Casane, D. and Philippe, H. (2002) 'Heterotachy, an important process of protein evolution', *Mol Biol Evol,* 19(1), 1-7.

Loughran, N. B., Hinde, S., McCormick-Hill, S., Leidal, K. G., Bloomberg, S., Loughran, S. T., O'Connor, B., O'Fagain, C., Nauseef, W. M. and O'Connell, M. J. (2012) 'Functional consequence of positive selection revealed through rational mutagenesis of human myeloperoxidase', *Mol Biol Evol,* 29(8), 2039-46.

Loughran, N. B., O'Connor, B., O'Fagain, C. and O'Connell, M. J. (2008) 'The phylogeny of the mammalian heme peroxidases and the evolution of their diverse functions', *BMC Evol Biol,* 8, 101.

Loytynoja, A. and Goldman, N. (2008) 'Phylogeny-aware gap placement prevents errors in sequence alignment and evolutionary analysis', *Science,* 320(5883), 1632-5.

Lunt, D. H. and Hyman, B. C. (1997) 'Animal mitochondrial DNA recombination', *Nature,* 387(6630), 247.

Lynch, M. (2007) 'The evolution of genetic networks by non-adaptive processes', *Nat Rev Genet,* 8(10), 803-13.

Lynch, M. and Conery, J. S. (2000) 'The evolutionary fate and consequences of duplicate genes', *Science,* 290(5494), 1151-5.

Lynch, M. and Force, A. (2000) 'The probability of duplicate gene preservation by subfunctionalization', *Genetics,* 154(1), 459-73.

Lynch, M. & Katju, V. (2004) 'The altered evolutionary trajectories of gene duplicates', *Trends Genet,* 20**,** 544-9.

Lynch, M., Hill W.G. (1986) 'Phenotypic Evolution by Neutral Mutation', *Evolution,* 40(5), 915-935.

Lyson, T. R., Sperling, E. A., Heimberg, A. M., Gauthier, J. A., King, B. L. and Peterson, K. J. (2012) 'MicroRNAs support a turtle + lizard clade', *Biol Lett,* 8(1), 104-7.

Madsen, O., Scally, M., Douady, C. J., Kao, D. J., DeBry, R. W., Adkins, R., Amrine, H. M., Stanhope, M. J., de Jong, W. W. and Springer, M. S. (2001) 'Parallel adaptive radiations in two major clades of placental mammals', *Nature,* 409(6820), 610-4.

Malay, M. C. and Paulay, G. (2010) 'Peripatric speciation drives diversification and distributional pattern of reef hermit crabs (Decapoda: Diogenidae: Calcinus)', *Evolution,* 64(3), 634-62.

Mallet, J., Meyer, A., Nosil, P. and Feder, J. L. (2009) 'Space, sympatry and speciation', *J Evol Biol,* 22(11), 2332-41.

Marais, G., Mouchiroud, D. and Duret, L. (2001) 'Does recombination improve selection on codon usage? Lessons from nematode and fly complete genomes', *Proc Natl Acad Sci U S A,* 98(10), 5688-92.

Markova-Raina, P. and Petrov, D. (2011) 'High sensitivity to aligner and high rate of false positives in the estimates of positive selection in the 12 Drosophila genomes', *Genome Res,* 21(6), 863-74.

Martin, A. P. and Palumbi, S. R. (1993) 'Body size, metabolic rate, generation time, and the molecular clock', *Proc Natl Acad Sci U S A,* 90(9), 4087-91.

Martin, D. and Rybicki, E. (2000) 'RDP: detection of recombination amongst aligned sequences', *Bioinformatics,* 16(6), 562-3.

Martin, D. P., Lemey, P., Lott, M., Moulton, V., Posada, D. and Lefeuvre, P. (2010) 'RDP3: a flexible and fast computer program for analyzing recombination', *Bioinformatics,* 26(19), 2462-3.

Martin, D. P., Posada, D., Crandall, K. A. and Williamson, C. (2005) 'A modified bootscan algorithm for automated identification of recombinant sequences and recombination breakpoints', *AIDS Res Hum Retroviruses,* 21(1), 98-102.

Massingham, T. and Goldman, N. (2005) 'Detecting amino acid sites under positive selection and purifying selection', *Genetics,* 169(3), 1753-62.

Matsen, F. A. and Steel, M. (2007) 'Phylogenetic mixtures on a single tree can mimic a tree of another topology', *Systematic Biology,* 56(5), 767-775.

Mau, B. (1996) *Bayesian phylogenetic inference via Markov chain Monte Carlo methods*, unpublished thesis Univ. Wisconsin.

Mayr, E. (1942) *Systematics and the origin of species.,* New York: Columbia University Press.

Mayr, E. (1970) *Populations, Species and Evolution,* Cambridge: The Belknap Press of Harvard University Press,.

Mayrose, I., Friedman, N. and Pupko, T. (2005) 'A gamma mixture model better accounts for among site rate heterogeneity', *Bioinformatics,* 21 Suppl 2, ii151-8.

McDonald, J. H. and Kreitman, M. (1991) 'Adaptive protein evolution at the Adh locus in Drosophila', *Nature,* 351(6328), 652-4.

McInerney, J. O., Pisani, D., Bapteste, E. and O'Connell, M. J. (2011) 'The public goods hypothesis for the evolution of life on Earth', *Biol Direct,* 6(1), 41.

Meredith, R. W., Janecka, J. E., Gatesy, J., Ryder, O. A., Fisher, C. A., Teeling, E. C., Goodbla, A., Eizirik, E., Simao, T. L., Stadler, T., Rabosky, D. L., Honeycutt, R. L., Flynn, J. J., Ingram, C. M., Steiner, C., Williams, T. L., Robinson, T. J., Burk-Herrick, A., Westerman, M., Ayoub, N. A., Springer, M. S. and Murphy, W. J. (2011) 'Impacts of the Cretaceous Terrestrial Revolution and KPg extinction on mammal diversification', *Science,* 334(6055), 521-4.

Messier, W. and Stewart, C. B. (1997) 'Episodic adaptive evolution of primate lysozymes', *Nature,* 385(6612), 151-4.

Metropolis, N., Rosenbluth, M. N., Rosenbluth, A. H., Teller, A. H. and Teller, E. (1953) 'Equations of state calculations by fast computing machines', *J. Chem. Phys.,* 21, 1087-1091

.

Milinkovitch, M. C., Orti, G. and Meyer, A. (1993) 'Revised phylogeny of whales suggested by mitochondrial ribosomal DNA sequences', *Nature,* 361(6410), 346-8.

Montgelard, C., Forty, E., Arnal, V. and Matthee, C. A. (2008) 'Suprafamilial relationships among Rodentia and the phylogenetic effect of removing fast-evolving nucleotides in mitochondrial, exon and intron fragments', *BMC Evol Biol,* 8, 321.

Morgan, C. C., Loughran, N. B., Walsh, T. A., Harrison, A. J. and O'Connell, M. J. (2010) 'Positive selection neighboring functionally essential sites and disease-implicated regions of mammalian reproductive proteins', *BMC Evol Biol,* 10, 39.

Morgan, C. C., Shakya, K., Webb, A., Walsh, T. A., Lynch, M., Loscher, C. E., Ruskin, H. J. and O'Connell, M. J. (2012) 'Colon cancer associated genes exhibit signatures of positive selection at functionally significant positions', *BMC Evol Biol,* 12(1), 114.

Morrison, D. A. and Ellis, J. T. (1997) 'Effects of nucleotide sequence alignment on phylogeny estimation: a case study of 18S rDNAs of apicomplexa', *Mol Biol Evol,* 14(4), 428-41.

Moury, B. and Simon, V. (2011) 'dN/dS-based methods detect positive selection linked to trade-offs between different fitness traits in the coat protein of potato virus Y', *Mol Biol Evol,* 28(9), 2707-17.

Muller, H. J. (1963) 'The need for recombination to prevent genetic deterioration', *Genetics,* 48, 903-903.

Muller, J., Creevey, C. J., Thompson, J. D., Arendt, D. and Bork, P. (2010) 'AQUA: automated quality improvement for multiple sequence alignments', *Bioinformatics,* 26(2), 263-5.

Muller, J., Oma, Y., Vallar, L., Friederich, E., Poch, O. and Winsor, B. (2005) 'Sequence and comparative genomic analysis of actin-related proteins', *Mol Biol Cell,* 16(12), 5736-48.

Muller, T. and Vingron, M. (2000) 'Modeling amino acid replacement', *J Comput Biol,* 7(6), 761-76.

Murphy, W. J., Eizirik, E., Johnson, W. E., Zhang, Y. P., Ryder, O. A. and O'Brien, S. J. (2001a) 'Molecular phylogenetics and the origins of placental mammals', *Nature,* 409(6820), 614-618.

Murphy, W. J., Eizirik, E., O'Brien, S. J., Madsen, O., Scally, M., Douady, C. J., Teeling, E., Ryder, O. A., Stanhope, M. J., de Jong, W. W. and Springer, M. S. (2001b) 'Resolution of the early placental mammal radiation using Bayesian phylogenetics', *Science,* 294(5550), 2348-51.

Murphy, W. J., Pringle, T. H., Crider, T. A., Springer, M. S. and Miller, W. (2007) 'Using genomic data to unravel the root of the placental mammal phylogeny', *Genome Research,* 17(4), 413-21.

Nehrt, N. L., Clark, W. T., Radivojac, P. and Hahn, M. W. (2011) 'Testing the ortholog conjecture with comparative functional genomic data from mammals', *PLoS Comput Biol,* 7(6), e1002073.

Nei, M. and Gojobori, T. (1986) 'Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions', *Mol Biol Evol,* 3(5), 418-26.

Nei, M., Rogozin, I. B. and Piontkivska, H. (2000) 'Purifying selection and birth-and-death evolution in the ubiquitin gene family', *Proc Natl Acad Sci U S A,* 97(20), 10866-71.

Nekrutenko, A. and Li, W. H. (2001) 'Transposable elements are found in a large number of human protein-coding genes', *Trends Genet,* 17(11), 619-21.

Nesnidal, M. P., Helmkampf, M., Bruchhaus, I. and Hausdorf, B. (2010) 'Compositional heterogeneity and phylogenomic inference of metazoan relationships', *Mol Biol Evol,* 27(9), 2095-104.

Nesse, R. M., Stearns, S. C. and Omenn, G. S. (2006) 'Medicine needs evolution', *Science,* 311(5764), 1071.

Newton, M. A., Raftery, A. E., Davison, A. C., Bacha, M., Celeux, G., Carlin, B. P., Clifford, P., Lu, C., Sherman, M., Tanner, M. A., Gelfand, A. E., Mallick, B. K., Gelman, A., Grieve, A. P., Kunsch, H. R., Leonard, T., Hsu, J. S. J., Liu, J. S., Rubin, D. B., Lo, A. Y., Louis, T. A., Neal, R. M., Owen, A. B., Tu, D. S., Gilks, W. R., Roberts, G., Sweeting, T., Bates, D., Ritter, G., Worton, B. J., Barnard, G. A., Gibbens, R. and Silverman, B. (1994) 'Approximate Bayesian-Inference with the Weighted Likelihood Bootstrap', *Journal of the Royal Statistical Society Series B-Methodological,* 56(1), 3-48.

Nielsen, R. and Yang, Z. (1998) 'Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene', *Genetics,* 148(3), 929-36.

Nielsen, R. G., Husby, S. and Kruse-Andersen, S. (2005) 'Premature closure of the upper esophageal sphincter as a cause of severe deglutition disorder in infancy', *J Pediatr Surg,* 40(4), 721-4.

Niemiller, M. L., Fitzpatrick, B. M. and Miller, B. T. (2008) 'Recent divergence with gene flow in Tennessee cave salamanders (Plethodontidae: Gyrinophilus) inferred from gene genealogies', *Mol Ecol,* 17(9), 2258-75.

Nikolaev, S., Montoya-Burgos, J. I., Margulies, E. H., Rougemont, J., Nyffeler, B. and Antonarakis, S. E. (2007) 'Early history of mammals is elucidated with the ENCODE multiple species sequencing data', *Plos Genetics,* 3(1), e2.

Nishihara, H., Hasegawa, M. and Okada, N. (2006) 'Pegasoferae, an unexpected mammalian clade revealed by tracking ancient retroposon insertions', *Proc Natl Acad Sci U S A,* 103(26), 9929-34.

Nishihara, H., Okada, N. and Hasegawa, M. (2007) 'Rooting the eutherian tree: the power and pitfalls of phylogenomics', *Genome Biol,* 8(9), R199.

Nishihara, H., Satta, Y., Nikaido, M., Thewissen, J. G., Stanhope, M. J. and Okada, N. (2005) 'A retroposon analysis of Afrotherian phylogeny', *Mol Biol Evol,* 22(9), 1823-33.

Noor, M. A., Johnson, N. A. and Hey, J. (2000) 'Gene flow between Drosophila pseudoobscura and D. persimilis', *Evolution,* 54(6), 2174-5; discussion 2176-7.

Novacek, M. J. (1992) 'Mammalian phylogeny: shaking the tree', *Nature,* 356(6365), 121-5.

Nylander, J. A., Ronquist, F., Huelsenbeck, J. P. and Nieves-Aldrey, J. L. (2004) 'Bayesian phylogenetic analysis of combined data', *Syst Biol,* 53(1), 47-67.

O'Connell, J. F., Hawkes, K. and Blurton Jones, N. G. (1999) 'Grandmothering and the evolution of homo erectus', *J Hum Evol,* 36(5), 461-85.

O'Connell, M. J. (2010) 'Selection and the cell cycle: positive Darwinian selection in a well-known DNA damage response pathway', *J Mol Evol,* 71(5-6), 444-57.

Ochman, H. (2001) 'Lateral and oblique gene transfer', *Curr Opin Genet Dev,* 11(6), 616-9.

Ohno, S. (1970) *Evolution by Gene Duplication,* Heidelberg: Springer-Verlag.

Ohta, T. (1973) 'Slightly deleterious mutant substitutions in evolution', *Nature,* 246(5428), 96-8.

Ohta, T. and Gillespie, J. H. (1996) 'Development of Neutral and Nearly Neutral Theories', *Theor Popul Biol,* 49(2), 128-42.

Orsi, R. H., Ripoll, D. R., Yeung, M., Nightingale, K. K. and Wiedmann, M. (2007) 'Recombination and positive selection contribute to evolution of Listeria monocytogenes inlA', *Microbiology,* 153(Pt 8), 2666-78.

Ozdil, F. and Ilhan, F. (2012) 'Phylogenetic relationship of Turkish Apis mellifera subspecies based on sequencing of mitochondrial cytochrome C oxidase I region', *Genet Mol Res,* 11(2), 1130-41.

Padidam, M., Sawyer, S. and Fauquet, C. M. (1999) 'Possible emergence of new geminiviruses by frequent recombination', *Virology,* 265(2), 218-25.

Pagel, M. (1999) 'Inferring the historical patterns of biological evolution', *Nature,* 401(6756), 877-84.

Pagel, M. and Meade, A. (2004) 'A phylogenetic mixture model for detecting pattern-heterogeneity in gene sequence or character-state data', *Syst Biol,* 53(4), 571-81.

Patterson, N., Richter, D. J., Gnerre, S., Lander, E. S. and Reich, D. (2006) 'Genetic evidence for complex speciation of humans and chimpanzees', *Nature,* 441(7097), 1103-8.

Patthy, L. (1996) 'Exon shuffling and other ways of module exchange', *Matrix Biol,* 15(5), 301-10; discussion 311-2.

Peden, J. F. (1999) *Analysis of Codon Usage*, unpublished thesis University of Nottingham f.

Penny, D., Hendy, M., Zimmer, E. and Hamby, R. (1990) 'Trees from sequences: panacea or pandora's box?', *Australian Systematic Botany,* 3(1), 21-38.

Pereira, S. L. (2000) 'Mitochondrial genome organization and vertebrate phylogenetics', *Genetics and Moleculat Biology,* 23(4), 754-752.

Perelman, P., Johnson, W. E., Roos, C., Seuanez, H. N., Horvath, J. E., Moreira, M. A., Kessing, B., Pontius, J., Roelke, M., Rumpler, Y., Schneider, M. P., Silva, A., O'Brien, S. J. and Pecon-Slattery, J. (2011) 'A molecular phylogeny of living primates', *Plos Genetics,* 7(3), e1001342.

Pessia, E., Popa, A., Mousset, S., Rezvoy, C., Duret, L. and Marais, G. A. (2012) 'Evidence for widespread GC-biased gene conversion in eukaryotes', *Genome Biol Evol,* 4(7), 675-82.

Peterson, M. E., Chen, F., Saven, J. G., Roos, D. S., Babbitt, P. C. and Sali, A. (2009) 'Evolutionary constraints on structural similarity in orthologs and paralogs', *Protein Sci,* 18(6), 1306-15.

Peto, R., Roe, F. J., Lee, P. N., Levy, L. and Clack, J. (1975) 'Cancer and ageing in mice and men', *Br J Cancer,* 32(4), 411-26.

Petsko, G. A. (2001) 'Homologuephobia', *Genome Biol,* 2(2), COMMENT1002.

Philippe, H. (2000) 'Opinion: long branch attraction and protist phylogeny', *Protist,* 151(4), 307-16.

Philippe, H., Brinkmann, H., Lavrov, D. V., Littlewood, D. T., Manuel, M., Worheide, G. and Baurain, D. (2011) 'Resolving difficult phylogenetic questions: why more sequences are not enough', *PLoS Biol,* 9(3), e1000602.

Philippe, H., Delsuc, F., Brinkmann, H. and Lartillot, N. (2005a) 'Phylogenomics', *Annu Rev Ecol Evol Syst*, 541-562.

Philippe, H., Derelle, R., Lopez, P., Pick, K., Borchiellini, C., Boury-Esnault, N., Vacelet, J., Renard, E., Houliston, E., Queinnec, E., Da Silva, C., Wincker, P., Le Guyader, H., Leys, S., Jackson, D. J., Schreiber, F., Erpenbeck, D., Morgenstern, B., Worheide, G. and Manuel, M. (2009) 'Phylogenomics revives traditional views on deep animal relationships', *Curr Biol,* 19(8), 706-12.

Philippe, H. and Douady, C. J. (2003) 'Horizontal gene transfer and phylogenetics', *Curr Opin Microbiol,* 6(5), 498-505.

Philippe, H. and Lopez, P. (2001) 'On the conservation of protein sequences in evolution', *Trends in Biochemical Sciences,* 26(7), 414-6.

Philippe, H., Snell, E. A., Bapteste, E., Lopez, P., Holland, P. W. and Casane, D. (2004) 'Phylogenomics of eukaryotes: impact of missing data on large alignments', *Mol Biol Evol,* 21(9), 1740-52.

Philippe, H., Zhou, Y., Brinkmann, H., Rodrigue, N. and Delsuc, F. (2005b) 'Heterotachy and long-branch attraction in phylogenetics', *BMC Evol Biol,* 5, 50.

Phillips, M. J., Delsuc, F. and Penny, D. (2004) 'Genome-scale phylogeny and the detection of systematic biases', *Molecular Biology and Evolution,* 21(7), 1455-1458.

Pisani, D., Benton, M. J. and Wilkinson, M. (2007) 'Congruence of morphological and molecular phylogenies', *Acta Biotheor,* 55(3), 269-81.

Poe, S. and Wiens, J. J. (2000) 'Character Selection and the Methodology of Morphological Phylogenetics' in Wiens, J. J., ed. *Phylogenetic Analysis of Morphological Data*, Washington and London: Smithsonian Institution Press.

Pollock, D. D., Zwickl, D. J., McGuire, J. A. and Hillis, D. M. (2002) 'Increased taxon sampling is advantageous for phylogenetic inference', *Syst Biol,* 51(4), 664-71.

Popadin, K., Polishchuk, L. V., Mamirova, L., Knorre, D. and Gunbin, K. (2007) 'Accumulation of slightly deleterious mutations in mitochondrial protein-coding genes of large versus small mammals', *Proc Natl Acad Sci U S A,* 104(33), 13390-5.

Posada, D. (2002) 'Evaluation of methods for detecting recombination from DNA sequences: empirical data', *Mol Biol Evol,* 19(5), 708-17.

Posada, D. and Crandall, K. A. (2001) 'Evaluation of methods for detecting recombination from DNA sequences: computer simulations', *Proc Natl Acad Sci U S A,* 98(24), 13757-62.

Posada, D. and Crandall, K. A. (2002) 'The effect of recombination on the accuracy of phylogeny estimation', *J Mol Evol,* 54(3), 396-402.

Prasad, A. B., Allard, M. W. and Green, E. D. (2008) 'Confirming the phylogeny of mammals by use of large comparative sequence data sets', *Mol Biol Evol,* 25(9), 1795-808.

Presgraves, D. C. and Yi, S. V. (2009) 'Doubts about complex speciation between humans and chimpanzees', *Trends Ecol Evol,* 24(10), 533-40.

Quang, L. S., Gascuel, O. and Lartillot, N. (2008) 'Empirical profile mixture models for phylogenetic reconstruction', *Bioinformatics,* 24(20), 2317-2323.

Ragan, M. A. (1992) 'Phylogenetic inference based on matrix representation of trees', *Mol Phylogenet Evol,* 1(1), 53-8.

Rambaut, A. (2001) Se-AL, email to [accessed

Rannala, B. and Yang, Z. (1996) 'Probability distribution of molecular evolutionary trees: a new method of phylogenetic inference', *J Mol Evol,* 43(3), 304-11.

Ratnakumar, A., Mousset, S., Glemin, S., Berglund, J., Galtier, N., Duret, L. and Webster, M. T. (2010) 'Detecting positive selection within genomes: the

problem of biased gene conversion', *Philos Trans R Soc Lond B Biol Sci,* 365(1552), 2571-80.

Reed, R. D. and Sperling, F. A. (1999) 'Interaction of process partitions in phylogenetic analysis: an example from the swallowtail butterfly genus Papilio', *Mol Biol Evol,* 16(2), 286-97.

Reeves, J. H. (1992) 'Heterogeneity in the substitution process of amino acid sites of proteins coded for by mitochondrial DNA', *J Mol Evol,* 35(1), 17-31.

Ren, F., Tanaka, H. and Yang, Z. (2009) 'A likelihood look at the supermatrix-supertree controversy', *Gene,* 441(1-2), 119-25.

Reyes, A., Gissi, C., Catzeflis, F., Nevo, E., Pesole, G. and Saccone, C. (2004) 'Congruent mammalian trees from mitochondrial and nuclear genes using Bayesian methods', *Mol Biol Evol,* 21(2), 397-403.

Reyes, A., Pesole, G. and Saccone, C. (2000) 'Long-branch attraction phenomenon and the impact of among-site rate variation on rodent phylogeny', *Gene,* 259(1-2), 177-87.

Rieseberg, L. H., Baird, S. J. and Gardner, K. A. (2000) 'Hybridization, introgression, and linkage evolution', *Plant Mol Biol,* 42(1), 205-24.

Roberts, R., Wells, G. A., Stewart, A. F., Dandona, S. and Chen, L. (2010) 'The genome-wide association study--a new era for common polygenic disorders', *J Cardiovasc Transl Res,* 3(3), 173-82.

Roca, A. L., Bar-Gal, G. K., Eizirik, E., Helgen, K. M., Maria, R., Springer, M. S., O'Brien, S. J. and Murphy, W. J. (2004) 'Mesozoic origin for West Indian insectivores', *Nature,* 429(6992), 649-51.

Rocha-Olivares, A., Fleeger, J. W. and Foltz, D. W. (2001) 'Decoupling of molecular and morphological evolution in deep lineages of a meiobenthic harpacticoid copepod', *Mol Biol Evol,* 18(6), 1088-102.

Rokas, A. and Carroll, S. B. (2008) 'Frequent and widespread parallel evolution of protein sequences', *Mol Biol Evol,* 25(9), 1943-53.

Rokas, A., Kruger, D. and Carroll, S. B. (2005) 'Animal evolution and the molecular signature of radiations compressed in time', *Science,* 310(5756), 1933-8.

Rokas, A., Williams, B. L., King, N. and Carroll, S. B. (2003) 'Genome-scale approaches to resolving incongruence in molecular phylogenies', *Nature,* 425(6960), 798-804.

Romiguier, J., Ranwez, V., Douzery, E. J. and Galtier, N. (2010) 'Contrasting GC-content dynamics across 33 mammalian genomes: relationship with life-history traits and chromosome sizes', *Genome Res,* 20(8), 1001-9.

Rosenbaum, H. C., Pomilla, C., Mendez, M., Leslie, M. S., Best, P. B., Findlay, K. P., Minton, G., Ersts, P. J., Collins, T., Engel, M. H., Bonatto, S. L., Kotze, D. P.,

Meyer, M., Barendse, J., Thornton, M., Razafindrakoto, Y., Ngouessono, S., Vely, M. and Kiszka, J. (2009) 'Population structure of humpback whales from their breeding grounds in the South Atlantic and Indian Oceans', *PLoS One,* 4(10), e7318.

Rosenberg, M. S. and Kumar, S. (2001) 'Incomplete taxon sampling is not a problem for phylogenetic inference', *Proc Natl Acad Sci U S A,* 98(19), 10751-6.

Rosenberg, M. S. and Kumar, S. (2003) 'Taxon sampling, bioinformatics, and phylogenomics', *Syst Biol,* 52(1), 119-24.

Rota-Stabelli, O., Campbell, L., Brinkmann, H., Edgecombe, G. D., Longhorn, S. J., Peterson, K. J., Pisani, D., Philippe, H. and Telford, M. J. (2011) 'A congruent solution to arthropod phylogeny: phylogenomics, microRNAs and morphology support monophyletic Mandibulata', *Proc Biol Sci,* 278(1703), 298-306.

Rota-Stabelli, O., Lartillot, N., Philippe, H. and Pisani, D. (2012) 'Serine codon usage bias in deep phylogenomics: pancrustacean relationships as a case study', *Systematic Biology,* Advance Access.

Roy, S. W. and Gilbert, W. (2005) 'Resolution of a deep animal divergence by the pattern of intron conservation', *Proc Natl Acad Sci U S A,* 102(12), 4403-8.

Rubinoff, D. and Holland, B. S. (2005) 'Between two extremes: mitochondrial DNA is neither the panacea nor the nemesis of phylogenetic and taxonomic inference', *Syst Biol,* 54(6), 952-61.

Rzhetsky, A. and Nei, M. (1995) 'Tests of applicability of several substitution models for DNA sequence data', *Mol Biol Evol,* 12(1), 131-51.

Sadava, D., Heller, C.H., Orians, G.H., Purves, W.K., Hillis, D.M. (2006) *LIFE: The Science of Biology, Eight Edition,* U.S.A.

Saitou, N. and Nei, M. (1987) 'The neighbor-joining method: a new method for reconstructing phylogenetic trees', *Mol Biol Evol,* 4(4), 406-25.

Salemi, M. and Vandamme, A., eds. (2003) *The Phylogenetic Handbook,* Cambridge: Cambridge University Press.

Salminen, M. O., Carr, J. K., Burke, D. S. and McCutchan, F. E. (1995) 'Identification of breakpoints in intergenotypic recombinants of HIV type 1 by bootscanning', *AIDS Res Hum Retroviruses,* 11(11), 1423-5.

Samsonova, J. V., Baxter, G. A., Crooks, S. R., Small, A. E. and Elliott, C. T. (2002) 'Determination of ivermectin in bovine liver by optical immunobiosensor', *Biosens Bioelectron,* 17(6-7), 523-9.

Sanderson, M. J. (2002) 'Estimating absolute rates of molecular evolution and divergence times: a penalized likelihood approach', *Mol Biol Evol,* 19(1), 101-9.

Sanderson, M. J., Donoghue, M. J. (1989) 'Patterns of variation in levels of homoplasy.', *Evolution,* 43(8), 1781-1795.

Sanderson, M. J. and Hufford, L. (1996) *Homoplasy: Recurrence of Similarity in Evolution,* Academic Press.

Sanderson, M. J., Purvis, A. and Henze, C. (1998) 'Phylogenetic supertrees: Assembling the trees of life', *Trends Ecol Evol,* 13(3), 105-9.

Sankoff, D., Leduc, G., Antoine, N., Paquin, B., Lang, B. F. and Cedergren, R. (1992) 'Gene order comparisons for phylogenetic inference: evolution of the mitochondrial genome', *Proc Natl Acad Sci U S A,* 89(14), 6575-9.

Scally, M. (2001) 'Molecular evidence for the major clades of placental mammals', *Journal of Mammalian Evolution,* 8(4).

Schierwater, B., Eitel, M., Jakob, W., Osigus, H. J., Hadrys, H., Dellaporta, S. L., Kolokotronis, S. O. and Desalle, R. (2009) 'Concatenated analysis sheds light on early metazoan evolution and fuels a modern "urmetazoon" hypothesis', *Plos Biology,* 7(1), e20.

Schmid, K. and Yang, Z. (2008) 'The trouble with sliding windows and the selective pressure in BRCA1', *PLoS One,* 3(11), e3746.

Schmidt, H. A., Strimmer, K., Vingron, M. and von Haeseler, A. (2002) 'TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing', *Bioinformatics,* 18(3), 502-504.

Schneider, A., Souvorov, A., Sabath, N., Landan, G., Gonnet, G. H. and Graur, D. (2009) 'Estimates of positive Darwinian selection are inflated by errors in sequencing, annotation, and alignment', *Genome Biol Evol,* 1, 114-8.

Seluanov, A., Chen, Z., Hine, C., Sasahara, T. H., Ribeiro, A. A., Catania, K. C., Presgraves, D. C. and Gorbunova, V. (2007) 'Telomerase activity coevolves with body mass not lifespan', *Aging Cell,* 6(1), 45-52.

Seluanov, A., Hine, C., Azpurua, J., Feigenson, M., Bozzella, M., Mao, Z., Catania, K. C. and Gorbunova, V. (2009) 'Hypersensitivity to contact inhibition provides a clue to cancer resistance of naked mole-rat', *Proc Natl Acad Sci U S A,* 106(46), 19352-7.

Semple, C. and Steel, M. (2000) 'A supertree method for rooted trees', *Discrete Appl. Math,* 105, 147.

Shapiro, B., Rambaut, A. and Drummond, A. J. (2006) 'Choosing appropriate substitution models for the phylogenetic analysis of protein-coding sequences', *Mol Biol Evol,* 23(1), 7-9.

Shaw, K. L. (2002) 'Conflict between nuclear and mitochondrial DNA phylogenies of a recent species radiation: what mtDNA reveals and conceals about modes of speciation in Hawaiian crickets', *Proc Natl Acad Sci U S A,* 99(25), 16122-7.

Shimodaira, H. and Hasegawa, M. (2001) 'CONSEL: for assessing the confidence of phylogenetic tree selection', *Bioinformatics,* 17(12), 1246-7.

Shoshani, J., Groves, C. P., Simons, E. L. and Gunnell, G. F. (1996) 'Primate phylogeny: morphological vs. molecular results', *Mol Phylogenet Evol,* 5(1), 102-54.

Sidow, A. and Wilson, A. C. (1990) 'Compositional statistics: an improvement of evolutionary parsimony and its application to deep branches in the tree of life', *J Mol Evol,* 31(1), 51-68.

Simmons, M. P., Carr, T. G. and O'Neill, K. (2004) 'Relative character-state space, amount of potential phylogenetic information, and heterogeneity of nucleotide and amino acid characters', *Mol Phylogenet Evol,* 32(3), 913-26.

Slatkin, M. (2004) 'A population-genetic test of founder effects and implications for Ashkenazi Jewish diseases', *Am J Hum Genet,* 75(2), 282-93.

Smith, G. J., Vijaykrishna, D., Bahl, J., Lycett, S. J., Worobey, M., Pybus, O. G., Ma, S. K., Cheung, C. L., Raghwani, J., Bhatt, S., Peiris, J. S., Guan, Y. and Rambaut, A. (2009) 'Origins and evolutionary genomics of the 2009 swine-origin H1N1 influenza A epidemic', *Nature,* 459(7250), 1122-5.

Smith, J. M. (1992) 'Analyzing the mosaic structure of genes', *J Mol Evol,* 34(2), 126-9.

Smith, N. G. and Eyre-Walker, A. (2002) 'Adaptive protein evolution in Drosophila', *Nature,* 415(6875), 1022-4.

Snel, B., Bork, P. and Huynen, M. A. (1999) 'Genome phylogeny based on gene content', *Nat Genet,* 21(1), 108-10.

Song, S., Liu, L., Edwards, S. V. and Wu, S. (2012) 'Resolving conflict in eutherian mammal phylogeny using phylogenomics and the multispecies coalescent model', *Proc Natl Acad Sci U S A,* 109(37), 14942-7.

Sonnhammer, E. L. and Koonin, E. V. (2002) 'Orthology, paralogy and proposed classification for paralog subtypes', *Trends Genet,* 18(12), 619-20.

Soulier, J., Leblanc, T., Larghero, J., Dastot, H., Shimamura, A., Guardiola, P., Esperou, H., Ferry, C., Jubert, C., Feugeas, J. P., Henri, A., Toubert, A., Socie, G., Baruchel, A., Sigaux, F., D'Andrea, A. D. and Gluckman, E. (2005) 'Detection of somatic mosaicism and classification of Fanconi anemia patients by analysis of the FA/BRCA pathway', *Blood,* 105(3), 1329-36.

Springer, M. S., Burk-Herrick, A., Meredith, R., Eizirik, E., Teeling, E., O'Brien, S. J. and Murphy, W. J. (2007) 'The adequacy of morphology for reconstructing the early history of placental mammals', *Syst Biol,* 56(4), 673-84.

Springer, M. S., DeBry, R. W., Douady, C., Amrine, H. M., Madsen, O., de Jong, W. W. and Stanhope, M. J. (2001) 'Mitochondrial versus nuclear gene sequences in deep-level mammalian phylogeny reconstruction', *Mol Biol Evol,* 18(2), 132-43.

Springer, M. S., Murphy, W. J., Eizirik, E. and O'Brien, S. J. (2003) 'Placental mammal diversification and the Cretaceous-Tertiary boundary', *Proc Natl Acad Sci U S A,* 100(3), 1056-61.

Springer, M. S., Stanhope, M. J., Madsen, O. and de Jong, W. W. (2004) 'Molecules consolidate the placental mammal tree', *Trends Ecol Evol,* 19(8), 430-8.

Stamatakis, A. (2006) 'RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models', *Bioinformatics,* 22(21), 2688-90.

Stamatakis, A., Auch, A. F., Meier-Kolthoff, J. and Goker, M. (2007) 'AxPcoords & parallel AxParafit: statistical co-phylogenetic analyses on thousands of taxa', *Bmc Bioinformatics,* 8, 405.

Stanhope, M. J., Waddell, V. G., Madsen, O., de Jong, W., Hedges, S. B., Cleven, G. C., Kao, D. and Springer, M. S. (1998) 'Molecular evidence for multiple origins of Insectivora and for a new order of endemic African insectivore mammals', *Proc Natl Acad Sci U S A,* 95(17), 9967-72.

Stearns, S. C. a. K., J.C. (2008) *Evolution in Heath and Disease,* 2 ed*.,* Oxford: Oxford University Press.

Stewart, C. B., Schilling, J. W. and Wilson, A. C. (1987) 'Adaptive evolution in the stomach lysozymes of foregut fermenters', *Nature,* 330(6146), 401-4.

Stewart, C. B. and Wilson, A. C. (1987) 'Sequence convergence and functional adaptation of stomach lysozymes from foregut fermenters', *Cold Spring Harb Symp Quant Biol,* 52, 891-9.

Stone, M. (1974) 'Cross-Validatory Choice and Assessment of Statistical Predictions', *Journal of the Royal Statistical Society. Series B (Methodological),*
, 36(2), 111-147.

Strimmer, K. and von Haeseler, A. (1997) 'Likelihood-mapping: a simple method to visualize phylogenetic content of a sequence alignment', *Proc Natl Acad Sci U S A,* 94(13), 6815-9.

Sullivan, J. and Swofford, D. L. (1997) 'Are guinea pigs rodents? The importance of adequate models in molecular phylogenies', *J. Mammal. Evol,* 4, 77-86.

Swofford, D. L. (2002) *Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4,* Sunderland, Massachusetts: SinUER ASSOCIATES.

Syring, J., Farrell, K., Businsky, R., Cronn, R. and Liston, A. (2007) 'Widespread genealogical nonmonophyly in species of Pinus subgenus Strobus', *Syst Biol,* 56(2), 163-81.

Tamura, K. and Nei, M. (1993) 'Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees', *Mol Biol Evol,* 10(3), 512-26.

Tavaré, S. (1996) 'Some probabilistic and statistical problems in the analysis of DNA sequences.', *Some mathematical questions in biology-DNA sequence anaylsis*, 29.

Taylor, M. S., Kai, C., Kawai, J., Carninci, P., Hayashizaki, Y. and Semple, C. A. (2006) 'Heterotachy in mammalian promoter evolution', *PLoS Genet,* 2(4), e30.

Tenesa, A., Navarro, P., Hayes, B. J., Duffy, D. L., Clarke, G. M., Goddard, M. E. and Visscher, P. M. (2007) 'Recent human effective population size estimated from linkage disequilibrium', *Genome Res,* 17(4), 520-6.

Tennessen, J. A. (2008) 'Positive selection drives a correlation between non-synonymous/synonymous divergence and functional divergence', *Bioinformatics,* 24(12), 1421-5.

Theissen, G. (2002) 'Secret life of genes', *Nature,* 415(6873), 741.

Thomas, J. A., Welch, J. J., Woolfit, M. and Bromham, L. (2006) 'There is no universal molecular clock for invertebrates, but rate variation does not scale with body size', *Proc Natl Acad Sci U S A,* 103(19), 7366-71.

Thompson, J. D., Plewniak, F., Ripp, R., Thierry, J. C. and Poch, O. (2001) 'Towards a reliable objective function for multiple sequence alignments', *J Mol Biol,* 314(4), 937-51.

Thompson, J. D., Thierry, J. C. and Poch, O. (2003) 'RASCAL: rapid scanning and correction of multiple sequence alignments', *Bioinformatics,* 19(9), 1155-61.

Thomson, J. D., Wilson, P. (2008) 'Explaining Evolutionary Shifts between Bee and Humming Bird pollination: Convergence, Divergence and Directionality', *International Journal of Plant Sciences,* 169(1), 23-38.

Thomson, T. M., Lozano, J. J., Loukili, N., Carrio, R., Serras, F., Cormand, B., Valeri, M., Diaz, V. M., Abril, J., Burset, M., Merino, J., Macaya, A., Corominas, M. and Guigo, R. (2000) 'Fusion of the human gene for the polyubiquitination coeffector UEV1 with Kua, a newly identified gene', *Genome Res,* 10(11), 1743-56.

Tibbetts, R. S., Cortez, D., Brumbaugh, K. M., Scully, R., Livingston, D., Elledge, S. J. and Abraham, R. T. (2000) 'Functional interactions between BRCA1 and the checkpoint kinase ATR during genotoxic stress', *Genes Dev,* 14(23), 2989-3002.

Tobe, S. S., Kitchener, A. C. and Linacre, A. M. (2010) 'Reconstructing mammalian phylogenies: a detailed comparison of the cytochrome B and cytochrome oxidase subunit I mitochondrial genes', *PLoS One,* 5(11), e14156.

Tommerup, H., Dousmanis, A. and de Lange, T. (1994) 'Unusual chromatin in human telomeres', *Mol Cell Biol,* 14(9), 5777-85.

Tuffley, C. and Steel, M. (1998) 'Modeling the covarion hypothesis of nucleotide substitution', *Math Biosci,* 147(1), 63-91.

UniProt (2012) 'Reorganizing the protein space at the Universal Protein Resource (UniProt)', *Nucleic Acids Res,* 40(Database issue), D71-5.

Usanga, E. A. and Luzzatto, L. (1985) 'Adaptation of Plasmodium falciparum to glucose 6-phosphate dehydrogenase-deficient host red cells by production of parasite-encoded enzyme', *Nature,* 313(6005), 793-5.

van de Lagemaat, L. N., Gagnier, L., Medstrand, P. and Mager, D. L. (2005) 'Genomic deletions and precise removal of transposable elements mediated by short identical DNA segments in primates', *Genome Research,* 15(9), 1243-9.

van Rheede, T., Bastiaans, T., Boone, D. N., Hedges, S. B., de Jong, W. W. and Madsen, O. (2006) 'The platypus is in its place: nuclear genes and indels confirm the sister group relation of monotremes and Therians', *Mol Biol Evol,* 23(3), 587-97.

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., Gocayne, J. D., Amanatides, P., Ballew, R. M., Huson, D. H., Wortman, J. R., Zhang, Q., Kodira, C. D., Zheng, X. H., Chen, L., Skupski, M., Subramanian, G., Thomas, P. D., Zhang, J., Gabor Miklos, G. L., Nelson, C., Broder, S., Clark, A. G., Nadeau, J., McKusick, V. A., Zinder, N., Levine, A. J., Roberts, R. J., Simon, M., Slayman, C., Hunkapiller, M., Bolanos, R., Delcher, A., Dew, I., Fasulo, D., Flanigan, M., Florea, L., Halpern, A., Hannenhalli, S., Kravitz, S., Levy, S., Mobarry, C., Reinert, K., Remington, K., Abu-Threideh, J., Beasley, E., Biddick, K., Bonazzi, V., Brandon, R., Cargill, M., Chandramouliswaran, I., Charlab, R., Chaturvedi, K., Deng, Z., Di Francesco, V., Dunn, P., Eilbeck, K., Evangelista, C., Gabrielian, A. E., Gan, W., Ge, W., Gong, F., Gu, Z., Guan, P., Heiman, T. J., Higgins, M. E., Ji, R. R., Ke, Z., Ketchum, K. A., Lai, Z., Lei, Y., Li, Z., Li, J., Liang, Y., Lin, X., Lu, F., Merkulov, G. V., Milshina, N., Moore, H. M., Naik, A. K., Narayan, V. A., Neelam, B., Nusskern, D., Rusch, D. B., Salzberg, S., Shao, W., Shue, B., Sun, J., Wang, Z., Wang, A., Wang, X., Wang, J., Wei, M., Wides, R., Xiao, C., Yan, C., et al. (2001) 'The sequence of the human genome', *Science,* 291(5507), 1304-51.

Vernooy, R., Haribabu, E., Muller, M. R., Vogel, J. H., Hebert, P. D., Schindel, D. E., Shimura, J. and Singer, G. A. (2010) 'Barcoding life to conserve biological diversity: beyond the taxonomic imperative', *PLoS Biol,* 8(7), e1000417.

Vitkup, D., Kharchenko, P. and Wagner, A. (2006) 'Influence of metabolic network structure and function on enzyme evolution', *Genome Biol,* 7(5), R39.

Waddell, P. J., Cao, Y., Hauf, J. and Hasegawa, M. (1999) 'Using novel phylogenetic methods to evaluate mammalian mtDNA, including amino acid-invariant sites-LogDet plus site stripping, to detect internal conflicts in the data, with special reference to the positions of hedgehog, armadillo, and elephant', *Syst Biol,* 48(1), 31-53.

Waddell, P. J. and Shelley, S. (2003) 'Evaluating placental inter-ordinal phylogenies with novel sequences including RAG1, gamma-fibrinogen, ND6, and mt-tRNA, plus MCMC-driven nucleotide, amino acid, and codon models', *Mol Phylogenet Evol,* 28(2), 197-224.

Waeschenbach, A., Taylor, P. D. and Littlewood, D. T. (2012) 'A molecular phylogeny of bryozoans', *Mol Phylogenet Evol,* 62(2), 718-35.

Wang, H. F., Takenaka, K., Nakanishi, A. and Miki, Y. (2011) 'BRCA2 and nucleophosmin coregulate centrosome amplification and form a complex with the Rho effector kinase ROCK2', *Cancer Res,* 71(1), 68-77.

Wang, L., McAllan, B. M. and He, G. (2011) 'Telomerase activity in the bats Hipposideros armiger and Rousettus leschenaultia', *Biochemistry (Mosc),* 76(9), 1017-21.

Wang, W., Brunet, F. G., Nevo, E. and Long, M. (2002) 'Origin of sphinx, a young chimeric RNA gene in Drosophila melanogaster', *Proc Natl Acad Sci U S A,* 99(7), 4448-53.

Watters, J. W. and McLeod, H. L. (2003) 'Cancer pharmacogenomics: current and future applications', *Biochim Biophys Acta,* 1603(2), 99-111.

Webster, M. T., Axelsson, E. and Ellegren, H. (2006) 'Strong regional biases in nucleotide substitution in the chicken genome', *Mol Biol Evol,* 23(6), 1203-16.

Weiller, G. F. (1998) 'Phylogenetic profiles: a graphical method for detecting genetic recombinations in homologous sequences', *Mol Biol Evol,* 15(3), 326-35.

Whelan, S. and Blackburne, B. P. (2012) *Characterizing the effect of multiple sequence alignment on downstream analyses,* translated by Dublin.

Whelan, S. and Goldman, N. (2001) 'A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach', *Mol Biol Evol,* 18(5), 691-9.

White, T. R., Conrad, M. M., Tseng, R., Balayan, S., Golding, R., de Frias Martins, A. M. and Dayrat, B. A. (2011) 'Ten new complete mitochondrial genomes of pulmonates (Mollusca: Gastropoda) and their impact on phylogenetic relationships', *BMC Evol Biol,* 11, 295.

Whitlock, M. C. and Bürger, R. (2004) 'Fixation of New Mutations in Small Populations', *Evolutionary Conservation Biology*, 155-170.

Wiens, J. J. (2003) 'Missing data, incomplete taxa, and phylogenetic accuracy', *Syst Biol,* 52(4), 528-38.

Wiens, J. J. (2006) 'Missing data and the design of phylogenetic analyses', *J Biomed Inform,* 39(1), 34-42.

Wiens, J. J., Chippindale, P. T. and Hillis, D. M. (2003) 'When are phylogenetic analyses misled by convergence? A case study in Texas cave salamanders', *Syst Biol,* 52(4), 501-14.

Wiens, J. J., Kuczynski, C. A., Arif, S. and Reeder, T. W. (2010) 'Phylogenetic relationships of phrynosomatid lizards based on nuclear and mitochondrial data,

and a revised phylogeny for Sceloporus', *Molecular Phylogenetics and Evolution,* 54(1), 150-61.

Wildman, D. E., Uddin, M., Opazo, J. C., Liu, G., Lefort, V., Guindon, S., Gascuel, O., Grossman, L. I., Romero, R. and Goodman, M. (2007) 'Genomics, biogeography, and the diversification of placental mammals', *Proc Natl Acad Sci U S A,* 104(36), 14395-400.

Wilkinson, T. N., Speed, T. P., Tregear, G. W. and Bathgate, R. A. (2005) 'Evolution of the relaxin-like peptide family', *BMC Evol Biol,* 5, 14.

Willi, Y., Buskirk, J.V., Hoffmann, A.A. (2006) 'Limits to the Adaptive Potential of Small Populations', *Annual Review of Ecology, Evolution, and Systematics,* 37, 433-458.

Wilson, A. C., Carlson, S. S. and White, T. J. (1977) 'Biochemical evolution', *Annu Rev Biochem,* 46, 573-639.

Wolf, Y. I., Rogozin, I. B., Kondrashov, A. S. and Koonin, E. V. (2001) 'Genome alignment, evolution of prokaryotic genome organization, and prediction of gene function using genomic context', *Genome Res,* 11(3), 356-72.

Wong, W. S., Yang, Z., Goldman, N. and Nielsen, R. (2004) 'Accuracy and power of statistical methods for detecting adaptive evolution in protein coding sequences and for identifying positively selected sites', *Genetics,* 168(2), 1041-51.

Woolfit, M. and Bromham, L. (2003) 'Increased rates of sequence evolution in endosymbiotic bacteria and fungi with small effective population sizes', *Mol Biol Evol,* 20(9), 1545-55.

Wu, C. I. and Li, W. H. (1985) 'Evidence for higher rates of nucleotide substitution in rodents than in man', *Proc Natl Acad Sci U S A,* 82(6), 1741-5.

Yamamichi, M., Gojobori, J. and Innan, H. (2012) 'An autosomal analysis gives no genetic evidence for complex speciation of humans and chimpanzees', *Mol Biol Evol,* 29(1), 145-56.

Yang, S., Doolittle, R. F. and Bourne, P. E. (2005) 'Phylogeny determined by protein domain content', *Proc Natl Acad Sci U S A,* 102(2), 373-8.

Yang, Z. (1994) 'Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods', *J Mol Evol,* 39(3), 306-14.

Yang, Z. (1996) 'Among-site rate variation and its impact on phylogenetic analyses', *Trends Ecol Evol,* 11(9), 367-72.

Yang, Z. (1997) 'PAML: a program package for phylogenetic analysis by maximum likelihood', *Computer Applications in the Biosciences,* 13(5), 555-6.

Yang, Z. (1998) 'Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution', *Mol Biol Evol,* 15(5), 568-73.

Yang, Z. (2006) *Computational Molecular Evolution,* Oxford University Press.

Yang, Z. and Nielsen, R. (1998) 'Synonymous and nonsynonymous rate variation in nuclear genes of mammals', *J Mol Evol,* 46(4), 409-18.

Yang, Z. and Nielsen, R. (2002) 'Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages', *Mol Biol Evol,* 19(6), 908-17.

Yang, Z., Nielsen, R., Goldman, N. and Pedersen, A. M. (2000) 'Codon-substitution models for heterogeneous selection pressure at amino acid sites', *Genetics,* 155(1), 431-49.

Yang, Z., Nielsen, R. and Hasegawa, M. (1998) 'Models of amino acid substitution and applications to mitochondrial protein evolution', *Mol Biol Evol,* 15(12), 1600-11.

Yang, Z., Wong, W. S. and Nielsen, R. (2005) 'Bayes empirical bayes inference of amino acid sites under positive selection', *Mol Biol Evol*
, 22(4), 1107-18.

Yokoyama, S., Tada, T., Zhang, H. and Britt, L. (2008) 'Elucidation of phenotypic adaptations: Molecular analyses of dim-light vision proteins in vertebrates', *Proc Natl Acad Sci U S A,* 105(36), 13480-5.

Zhai, W., Nielsen, R., Goldman, N. and Yang, Z. (2012) 'Looking for Darwin in Genomic Sequences--Validity and Success of Statistical Methods', *Mol Biol Evol*.

Zhang, J. (2004) 'Frequent false detection of positive selection by the likelihood method with branch-site models', *Mol Biol Evol,* 21(7), 1332-9.

Zhang, J., Nielsen, R. and Yang, Z. (2005) 'Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level', *Mol Biol Evol,* 22(12), 2472-9.

Zharkikh, A. (1994) 'Estimation of evolutionary distances between nucleotide sequences', *J Mol Evol,* 39(3), 315-29.

Zhou, Y., Rodrigue, N., Lartillot, N. and Philippe, H. (2007) 'Evaluation of the models handling heterotachy in phylogenetic inference', *BMC Evol Biol,* 7, 206.

Zuckerkandl, E. and Pauling, L. (1965) 'Molecules as documents of evolutionary history', *J Theor Biol,* 8(2), 357-66.

Zuckerkandl, E., Pauling, L.B. (1962) 'Molecular disease, evolution, and genetic heterogeneity', *Horizons in Biochemistry*, 189-225.

Zwickl, D. J. and Hillis, D. M. (2002) 'Increased taxon sampling greatly reduces phylogenetic error', *Syst Biol,* 51(4), 588-98.

Publications

Manuscripts Published:

Morgan CC, Shakya K, Webb AE, Walsh TA, Lynch M, Loscher CE, Ruskin HJ, O'Connell MJ. "*Colon cancer associated genes exhibit signatures of positive selection at functionally important positions*". **BMC Evolutionary Biology.** 2012 Jul 12, 12:114. PMID: 22788692.

Contribution by Morgan CC: Homolog identification, MSA generation, phylogeny analyses, selective pressure analyses, *in silico* analyses of positively selected sites, designed and implemented all quality control measures and contributed to the drafting of the manuscript.

Morgan CC, Loughran NB, Walsh TA, Harrison AJ, O'Connell MJ. "*Positive selection neighboring functionally essential sites and disease-implicated regions of mammalian reproductive proteins*". **BMC Evolutionary Biology.** 2010 Feb 11, 10:39. PMID: 20149245.

Contributions by Morgan CC: Data assembly, homolog identification, MSA generation, data quality and phylogeny analyses, selective pressure analyses and contributed to the drafting of the manuscript.

BMC Evolutionary Biology

RESEARCH ARTICLE Open Access

Positive selection neighboring functionally essential sites and disease-implicated regions of mammalian reproductive proteins

Claire C Morgan, Noeleen B Loughran, Thomas A Walsh, Alan J Harrison, Mary J O'Connell*

Abstract

Background: Reproductive proteins are central to the continuation of all mammalian species. The evolution of these proteins has been greatly influenced by environmental pressures induced by pathogens, rival sperm, sexual selection and sexual conflict. Positive selection has been demonstrated in many of these proteins with particular focus on primate lineages. However, the *mammalia* are a diverse group in terms of mating habits, population sizes and germ line generation times. We have examined the selective pressures at work on a number of novel reproductive proteins across a wide variety of *mammalia*.

Results: We show that selective pressures on reproductive proteins are highly varied. Of the 10 genes analyzed in detail, all contain signatures of positive selection either across specific sites or in specific lineages or a combination of both. Our analysis of SP56 and Col1a1 are entirely novel and the results show positively selected sites present in each gene. Our findings for the Col1a1 gene are suggestive of a link between positive selection and severe disease type. We find evidence in our dataset to suggest that interacting proteins are evolving in symphony: most likely to maintain interacting functionality.

Conclusion: Our *in silico* analyses show positively selected sites are occurring near catalytically important regions suggesting selective pressure to maximize efficient fertilization. In those cases where a mechanism of protein function is not fully understood, the sites presented here represent ideal candidates for mutational study. This work has highlighted the widespread rate heterogeneity in mutational rates across the *mammalia* and specifically has shown that the evolution of reproductive proteins is highly varied depending on the species and interacting partners. We have shown that positive selection and disease are closely linked in the Col1a1 gene.

Background

Reproductive proteins are essential for success of sexually reproducing species and indeed for the emergence of new species. In the past it has been observed that reproductive proteins tend to be under positive selective pressure to change, i.e. adaptive evolution, a trend found in a variety of animal species from abalone to primates [1,2]. Adaptive evolution or positive selection is a selective pressure placed on a protein by a change in environment in order to improve the fitness of the organism in that environment.

With changes in environment, that can include mating system, there is a subsequent selective pressure on the protein sequences related to those functions to adapt accordingly. This variation can be detected using the well-known measurements of the rate of non-synonymous substitutions per non-synonymous site (Dn) and synonymous substitutions per synonymous site (Ds) and their ratio ω = Dn/Ds. The detection of adaptive evolution, where the ratio exceeds unity, is referred to as positive Darwinian selection. Detecting positive Darwinian selection in a region of a protein, or indeed in a lineage of a phylogeny, indicates that there is a selective advantage in changing the amino acid sequence in this region. These signals are essential for our understanding of functionally important residues in a protein sequence and protein functional shift.

In general, the rate of mutation that a gene undergoes is contingent on a number of factors including; protein

* Correspondence: mary.oconnell@dcu.ie
Bioinformatics and Molecular Evolution Group, School of Biotechnology, Dublin City University, Glasnevin, Dublin 9, Ireland

structure, presence of gene duplicates, location in the genome, effective population size, germ line generation time, and composition of the sequence (for review see [3]). It has recently been shown that the number of physical interactions of a particular protein also influences the intrinsic rate of evolution [4]. Evidence for the generation time effect has come from studies on various proteins and species including analyses of substitution rates in higher primates and rodents [5], substitution rates in higher grasses and in palms [6], in mammalian genomes [7] and in chloroplast and sex mutation rate ratios [5,6]. With recent advances in sequencing we have an opportunity to examine these effects using a wider selection of proteins and species. Documented selective pressures associated with positive selection in reproductive proteins include: (i) intense sperm competition whereby sperm from numerous males, ejaculated into the female reproductive tract, compete with one another for the prized fertilization of the egg [8]; (ii) evasion of the immune system, whereby surface layer reproductive proteins evolve to evade destruction by the host's immune system [8]; and finally (iii) selective pressures enforced by mating system, related of course to point (i) above. Species that are more promiscuous have increased levels of selective pressure acting on reproductive proteins than species that are monogamous. This later point is illustrated in the study of SEMG2, where adaptive evolution was found to correlate with mating system in primates [9].

In order to determine the variation in selective pressure in these proteins, there are a number of criteria that the data must meet. Firstly, the data must have a robust phylogenetic signal. Secondly, systematic biases that may exist in the data must be minimized, these include but are not limited to: long branch attraction (LBA), amino acid composition bias, base composition bias and unqualified ortholog predictions, all of which may lead to inaccurate estimates of phylogeny. Thirdly, sensitivity to taxa number is a known limitation of methods for detecting positive selection, therefore more than 6 taxa are needed to gain accurate estimations of selective pressure using the maximum likelihood (ML) method applied here [10].

In this study we have selected a subset of proteins that have roles to play in reproduction. Our dataset was composed of three major datatypes, (i) previously published reproductive proteins, (ii) interacting proteins, here we identified proteins shown to interact with (i), and finally (iii), genes identified from microarray experiments as being highly expressed in reproductive tissues. For group (iii) we assume that those proteins highly expressed in reproductive tissues are important for the function of that tissue. The previously untested reproductive proteins analysed here are from data types (ii) and (iii) outlined above. These novel proteins are SP56, Porimin and Col1a1. SP56 is sperm binding protein number 56, this protein is a representative of the interacting protein subset of sequences analysed. SP56 has been shown to interact with ZP3 - a well-studied reproductive protein. Both Porimin and Col1a1 have been identified from published microarray experiments on normal human tissue [11], and were selected for analysis due to their high levels of expression in reproductive tissues in that study. Porimin is a transmembrane protein that is highly expressed in the uterus, prostate and placenta and Col1a1 is highly expressed in the uterus. Further evidence for the link between Porimin and reproduction was not available in the literature and therefore results from this particular gene are taken with caution until this protein is further characterized. Col1a1 plays an important role during spermatogenesis where it mediates the detachment and migration of germ cells, thus adding further support for its role in reproduction [12].

We have analyzed these data with an approach sensitive to all the systematic biases and limitations of methods given above. A number of genes in our dataset have been analyzed previously but have not taken these limitations and considerations into account. We have expanded these datasets to include a greater number of taxa, we have analyzed all of these genes for evidence of systematic biases and we have used improved models of codon evolution. In this paper we have included models that allow for rate variation across the sequence and across the phylogeny.

Results and Discussion

We performed phylogenetic analyses on all 11 datasets. The resultant gene trees were found to conflict with the canonical phylogeny species ([13], as adapted in Figure 1. The only exception was the Catsper1 mammalian dataset. We postulate the following causes for this conflict: (1) amino acid and/or base composition bias, (2) lack of phylogenetic signal in the data, and finally (3), LBA caused by mixtures of long and short germ line generation times (see Figure 2 for a sample of species and their germ line generation times from our dataset). What follows is a summary of the results of the tests of data quality and bias we performed, see Table 1 for synopsis. We carried out these tests to determine in each case whether these conflicting phylogenies are accurate descriptions of history or whether the data are subject to these known issues listed 1-3 above. Subsequent statistical comparison of the gene trees and species phylogeny using the Shimodaira Hasegawa (SH) test [14] revealed that there is no statistical difference between the gene and species trees in each case, see Table 2 for results of SH tests. The only exceptions

Table 1 Summary of the analysis of quality and bias present in the data

GENE	DATA QUALITY			PHYLOGENETIC ANALYSIS		
	LM Category	AA Comp Bias	Base Comp Bias	Substitution Model	Gene v Species Tree	LBA Artifact
Adam2	1	Pass	Pass	JTT+G	Unresolved	No
Catsper1 Exon1	1	Pass	Pass	JTT+I+G+F	Unresolved	No
Catsper1 Mammals	1	Pass	Pass	JTT+G+F	Unresolved	No
Col1a1	1	Pass	Pass	JTT+G	Unresolved	No
Ph20	1	Pass	Pass	JTT+G+F	Resolved	Yes
Porimin	1	Pass	Pass	JTT+G+F	Unresolved	No
Prkar2a	2	Pass	Pass	JTT+I+G	Unresolved	No
Semg2	1	Pass	Pass	JTT+G+F	Unresolved	No
Sp56	2	Pass	Pass	JTT+I+G	Unresolved	No
Zp2	1	Pass	Pass	JTT+G	Unresolved	No
Zp3	1	Pass	Pass	JTT+G+F	Unresolved	No

Genes with significant signal are given in the Likelihood mapping, or, L.M. Category column, see text for explanation of the category 1 and 2 in this column. Results of the amino acid composition and nucleotide base composition bias tests, are shown in the A.A. Comp Bias and Base Comp Bias columns respectively. The phylogenetic trees for each gene are drawn using the substitution model described where G = gamma distributed rates across sites, I = invariable sites, F = frequency of amino acids, JTT = Jones Taylor Thornton model. In the case of LBA analysis, No = no evidence of LBA in the gene analysed, Yes = evidence of LBA in the gene analysed.

were Prkar2a and ZP3 where the presence of polytomies in the gene trees caused the preference of the unresolved nodes over the resolved nodes.

1. Tests of Data Quality and Bias

(i) Test for amino acid and base composition biases

We tested all multiple sequence alignments (MSAs) for evidence of significant levels of amino acid composition bias and base composition bias in each lineage using the TreePuzzle software [15]. We found that all alignments passed the significance test with p-values < 0.05, see Table 1 for summary. For full set of amino acid and base composition bias test results, see Additional Files 1 and 2 respectively. In summary the discordance between each of the gene trees and the canonical species phylogeny is not a result of amino acid or base composition biases providing evidence of false relationships.

Table 2 Summary of SH tests for complete gene datasets

Gene	SH - gene	SH - ideal	Best-fit Tree
Adam2	1.0000	0.1200	NS
Catsper1 Exon1	1.0000	0.1460	NS
Catsper1 mammals	0.5020	1.0000	NS
Col1a1	1.0000	0.2650	NS
Ph20	1.0000	0.3220	NS
Porimin	0.4040	1.0000	NS
Prkar2a	1.0000	0.0490	gene
Semg2	1.0000	0.1010	NS
Sp56	1.0000	0.2380	NS
Zp2	0.1620	1.0000	NS
Zp3	1.0000	0.0050	gene

For each gene, the likelihood of estimated Bayesian phylogeny (gene) and corresponding ideal species tree (ideal) to fit the dataset were determined with the SH test at a 5% significance level. Values equal to 1.0000 represent the tree with the lowest log likelihood, values less than 0.05 refer to those cases where there is a significant difference between the two topologies, and the gene tree is a significantly better fit to the data. NS = No Statistical significance between gene and species tree, in these cases the species tree was used.

(ii) Test for phylogenetic signal

We performed the likelihood mapping procedure implemented in the TreePuzzle software [15,16] to determine the level of phylogenetic signal/conflict present in each alignment, for more detail see the *Methods* section. Our initial dataset consisted of 27 genes, we used this filtering step to reduce our dataset to contain only those genes with phylogenetic signal. We categorized the results from the likelihood mapping analysis into 3 main categories of signal: category 1 had strong phylogenetic signal (see Figure 3a), category 2 had medium level of phylogenetic signal (see Figure 3b) and category 3 had low/no levels of phylogenetic signal (see Figure 3c). The results of the test for phylogenetic signal are summarized in Table 1 and in total 9 out of the 27 genes had strong phylogenetic signal (category 1), with an additional 2 genes with moderate levels of phylogenetic signal (category 2). The complete set of results for the likelihood mapping process is given in Additional File 3. The remaining 17 genes failed the test (category 3). The category 3 genes (with low or no levels of phylogenetic signal) were subsequently removed from the analysis, only 10 genes were retained for further analysis.

(iii) Long Branch Attraction (LBA) analysis

We assessed the data for evidence of LBA which would manifest itself in the data by drawing species with a greater number of mutations in the gene of interest together erroneously on the phylogenetic tree. The method applied uses the MSA and the corresponding phylogeny to categorise rates amongst sites, using an

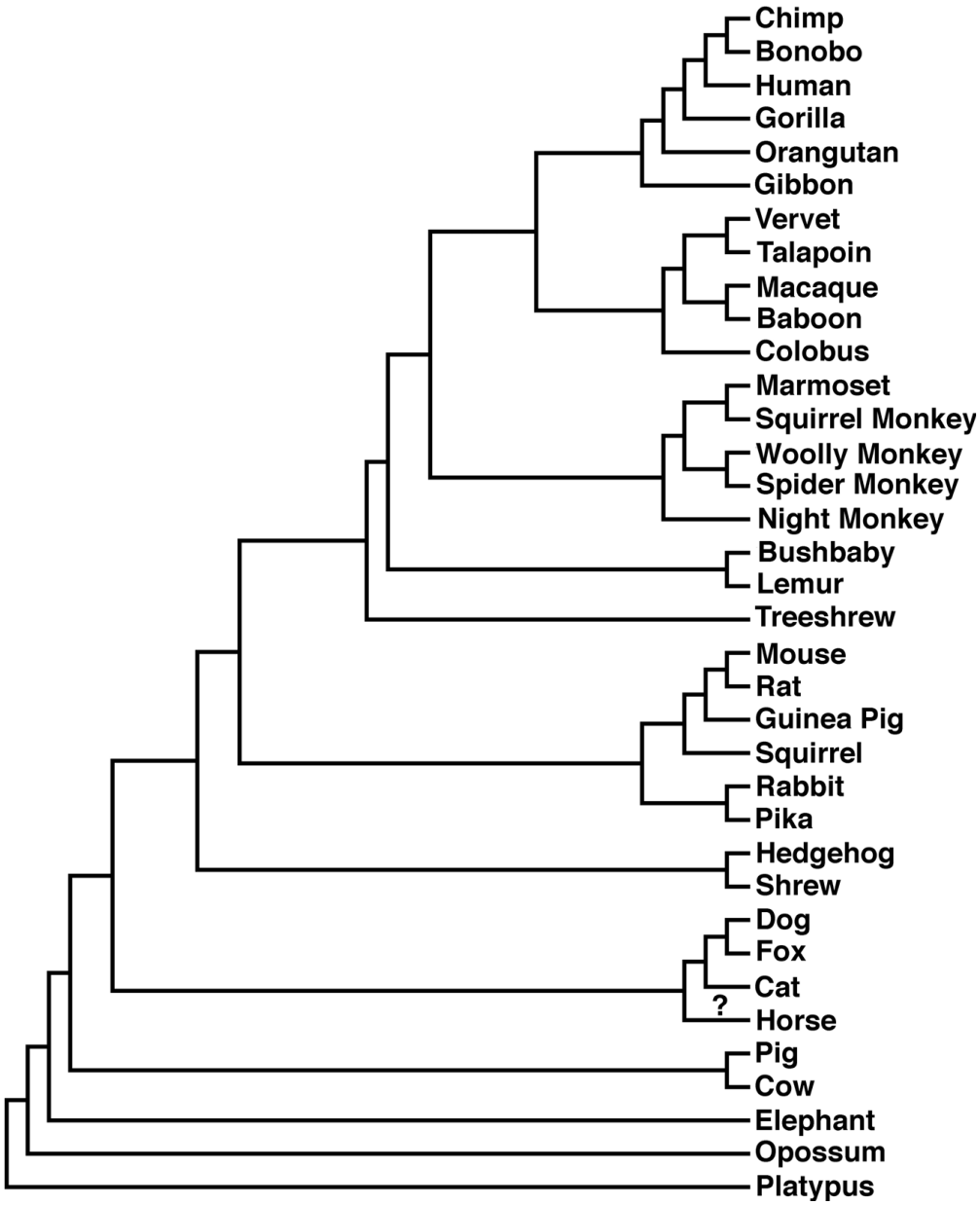


Figure 1 Canonical mammalian species phylogeny. Shown here is a representation of the agreed relationships amongst the *mammalia* for the species used in this analysis. The "?" on the lineage leading to horse indicates controversy over the position of this lineage on the phylogeny.

approach we have previsouly published for mammalian data [17], as described in detail the *Methods* section. In this method of site-stripping we apply the phylogenetic tree (estimated *ab initio* in this software) and the MSA to classify all sites in the alignment into one of eight categories of mutation rate. These are arbitrary categories from 1-8; with 1 being the most highly conserved sites and 8 being the most highly variable. Essentially, these estimates allow us to select only the most conserved sites for phylogeny reconstruction. Sites are sequentially stripped from the alignments based on their rate of evolution and phylogenies are created based on slower evolving sites. These site-stripped phylogenies are then compared to the species tree. Using two independent methods of comparison we determined whether the resultant stripped trees had topologies significantly similar to the species phylogeny. The "root mean squared deviation", or RMSD, method is restricted to binary trees [18], see Additional File 4 for full set of results. Therefore we also employed the SH method of

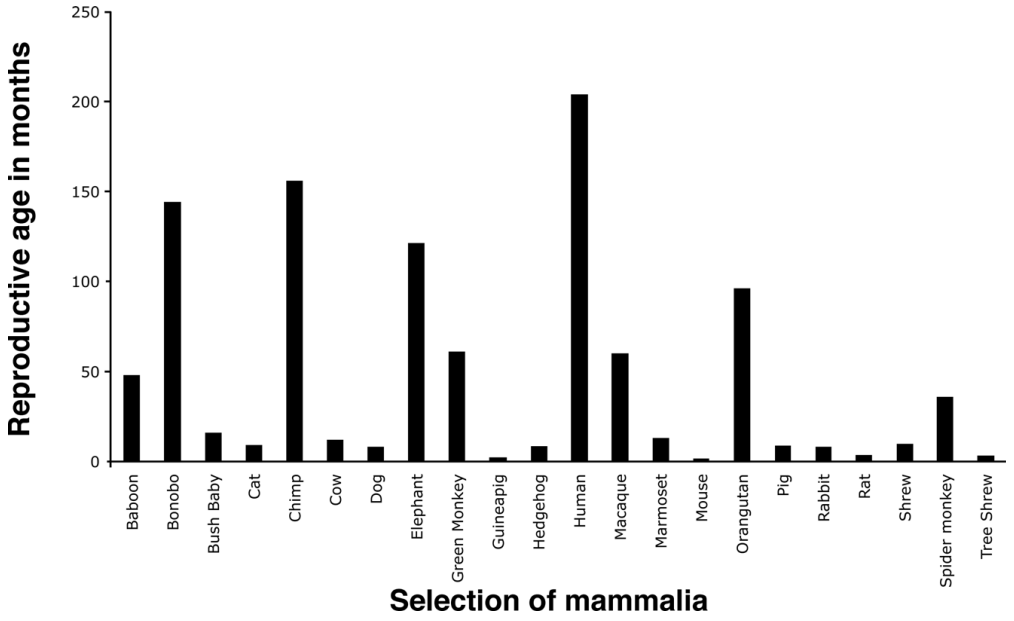


Figure 2 Selection of mammalia used in the analysis and the time to reproductive age in months. Species are shown on the X-axis in alphabetical order. On the Y-axis is the number of months it takes for each species to reach reproductive age.

comparing phylogenies [14], see Additional File 5 for full set of results. For a full description of the RMSD statistic used here [18], see the corresponding section in the *Methods*. Using this approach we could identify the signature of LBA in the Ph20 dataset alone, see Table 1 for summary.

2. Analysis of selective pressures using codon models of evolution

Following analysis of the phylogenies of these reproductive genes, we determined the selective forces at work on these 10 genes (11 datasets). Only those genes passing the data quality tests were analyzed here (*i.e.* 10 genes), see Table 1. In the case of Catsper1, we have analyzed the gene at two different evolutionary distances because it contains high levels of insertion and deletion events. The two datasets for Catsper1 are: exon 1 from the primates only, and, the entire gene from only distant mammalian groups. Hence the number of datasets is 11, and the number of genes tested is 10. The alignments in all cases reached significant levels following randomization tests (z-scores > 1000 in all cases, a z-score of greater than 5 is typically taken as significant).

In those cases where the genes had already been analyzed in previous studies, we expand upon the data in these studies and use more sophisticated models of evolution. ML methods are sensitive to sample size with a minimum of 6 taxa recommended from simulation studies [10]. For a summary of the site-specific and lineage-specific results, see Table 3 and Table 4

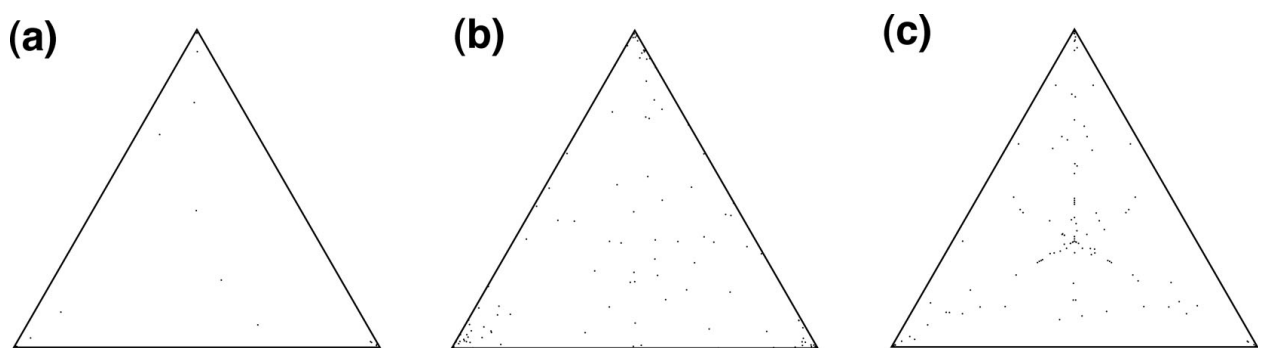


Figure 3 Example of likelihood mapping categories. **(a), category 1** genes with strong phylogenetic signal the example given here is the ZP2 gene, **(b) category 2** genes with intermediate levels of phylogenetic signal, the example given here is the Prkar2a gene, and **(c)**, **category 3** genes with low/no phylogenetic signal, the example given here is the CD9 gene.

respectively. For a summary of all likelihood ratio tests (LRTs) performed in the analyses of these genes see Table A9. In general the lineages tested in the lineage specific analysis for each gene were as follows: modern human; the primate ancestor; modern mouse, and the rodent ancestor, these are indicated in Figure 4(a-k). For certain datasets the species tested varied depending on those species for which high quality sequence data existed for that gene, these are discussed on a gene-by-gene basis below. In summary, for each of the 11 datasets tested, positive selection was detected. In the site-specific test between 7 and 94 sites per gene were identified as positively selected. In the lineage-specific analyses there were up to 2 lineages per gene identified as having evidence of positive selection. Below is a brief description of the results on a gene-by-gene basis, the complete set of all parameters, likelihood values and LRTs are given in Additional File 6.

Col1a1

Possibly the most intriguing result from our entire analysis is that from the Col1a1 protein. According to the microarray study employed here [11], Col1a1 is highly expressed in the uterus tissue. It is also found in most structural tissues including cartilage, bone, tendon, skin and part of the eye (sclera). It is a member of the group 1 collagen proteins involved in the development of the uterine fibroids [19]. There are two propeptide regions to the Col1a1 gene, denoted N- and C-terminal propeptides. According to studies on Col1a1 function, a role has been established for Col1a1 in spermatogenesis [12].

Our site-specific analysis shows 66 sites evolving with an ω value of 4.09, see Table 3. In summary 35/66 of our positively selected sites fall in the N-terminal propeptide region (23-161) and 9/66 positively selected sites fall in the C-terminal propeptide region (1219-1464), this can be seen clearly in Figure 5a. Position 162 in

Table 3 Summary of the results of the site-specific analysis: in each case the most significant model was M8

Gene	n	Parameter estimates	# Positively selected Sites
Adam2	12	p_0 = 0.92632 p = 0.37637 q = 0.60688 p_1 = 0.07368 ω = 3.94326	45>0.50 15>0.95 5>0.99
Catsper1_Exon1 (primates only)	16	p_0 = 0.82736 p = 0.13661 q = 0.03850 p_1 = 0.17264 ω = 3.13071	95>0.50 7>0.95 1>0.99
Catsper1_Mammals (non-primate mammals only)	8	p_0 = 0.83315 p = 0.34233 q = 0.51278 p_1 = 0.16685 ω = 3.26879	124>0.50 30>0.95 8>0.99
Col1a1	10	p0 = 0.98023 p = 0.04796 q = 0.32286 p1 = 0.01977 ω = 4.09285	66>0.50 21>0.95 8>0.99
Ph20	11	p_0 = 0.87658 p = 0.56141 q = 0.83349 p_1 = 0.12342 ω = 2.20500	39>0.50 3>0.95 0>0.99
Porimin	10	p_0 = 0.85067 p = 0.41864 q = 0.32952 p_1 = 0.14933 ω = 12.21841	30>0.50 13>0.95 5>0.99
Prkar2a	17	p_0 = 0.95102 p = 0.16339 q = 0.98823 p_1 = 0.04898 ω = 2.60992	19>0.50 4>0.95 0>0.99
Semg2	12	p_0 = 0.97236 p = 0.01163 q = 0.00500 p_1 = 0.02764 ω = 12.26405	41>0.50 5>0.95 2>0.99
Sp56	14	p_0 = 0.98807 p = 0.16114 q = 1.12262 p_1 = 0.01193 ω = 3.81710	8>0.50 2>0.95 2>0.99
Zp2	18	p_0 = 0.87339 p = 0.63945 q = 0.75356 p_1 = 0.12661 ω = 2.04655	52>0.50 9>0.95 6>0.99
Zp3	13	p_0 = 0.91489 p = 0.30029 q = 0.77328 p_1 = 0.08511 ω = 1.92305	48>0.50 0>0.95 0>0.99

Following LRT analysis M8 was chosen in each case as the most significant model. n refers to the number of taxa in each dataset. The proportion of sites (p), evolving under each corresponding selective pressures (ω) are shown. For example, p_0 refers to the proportion of the protein evolving under the selective pressure value given by ω_0. The parameters p and q describe the beta distribution. The final column gives the number of sites with posterior probability (PP) of 0.50, 0.95 and 0.99 that belong in the positively selected category or sites. The number before the ">" refers to the number of sites with a specific PP value.

Table 4 Summary of lineage-specific positive selection detected.

Species tested as Foreground	Significant LRT	Parameter estimates		
		P	Fwd ω	Bck ω
Adam2				
Macaque	ModelA v M1	9.57%	1.71	0.10/1
Catsper1 Mammals				
Ferungulata	ModelA v M1	4.46%	998.99	0.09/1
Rodents	ModelA v M1	5.45%	999.00	0.084/1
	ModelB v m3Discrtk2	4.47%	999.00	0.12/1.38
Col1a1				
Rodents	ModelA v M1	2.17%	72.73	0.013/1
	ModelB v m3Discrtk2	1.93%	72.77	0.02/1.35
PH-20				
Guinea Pig	ModelA v M1	6.3%	11.48	0.13/1
	ModelB v m3Discrtk2	6.14%	12.57	0.14/1.10
Prkar2a				
Macaque	ModelA v M1	2.37%	999.00	0.04/1
	ModelB v m3Discrtk2	2.53	999.00	0.04/1.22
Sp56				
Human	ModelB v m3Discrtk2	100%	62.40015	0.02/0.55
Glires	ModelB v m3Discrtk2	2.56%	1.03	0.02/0.55

Summary table of significant results for lineages specific analyses following LRT analyses. Lineages tested as foreground (Fwd) are shown in the first column. Only those lineages with significant LRT values for Model B or Model A and ω >1 are shown here. Parameter estimates are given for the LRT values highlighted in bold. P is the proportion of sites under selection the corresponding selective pressure as measured by ω. Fwd ω and Bck ω scores for the foreground species and background species respectively are given in the final column.

Col1a1 is cleaved and modified by an endopeptidase, position 162 is also modified by pyrrolidone carboxcylic acid (Swiss-Prot PO2452). A positively selected site at position 163 is neighboring this multifunctional site, suggesting that there has been an evolutionary effort to improve cleavage and/or modification in this protein.

Variations in Col1a1 are linked with Osteogenesis Imperfecta (OI), an autosomal dominant disease, resulting in an inability to make the correct collagen protein. There are a spectrum of OI conditions, the most severe is OI type 2 (OI-II) leading to death in the perinatal period. A recent extensive study of the Single Nucleotide Polymorphisms (SNPs) associated with OI has revealed a number of substitutions of glycine residues within the triple helical domains of the Col1a1 protein [20]. The total number of disease implicated sites in the Swiss-Prot entry P02452 for Col1a1 is 99: 4 of these are OI non-specific, 4 are OI-I, 59 are OI_II, 14 are OI-IV and 15 are SNPs (2 are associated with another disease). One third of the mutations that result in substitutions for glycine in Col1a1 are lethal whereas those between the start codon and position 200 are non-lethal. Only 1 of the sites we have identified as positively selected is in the non-lethal domain from position 1-200, this is site 195. This positively selected site is neighboring the SNP position 197 that causes a mild OI phenotype. In Table 5 we show a list of 11 positively selected sites that fall in close proximity to sites associated with disease and are located between 280 and 1456, spanning the important triple helix region. These positions are all within 1 to 5 amino acid residues of known disease variants, 8 of these disease variants are the severe/lethal OI-II disease form. Two exclusively lethal regions, helix positions 691-823 and 910-964 aligned with major binding regions [20] and we find a positively selected site in this region. Following a randomization test for the positively selected sites and disease implicated sites (as denoted by Swiss-Prot entry PO2452), we have found that the pattern we observe, i.e. finding positively selected sites in close proximity to disease implicated sites is significant in 3 out of the 11 cases examined here (at $P < 0.05$).

Lineage-specific analysis shows evidence for positive selection in this protein in the rodent ancestor. In total, 2.2% of the sites in the rodent ancestor have $\omega = 72.73$, while the rest of the species are evolving under purifying selection, $\omega = 0.013$. For a summary of site and lineage specific results for Col1a1, see Table 3 and 4. For complete set of results see Additional File 6(d).

Prkar2a (interacts with SEMG2)

Prkar2a is a cAMP dependent protein kinase that is attached to the sperm flagella via regulatory subunit (RII) [21]. Protein tyrosine phosphorylation has been linked with successful fertilization due to hyper-activated sperm motility [22]. This increase in phosphorylation is part of a cAMP dependent pathway that activates protein kinase A [22].

The PRKA families were previously tested for positive selection using 3 to 4 taxa and site-specific model M8 with no significant results for positive selection reported. With our 17 taxa dataset, we were able to detect that 4.7% of sites were evolving at a rate of ω = 2.60, see Table 3 for summary of details.

Positively selected sites detected in the site-specific analysis of Prkar2a were compared to the human Swiss-Prot sequence (P13861). In total 18 sites were predicted to be positively selected, 17 of these sites occur in the region of the protein associated with dimerization and phosphorylation (2-138), see Figure 5(c). In the Swiss-Prot entry there are a number of residues listed as being modified by phosphoserine. These are positions 58, 78, 80, 99 and phosphothreonine at position 54. The sites estimated to be positively selected from our analysis are: 58, 59, 61, 62, 63, 64, 65, 68, 70, 74, 75, these sites are at or in close proximity to these modified residues.

The regulatory subunit alpha 2 of Prkar2a has been shown *in vitro* to interact with Semg2. The phosphorylation of Semg2 may lead to its activation into forming a gel matrix in the female reproductive tract. From our analysis it is shown that while Semg2 has positively selected sites dispersed throughout its sequence, whereas the positively selected sites for Prkar2a are localized to the N-terminus region, and the remainder of the gene is under strong purifying selection. Literature has so far not specified an exact phosphorylation site for Semg2, which prevents us from commenting further on its interactions with Prkar2a.

Lineage-specific analysis shows that Prkar2a in the macaque has undergone a greater selective pressure to change when compared with other *mammalia* in the dataset, with 2.53% of sites evolving at ω = 1.22, see Table 4 for summary of results. For complete set of results for Prkar2a, see Additional File 6(g).

Ph20 (interacts with ZP2 and ZP3)

Ph20 is expressed in the testis and found in the acrosome of the sperm. It is also codes for a receptor that is involved in the sperm to zona pellucida (ZP) adhesion [23].

Previous analysis conducted on this protein involved 6 taxa [24]. Here we have increased the number of taxa to 11. We have omitted the carnivores from our analysis of Ph20 as the sequences were spurious. We found evidence for LBA in the Ph20 dataset. By removing fast evolving sites a fully resolved gene phylogeny is obtained. This gene tree now is in agreement with the ideal species phylogeny ([13].

Lineage-specific analysis shows that guinea pig is under positive selection, with 6.1% of sites with ω = 12.57 while all other species in the background are evolving at ω = 0.14 or neutrally, see Table 4. The 39 positively selected sites were then compared to the human Swiss-Prot sequence (P38567), see Figure 5(b) for results. Catalytically important resides 146, 148, 211 284 and 287 when mutated result in a reduction in, or loss of, activity [25]. It has been shown experimentally that mutations in the region of this active site significantly reduce or completely block the function of this protein [25]. Our results show that 3 of the positively selected sites, 155, 272, 273, are in close proximity to these regions. Another 5 positively selected sites: 83, 155, 252, 353 and 391 are close to glycosylation sites, see Figure 5 (b). These sites when modified are known to change the structure and function of the Ph20 protein. For complete set of results for Ph20 see Additional File 6(e). These results are of significance as the Ph20 protein changes position in the sperm during the different stages of the fertilization process. In guinea pig Ph20 protein is known to migrate from the post acrosomal membrane to the inner acrosomal membrane [26]. Thus finding these positively selected sites in close proximity to these glycosylation sites in guinea pig suggests that these sites have been selected to modify the Ph20 structure more effectively thus increasing the chance of capacitation.

SP56 (interacts with ZP2 and ZP3)

The binding of sperm to the zona pellucida (ZP) is crucial for gamete formation to take place. The exact mechanisms of this process are still to be uncovered therefore any predictions on important residues will greatly improve knowledge by directing mutational studies. SP56 has been shown through photoaffinity cross-linking experiments to have a specific binding affinity for ZP3 [27]. Therefore it is believed to play an important role in the binding of sperm to the ZP matrix. Experiments have shown that during capacitation SP56 is released from the acrosomal matrix and becomes situated in the sperm head membrane, enabling it to act as a ZP3 binding protein [28].

Here we have found 8 positions in the SP56 protein that are under positive selection (ω = 3.82) following site-specific analysis. These sites were compared to the human SP56 entry in Swiss-Prot (Q13228) to determine possible links to function. One of these 8 positively selected sites is position 122, regarded as a SNP number (rs35396382) in dbSNP database [29]. Although further experimental work needs to be conducted to decipher the clinical association of this position, it is extremely interesting that our most significant positively selected site also displays variation in the population, especially given the overall high level of conservation in this gene. For summary of results see Tables 3 and 4, and for full set of results for this gene see Additional File 6(i).

ZP2

Zona pellucida (ZP) proteins form the complex glycoprotein coat that surrounds the oocyte [30]. These ZP

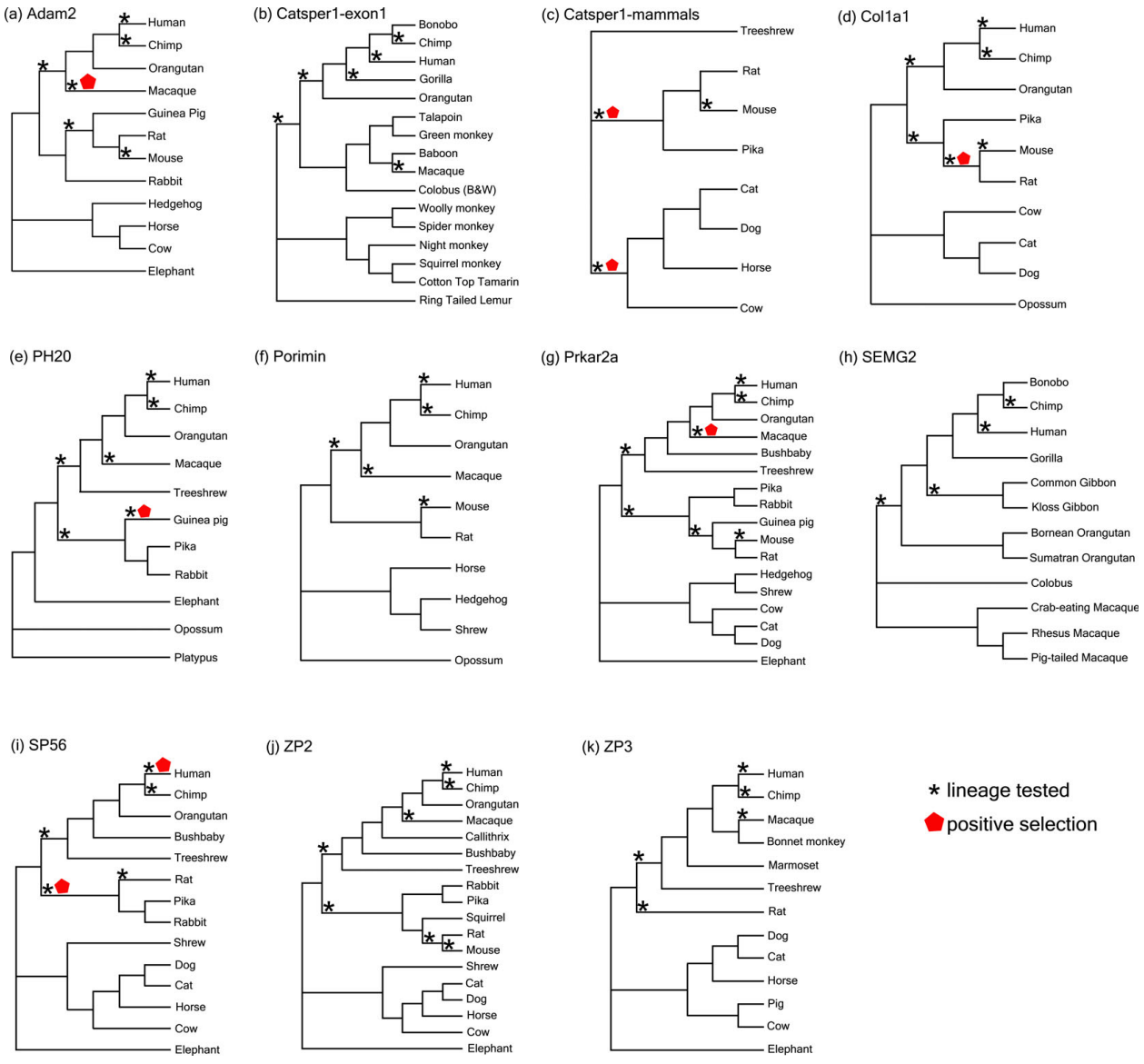


Figure 4 Results of lineage specific positive selection analysis on 11 datasets. The phylogeny used for each gene is a reduced version of the species phylogeny. The lineages labeled as foreground in the analysis are denoted in the diagram with asterix symbols. These are the results following LRT analysis. Those lineages where positive selection was determined are represented by red pentagons.

proteins have been shown to be under strong pressure to change, and results have been published on both site and lineage analyses [31]. Here we have expanded the analysis of ZP2 to include 18 taxa (maximum previously tested = 8 [31]). We have also applied more complex models of evolution and have sampled deeper branches on the phylogeny including a representative of the *Afrotheria* - elephant.

In this case, the results of our larger dataset and more complex models show that the values of ω determined here vary slightly when compared to previous analyses [31]. This previous test showed 4.7% of sites to have ω = 2.5, increasing the size of the dataset in this study results in 52 sites in ZP2 that have an ω value of 2.05. See Additional File 6(j) for complete results.

Positively selected sites were compared to the human Swiss-Prot entry for ZP2 (Q05996) to identify possible function for these sites, see Figure 5(d). ZP2 contains 7 carbohydrate chains situated between sites 87-462, these are important for the sperm to bind to the ZP of the egg coat [32]. Of the 46 sites identified to be under positive selection, 23 fall between positions 66-257, this region contains 5 of the binding domains of the carbohydrate chains. The clustering can be seen more clearly

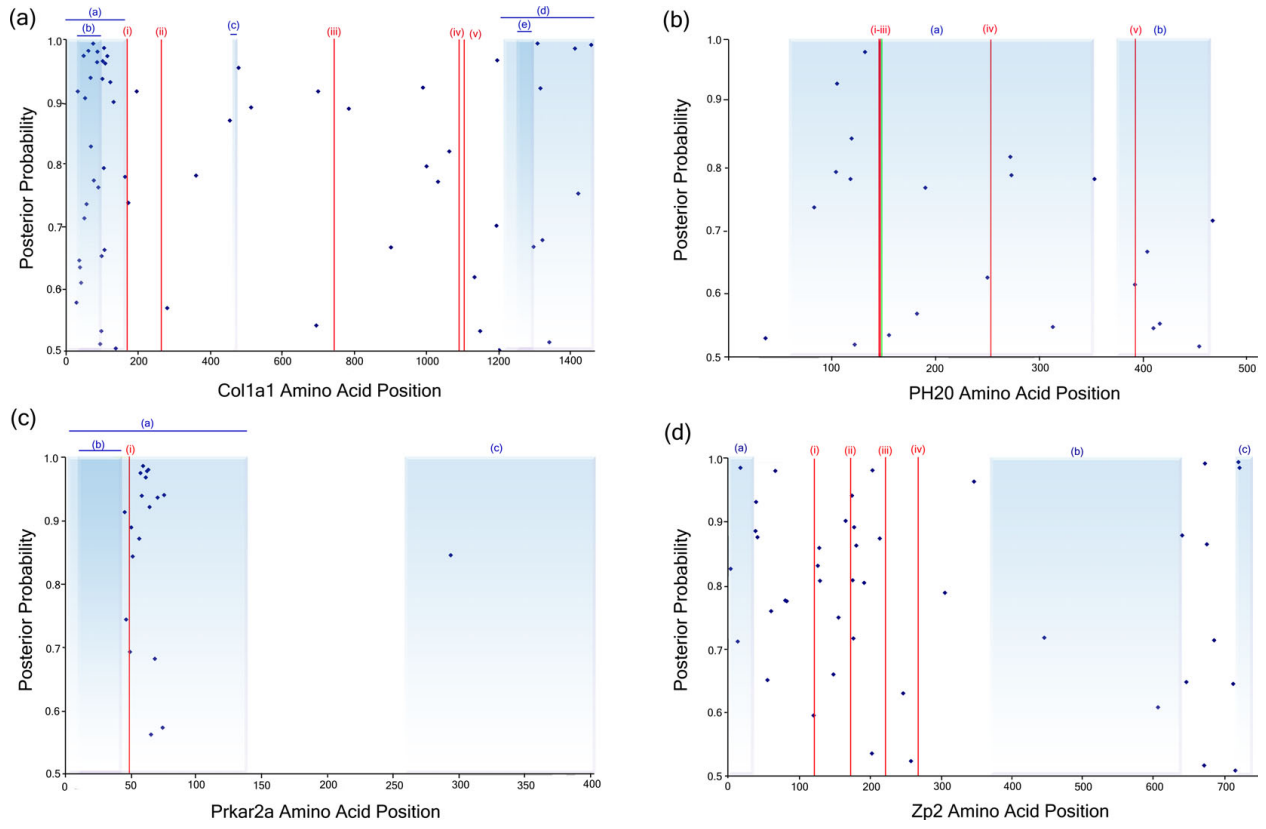


Figure 5 Results of positive selection analysis for 4 genes. Each of the four graphs represents the CDS of a gene from position 1 to the stop codon (X-axis). The Y-axis is the posterior probability of each of the sites belonging to the positively selected category. The dark blue data points are sites estimated to be under positive selection. Alternative pale blue and white regions depict alternative domains in the protein, this data is taken from Swiss-Prot. The vertical red bars in each case represent functionally important sites, these are specific to each gene as follows: **(a) Col1a1**, (i) Cleavage site by procollagen N-endopeptidase, (ii) O-linked Gylcosylation site, (iii) Cell attachment site, (iv) Cell Attachment site, and (v) O-linked Glycosylation site. **(b) PH20**, (i) active site, proton donor, (ii) and (iii) are positions when mutated result in loss of activity, and (iv) N-linked Glycosylation site, **(c) Prkar2a**, (i) and is a Phosphothreoinine modified residue, **(d) ZP2**, (i), (iii) and (iv) are N-linked Glycosylation sites, and (ii) is the cleavage site.

in Figure 5(d). Another cluster of positively selected sites (10 sites in total) occurs in the propeptide region (641-745). It has been suggested that upon the cleavage of the propeptide region, the mature ZP2 protein plays a role in the prevention of polyspermy [33].

ZP3

Analysis of site-specific evolution in ZP3 identified 48 positively selected sites. Of specific interest are positively selected positions 329, 330, 332, 336, 338, 339, as these sites were in close proximity to identified sperm binding sites (329-334) [34], see Table 3. The furin cleavage site is identified at position (350-353), and the propeptide domain at position (351-424). When cleavage takes place the ZP3 undergoes a conformational change that inhibits any further sperm binding to the coat thus preventing polyspermy [35]. Of the 48 positively selected sites identified, 10 fall within the propeptide domain, with an additional 12 occurring close to the vicinity of the furin cleavage and sperm binding sites, thus suggesting that there is a pressure to improve binding and prevent polyspermy. For complete set of results for ZP3, see Additional File 6(k).

Adam2 (Fertilin β)

Adam2 is a cell adhesion molecule that plays a fundamental role in the final binding of sperm to the oocyte membrane [36]. Indirect interactions have been shown with female proteins CD9 [37]. (We have not continued further analysis on CD9, as it failed the likelihood mapping test).

Previous results have been published reporting positive selection using site-specific analysis on 6 taxa [24]. Here we have included 12 taxa for Adam2 and we have investigated the possible functional implications of positively selected sites found. In the site-specific analysis we find 7.3% of sites with $\omega = 3.94$, this corresponds to 45 sites in total, see Table 3. Comparison of these positions to human Swiss-Prot Adam2 sequence (Q99965), we determine that 39/45 positively selected sites are

Table 5 Summary of the positively selected sites in the col1a1 gene, their clinical relevance, and, the probability of being located within distance "d" from the nearest disease-implicated site.

Positively selected sites	Posterior Probability	Human Variant: SNP position	Distance (*d*)	Probability of being *d* from nearest disease-implicated site	Genetic code distances between observed character states	Clinical Association	
195	0.926	197	2	0.04	A-N = 2	G → C	mild phenotype
280	0.588	275	5	0.26	A-S = 1; S-T = 1; T-A = 1	G → D	OI-II
478	0.959	476	2	0.128	A-S = 1; S-T = 1; T-A = 1	G → R	OI-II
784	0.968	776	8	0.396	A-S = 1; S-T = 1; T-A = 1	G → S	OI-II
1032	0.535	1025	7	0.364	A-P = 1	G → R	OI-II
1063	0.826	1061	2	0.128	N-S = 1	G → D	OI-II
		1061	2	0.032	N-S = 1	G → S	OI-IV
1149	0.623	1151	2	0.032	A-S = 1	G → S	OI-III
		1151	2	0.128	A-S = 1	G → V	OI-II
1194	0.675	1195	1	0.076	A-G = 1; G-S = 1; S-A = 1	G → C	OI-II mild form
1196	0.972				A-F = 2; F-Y = 1; Y-A = 2		
1316	0.928	1312	4	0.24	K-N = 1; N-P = 2; P-K = 2	W → C	OI-II
1456	0.997	1460	4	0.1	C-F = 1; C-L = 2; C-M = 2; F-L = 1; F-M = 2; L-M = 1	P → H	dbSNP: rs17853657

The sites under positive selection in the col1a1 protein and their associated posterior probabilities (PP) are shown. The third column shows variant positions (SNPs) as determined using Swiss-Prot human (PO2452) sequence. The fourth and fifth columns show the residue distance "d" of the positively selected site from its nearest genetic variant, and the probability of being located "d" residues from any disease implicated site by random chance alone. The sixth column uses single-letter amino acid symbols to show the genetic code distances between all observed character states at each positively selected site. "Clinical Association" show the replacement substitution at the human variant position and its clinical association with that human variant. OI = Osteolysis imperfecta, OI-I to -IV. The final entry for dbSNP is database entry number rs17853657 and as yet has not been associated with OI although it is in the same domain as the other disease-causing SNPs.

situated in the C-terminus region. On closer investigation of these sites we find that 12/45 positively selected sites occur in the disintegrin domain (position 384-473). The disintegrin domain has been shown to be involved in the binding of Adam2 to the oocyte [38]. A cysteine-rich domain occurs between (477-606), 16/45 positively selected sites fall in this region. It has been suggested for Adam12, (another member of the Adam family of proteins), that the cysteine-rich domain plays a role in mediating the cellular interactions via syndecans and integrin [39], a similar role for this domain in Adam2 can be postulated. Overall the results for Adam2 suggest a selective pressure for increased binding of Adam2 to the oocyte regardless of species of origin. For a complete set of results and LRTs for Adam2, see Additional File 6 (a).

Catsper1

Catsper1 is involved in regulating the calcium cation channel in sperm flagella, the result of which is movement of sperm [40]. Previous studies on Catsper1 exon 1 have been performed [41]. We intended to expand our analysis to span all exons and expand the data set to include a variety of *mammalia*. However, the exon 1 of non-primate *mammalia* is so highly variable that an accurate alignment cannot be constructed. The remaining exons were highly conserved across all species. We therefore split our catsper1 dataset into two sections each of which produced a good quality alignment for analysis, (1) exon1 of catsper1 for the primates, and (2) entire catsper1 gene for non-primate *mammalia*. ***(a) Catsper1 Exon 1 primates***Site-specific analysis of this protein identified 17% of the protein under positive selection with ω = 3.13. Previous analysis of this exon showed positive selection on indel substitutions in this gene [41]. The positively selected sites are situated throughout exon1, little is known about the functional significance of these sites. However, it is known that exon 1 has a significant role to play in altering the rate

of calcium ion channel inactivation. Different lengths in the N-terminus result in different rates of channel inactivation, where a long terminus results in a longer time to activation than the shorter terminus. This is described most effectively by the ball and chain mechanism described in [41]. See Additional File 6(b) for complete results. These results show the importance of this protein, and specifically the first exon, for reproductive success.

(b) Catsper1 entire gene non-primate mammalsOur site-specific analysis identified 16.7% of the sites under positive selection with an ω = 3.27, see Table 3. These sites all cluster in exon 1. While the rodent ancestor appears to be under positive selection with 4.47% of its sites evolving at ω = 999, see Additional File 6(c) for complete set of results. A previous study of 9 rodent species, including *Mus musculus* individuals from 4 different populations, has shown that within the rodent order there has been a continued pressure to evolve, with positive selection for indel substitutions in exon1 of the Catsper1 gene [43].

Semg2

A member of the family of semenogelin genes, Semg2 is involved in the formation of a postcopulatory plug [44]. Previously, positive selection has been reported for both site-specific and lineage-specific analysis for Semg2 [9,45]. We have expanded the data set from previous analyses to incorporate more species.

In our site-specific analysis, we found that 2.7% of our sites had an ω value of 12.26, see Table 3.

We have performed a novel functional analysis of these positively selected sites by comparing them to the human Semg2 sequence (Q02383) in the Swiss-Prot database. This is a step not previously taken by other studies of Semg2. A striking pattern emerged - all known domains of this protein have several positively selected sites. There is a probable glycosylation site at position 272, which is located close to a large stretch of positively selected sites (positions 262 to 289). It is so far unknown how significant this glycosylation site is in Semg2 and whether it plays a role in modifying the protein to form a copulatory plug. However, the results indicate that this protein, and in particular the region around the glycosylation site, has been under significant pressure to change.

A complete set of results for Semg2 is given in Additional File 6(h). The lineage-specific results are not described here in detail as lineage analyses have been carried out previously on the primate Semg2 gene [9,45]. It has been shown recently that the rate of evolution for this protein varies depending on the level of sperm competition [9]. Our results are in agreement with this finding, thus further verifying our approach.

Porimin

Two isoforms of this protein have been identified; we have focused on isoform 1 in the *mammalia*, as isoform 2 contains an additional human specific region between residues 34-52. To date the exact mechanisms of this transmembrane receptor are unknown. This protein is not well characterized biochemically and its function cannot be verified as reproduction related, therefore we only discuss the results briefly below.

On site-specific analysis of this protein we determined that 30 of the sites are under positive selection (ω = 12.22), see Table 3. From analysis of the sites on the Swiss-Prot entry for human Porimin (Q8N131), we could determine that two positively selected sites (146 and 147), were found in a highly conserved region and fall in close proximity to the N-linked glycosylation site. For complete set of results for Porimin, see Additional File 6(f).

Conclusion

Testing for phylogenetic signal and biases, such as amino acid composition bias and LBA, indicated that there was adequate phylogenetic signal for 10 of the genes and in general no evidence of systematic biases. On testing for LBA, Ph20 was the only protein in this dataset that displayed the typical signature of this bias with gene and species tree agreement being maximized with the removal of the fastest evolving categories. This would suggest that while germ line generation times vary greatly in the dataset, the effect of the resultant LBA does not impact on the sequence data to any great extent (1/11datasets).

Selective pressures for the reproductive proteins studied here are heterogeneous. All proteins exhibited regions of strong conservation proving the importance of maintaining structural stability and overall function in these proteins. All but 1 protein (Adam2) exhibited evidence of positive selection in specific lineages, and all proteins without exception exhibited positive selection in regions of catalytic/functional importance. For SP56 and Col1a1 the site-specific results are entirely novel. The lineage-specific results described here for Prkar2a and Catsper1 exon 1 in primates, are also novel. We have shown that, in the case of Catsper1, there is a fundamental protein functional shift between new world monkeys and old world monkeys. The Dn/Ds measurement applied here assumes that neutral substitution rate is akin to Ds, therefore no selection on silent sites. There have been many publications of late to the contrary therefore we are mindful of examining the rate of silent substitution in all our analyses [46,47].

For the reproductive genes in our dataset, we show that lineages evolve at unique rates and at functionally

crucial sites, specifically those involved in phosphorylation. We have also shown that a number of these proteins (Col1a1 and Catsper1) show positive selection for example in the ancestral rodent lineage and evidence of purifying selection in the subsequent divergent species.

Overall our analyses of these reproductive proteins show how important it is to carefully examine data for systematic biases prior to testing for lineage and/or site specific positive selection. We have also demonstrated the importance of including large numbers of taxa/lineages in these analyses. This finding was highlighted in our analysis of Prkar2a where previous analysis of this protein had included only 4 taxa and therefore reported a negative result. We do not observe any large-scale effect of germ line generation time in our dataset, with only 1 protein (Ph20) with evidence of long branch attraction. The results of Col1a1 indicate that the positively selected sites may have been of such importance for this protein that neighboring mutated sites may have been maintained in the population despite their propensity for causing disease. The location of positively selected sites determined using this approach and in regions of functional importance in the proteins in this dataset, provides us with further evidence of the link between functional shift and positive selection.

Methods

The data analyzed in this study consist of homologous reproductive genes from a variety of mammalian genomes. Genes were identified as being reproduction related from literature searches, analysis of protein interaction networks (iHOP) [48] and expression (microarray) data [11]. The microarray expression data used is from normal human tissues. We have also included a more in-depth analysis of previously identified cases of positive selection in reproductive proteins. A list of all data used in this study are available in Additional File 7, the total number of genes analyzed was 10. Homologs of all 10 reproduction related genes were identified in mammalian genomes that span the entire phylogeny of mammals, see Figure 1. For each of the reproduction related genes, the alignment of homologs contained between 10 and 18 species, and the alignment length varied between 351 and 4374 base pairs.

Sequence Data

Protein coding sequences for the reproductive proteins were retrieved by the combination of two methods; Ensembl and Blast searches. Orthologous coding sequences from all available completed mammalian genomes were retrieved from the Ensembl database [49]. These orthologs had been identified previously by performing a genome-wide reciprocal WUBlastp +SmithWaterman search of each gene across all completed genomes. To include those *mammalia* that were not present in Ensembl a BlastP search was conducted on all the human amino acid sequences from each gene against the Swiss-Prot database.

Mammalian Species

Primates: Human *(Homo sapiens)*, Chimp (*Pan troglodytes)*, Bonobo *(Pan paniscus)*, Bornean Orangutan *(Pongo pygmaeus)*, Sumatran Orangutan (*Pongo abelii*), *Gorilla (Gorilla gorilla)*, Rhesus Macaque *(Macaca mulatta)*, Crab eating *Macaque* (*Macaca fascicularis*), Pigtailed Macaque (*Macaca nemestrina)*, Bonnet monkey *(Macaca radiata)*, Baboon (*Papio hamadryas*), Mantled Guereza *(Colobus guereza)*, Vervet Monkey (*Cercopithecus aethiops*), Angolan Talapoin (*Miopithecus talapoin*), Squirrel Monkey (*Saimiri sciureus*), Cotton top tamarin (*Saguinus oedipus*), Common Marmoset (*Callithrix jacchus)*, Marmoset/Callithrix *(Callithrix-jacchus)*, Spider Monkey *(Ateles geoffroyi*), Bushbaby *(Otolemur garnettii*), Common woolly monkey (*Lagothrix lagotricha*), Ringtailed lemurs (*Lemur catta)*, Kloss Gibbon *(Hylobates klossii)*, Common/Lar Gibbon *(Hylobates lar)*, Night/owl Monkey *(Aotus trivirgatus boliviensis)*. Scandentia: Treeshrew *(Tupaia belangeri)*. Rodents: Mouse *(Mus musculus*), Rat *(Rattus norvegicus)*, Guinea pig *(Cavia porcellus)*, Ground Squirrel/Squirrel *(Spermophilus tridecemlineatus)*. Lagomorpha: Rabbit (*Oryctolagus cuniculus*), Pika (*Ochotona princes*). Eulipotyphila: Hedgehog *(Erinaceus europaeus)*, Shrew (*Sorex araneus*). Carnivores: Cat *(Felis catus)*, Dog *(Canis familiaris)*. Artiodactyla: Cow *(Bos taurus*), Pig *(Sus scrofa)*. Perisodactyla: Horse *(Equus caballus)*. Proboscidea: Elephant *(Loxodonta africana)*. Monotremata: Platypus *(Ornithorhynchus anatinus)*. Didelphimorphia: Opossum *(Monodelphis domestica)*.

Multiple Sequence Alignment (MSA)

All coding sequences were translated into their corresponding amino acid sequences using in-house translation software. Gene family alignments were generated at protein level using ClustalX 1.83.1 using default parameter settings [50]. The corresponding nucleotide gene family datasets were aligning based on their protein alignments using in-house software. Each gene family alignment was manually edited using Se-Al [51] to remove any ambiguous regions.

Nucleotide composition bias, amino acid composition bias and likelihood mapping tests

TreePuzzle 5.2 [15] performs a chi-square test that compares the amino acid composition of each

sequence to the frequency distribution assumed in the General Time Reversible (GTR) and Jones Taylor Thornton (JTT) models [52]. Ideally no species should fail this test, however, where two species fail and are thus drawn together on a tree, these sequences are excluded. Using the likelihood mapping method, each tree is disassembled into its constituent quartets and the support for each possible quartet is assessed. If the data contains phylogenetic signal then the likelihood of all three possible relationships for that quartet will be equally likely, these are represented by the three tips of the triangle, and the majority of the signal will be in these tip regions. Otherwise, the vertices and central region will be most heavily populated by supporting quartets.

Phylogeny Reconstruction

Phylogenetic trees were constructed using MrBayes v3.2.1 [53] and the amino acid sequences. Amino acid sequences were used in order to vitiate the effects of base and codon compositional biases. The substitution model was selected following model testing using Modelgenerator version 85 [54]. The selected model was JTT, the GTR rate model was implemented and the first 20000 trees for each gene were discarded as "burnin". A majority rule consensus tree from the remaining trees sampled was constructed for each gene. The parameter settings for each gene phylogeny are summarized in Additional File 8.

Site-stripping for significance

To test for long branch attraction (LBA) we applied the slow-fast approach of Brinkman and Phillipe [55]. We implemented the rate categorisation in a maximum likelihood framework in TreePuzzle 5.2 [15]. This software takes the alignment as input and generates *ab initio* phylogenetic trees. It then calculates the rate of mutation for each site in the alignment. The software specifies 8 arbitrary categories of site: each one of these categories contains some portion of the alignment. In this manuscript 8 is the most rapidly evolving (for example every lineage has a different character state for that character), and category 1 is the most slowly evolving (for example each lineage has the same/identical character state for that character). Sites are then progressively removed from the protein MSA according to their evolutionary rate, and at each stage a new phylogenetic tree is constructed based on this slightly reduced dataset. The difference between the new topology created on a reduced alignment and the original topology reconstructed based on the entire alignment are then compared in a statistical framework to determine which fits the data best (SH Test 2, see below) or which is most similar to the species phylogeny (RMSD Test 1, see below). At each stage we employ MrBayes [56] to perform the phylogenetic reconstruction using the aforementioned settings.

Tests of the difference between two trees

Test 1: Nodal distance calculation

TOPD/FMTS v 3.3 [18] calculates the distance between the site-stripped trees and the 'ideal' tree. The 'ideal' tree used for each gene was a pruned version of the canonical species tree as seen in Figure 1. A distance matrix is derived by counting the number of nodes that separate each of the taxa in a tree. A distance matrix is calculated for each site-stripped tree as compared to the ideal species tree. The nodal distance score is obtained by calculating the RMSD of the matrices. If both trees are identical the RMSD value would be 0, indicating no distance between them. This figure increases the more distance there is between the two trees.

Test 2: Shimodaira-Hasegawa (SH) statistical test of two trees

For each gene MSA, complete and site-stripped, a comparison of the likelihood of the estimated Bayesian phylogeny for that alignment with the likelihood of its corresponding 'ideal' species tree was carried out using the SH test [14] implemented in TreePuzzle 5.2 [15] to determine which tree was significantly the best-fit tree for the alignment.

Selective Pressure Analysis

PAML 4.3 [57,58] uses a ML method of calculating ω for site-specific and lineage-site specific changes. Codeml, part of the PAML 4.3 package [57,58], applies a series of models to our data, with each model differing from the previous with the addition of more complex parameters. The simplest model is M0, and it calculates an ω value over the entire alignment. This model assumes that all sites and all lineages are evolving at the same rate. Model M3 is an extension of M0 and allows all ω values to vary freely. There are two variations of the M3 model, m3(k = 2) discrete which allows two variable classes of sites and m3(k = 3) which allows three classes of site. M1 is a neutral model that allows two parameter estimates for proportion of sites where ω = 0 or ω = 1. M2 is the selection model, it allows three parameters where ω = 0 or ω = 1 or ω is estimated and free to be greater than 1. M7, is the beta model, it allows ten different site classes for ω between 0 and 1. M7 is compared against the more parameter rich M8 (beta &omega >1). M8 allows 10 different site classes but contains an additional parameter whereby the 11^{th} ω is free to vary between 0 and >1. M8a(beta &omega = 1) is null hypothesis of model 8. Model A & Model B are models that allow

testing of ω variation in lineage-site analyses. Model A is an extension of M1 and Model B is a more parameter rich extension of m3(k = 2). We have also implemented model A null which is denoted as modelA1 elsewhere. Model A null is compared to model A in an LRT as per Additional File 9. Only statistically significant models for the data are taken into account. Statistically significant results were decided by calculating the difference in log likelihood or, lnL, scores between models and their more parameter rich extensions in a likelihood ratio test (LRT) as described previously in [17,58]. If the likelihood score was exceeded the critical χ^2 values, then the result was significant. See Additional File 9 for full set of LRTs performed.

In silico analysis of positively selected sites

Sites under positive selection (ω > 1) were estimated using the empirical Bayes methods in the site-specific and lineage specific analysis performed. The methods used were naúve empirical Bayes (NEB) and Bayes empirical Bayes (BEB) [58]. Swiss-Prot is a protein sequence database that provides description of the function of a protein, the domain structures, post-translational modifications and variants. Significant sites, verified through close examination of the MSAs and codeml output using alignment visualisation software Se-AL [51], were compared with unaligned human amino acid sequence taken from Swiss-Prot. These sites were examined to see whether or not they lay in catalytically important regions of the protein.

Additional file 1: Additional Table 1 - Results of amino acid composition bias per gene. Results of the amino acid composition bias test and shown here on a per gene basis. We would expect that if two species have similarly and significantly (P < 0.05) biased amino acid composition that they would be drawn together on the phylogeny. Those with P < 0.05 scores are highlighted but are dispersed throughout different genes. The frequency distribution assumed in the maximum likelihood model calculated by Tree-Puzzle (5% chi-square p-values) was used. N/A = species not represented in the gene dataset.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2148-10-39-S1.DOC]

Additional file 2: Additional Table 2 - Results of base composition bias per gene. Results of the base composition bias test and shown here on a per gene basis. We would expect that if two species have similarly and significantly (P < 0.05) biased base composition that they would be drawn together on the phylogeny. Those with P < 0.05 scores are highlighted but are dispersed throughout different genes. The frequency distribution assumed in the maximum likelihood model calculated by Tree-Puzzle (5% chi-square p-values) was used. N/A = species not represented in the gene dataset.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2148-10-39-S2.DOC]

Additional file 3: Additional Table 3 - Results of likelihood mapping test for phylogenetic support and conflict estimated for each gene. Results of Likelihood mapping test are shown here on a gene-by-gene basis. This table summarizes the amount of phylogenetic signal and conflict in each alignment. The three possible topologies for each quartet of species are represented by the corners of the triangle, these corners represent strong support for phylogenetic signal. Quartets present on the vertices represent incongruence in the phylogenetic signal. Quartets at the centre of the triangle represents those quartets where all three topologies are equally likely, i.e. phylogenetic signal completely lacking. Each gene is subsequently given a category based on the quality of the data, only categories 1 and 2 were used.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2148-10-39-S3.DOC]

Additional file 4: Additional Table 4 - Results of root mean squared deviation (RMSD) analysis for comparing binary trees. This table summarizes the results of comparing the site stripped phylogenies with the ideal species phylogeny. In the first column is the gene name. Each of the subsequent columns represents a category of site variation that is removed (1 is the slowest evolving, 8 the most rapid). The values given for each category removed is the RMSD statistic and represents how similar the resultant site stripped topology is to the canonical species phylogeny. NB - non-binary tree, N/A - not applicable (site category not estimated for alignment).
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2148-10-39-S4.DOC]

Additional file 5: Additional Table 5 - Results of the SH test for site-stripped gene versus ideal species phylogeny. This table summarizes the results of comparing the site stripped phylogenies with the ideal species phylogeny using the SH test, this is a more statistically robust approach and more suited to multi-furcating topologies such as those in the dataset. Each of the rows represents a category of site variation that is removed. For each site stripped site dataset the resultant gene tree is compared to the species phylogeny. The values given for each category removed denotes whether there is a significant difference between the site stripped tree and the species phylogeny, values of less than 0.05 represent those cases where there is a significant difference between the phylogenies. NS = No Statistical significance between gene and species tree, the species tree was taken in these cases.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2148-10-39-S5.DOC]

Additional file 6: Additional Table 6(a-k) - Complete results of Maximum likelihood analysis for selective pressure variation per gene. For each gene analyzed (a-k) the results are shown in full on a gene-by-gene basis (in alphabetical order). The layout of each table is identical for each gene. The corresponding LRTs performed and all scores and values computed are shown below. The models used are given in the left-most column (Model), followed by the number of parameters associated with that model (P). The Log Likelihood or each model is given in the column (L), and the estimates of the parameters for the proportion of sites (p) and the ratio of Dn/Ds (ω) are given. Sites identified by each model as being positively selected are shown in the final column.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2148-10-39-S6.DOC]

Additional file 7: Additional Table 7 - Summary of data used in the analysis. Species names, unique identifiers and sequence lengths are given for all data. Summary of data used in the analysis. Species names, unique identifiers for Ensembl (ENS) or Swiss-Prot and database versions are given. The sequence length per species are given for all genes.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2148-10-39-S7.DOC]

Additional file 8: Additional Table 8 - Parameters for Phylogeny Reconstruction per gene. The parameters used to reconstruct each gene tree in MrBayes are shown. The model of rate heterogeneity for each gene is shown, along with the number of generations required, and the number of markov chains (these values vary based on the size of the dataset).
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2148-10-39-S8.DOC]

Additional file 9: Additional Table 9 - Likelihood ratio tests (LRTs) performed using all evolutionary models used in selection analysis. Details on all likelihood ratio tests performed in the analysis. The models are denoted by their abbreviated names, Model A1 is denoted as Model A null throughout the manuscript. The number of degrees of freedom (df) are shown, this is relevant for the chi-squared test for significance, the critical values in each instance are given in the final column.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2148-10-39-S9.DOC]

Abbreviations

A.A.: Amino Acid; Bck: Background lineage/s; BEB: Bayes Empirical Bayes; CDS: Coding DNA sequence; Dn: Non-synonymous substitution per non-synonymous site; Ds: Synonymous substitution per synonymous site; F: Frequency of amino acids; Fwd: Foreground lineage/s; G: gamma distributed sites rates across sites; GTR: General Time Reversible; I: invariable; JTT: Jones, Taylor and Thornton; LBA: Long Branch Attraction; LM: Likelihood mapping; LRT: Likelihood Ratio Test; ML: Maximum Likelihood; MSA: Multiple Sequence Alignment; N/A: data not available; NB: Non-binary tree; NEB: Naïve Empirical Bayes; NS: No statistical difference; OI: Osteogenesis imperfecta; OI-II/-III/-IV: Osteogenesis imperfecta type -2/-3/-4; P: probability; PP: Posterior Probability; RMSD: Root Mean Squared Deviation; SH: Shimodaira Hasegawa; SNP: Single nucleotide polymorphism.

Acknowledgements

We would like to thank the Irish Research Council for Science, Engineering and Technology (Embark Initiative Postgraduate Scholarship to NBL and CCM) for financial support and DCU School of Biotechnology scholarship (for TAW). We would like to thank the SFI/HEA Irish Centre for High-End Computing (ICHEC) for processor time and technical support for both phylogeny reconstruction and selection analysis. We would like to thank the SCI-SYM centre for processor time.

Authors' contributions

CCM carried out all data assembly, including searches of (i) literature, (ii) microarray studies, and (iii) protein interaction databases. CCM carried out all homolog identification and MSAs. NBL and CCM carried out all data quality and phylogeny analyses. TAW designed and performed randomization tests, designed bespoke software for the analyses and contributed to the preparation of the manuscript. CCM, NBL and MJO'C carried out all selective pressure analyses. NBL and CCM participated in drafting the manuscript. AJH analysed reproductive age data and genestational times for all mammals in the study, and helped to draft the manuscript. MJO'C conceived of the study, its design and coordination and drafted the manuscript. All authors read and approved the final draft.

Received: 20 July 2009
Accepted: 11 February 2010 Published: 11 February 2010

References

1. Aagaard JE, *et al*: **Rapidly evolving zona pellucida domain proteins are a major component of the vitelline envelope of abalone eggs.** *Proceedings of the National Academy of Sciences of the United States of America* 2006, **103(46)**:1730-17307.
2. Wyckoff GJ, Wang W, Wu CI: **Rapid evolution of male reproductive genes in the descent of man.** *Nature* 2000, **403(6767)**:304-309.
3. McInerney JO: **The causes of protein evolutionary rate variation.** *Trends Ecol Evol* 2006, **21(5)**:230-2.
4. Zhou T, Drummond DA, Wilke CO: **Contact density affects protein evolutionary rate from bacteria to animals.** *J Mol Evol* 2008, **66(4)**:395-404.
5. Li WH, *et al*: **Rates of nucleotide substitution in primates and rodents and the generation-time effect hypothesis.** *Mol Phylogenet Evol* 1996, **5(1)**:182-7.
6. Gaut BS, *et al*: **Relative rates of nucleotide substitution at the rbcL locus of monocotyledonous plants.** *J Mol Evol* 1992, **35(4)**:292-303.
7. Ohta T: **An examination of the generation-time effect on molecular evolution.** *Proc Natl Acad Sci USA* 1993, **90(22)**:10676-80.
8. Swanson WJ, Vacquier VD: **The rapid evolution of reproductive proteins.** *Nature reviews Genetics* 2002, **3(2)**:137-144.
9. Dorus S, *et al*: **Rate of molecular evolution of the seminal protein gene SEMG2 correlates with levels of female promiscuity.** *Nature genetics* 2004, **36(12)**:1326-1329.
10. Anisimova M, Bielawski JP, Yang Z: **Accuracy and power of bayes prediction of amino acid sites under positive selection.** *Molecular biology and evolution* 2002, **19(6)**:950-958.
11. Shyamsundar R, *et al*: **A DNA microarray survey of gene expression in normal human tissues.** *Genome biology* 2005, **6(3)**:R22.
12. He Z, *et al*: **Expression of Col1a1, Col1a2 and procollagen I in germ cells of immature and adult mouse testis.** *Reproduction* 2005, **130(3)**:333-41.
13. Murphy WJ, *et al*: **Resolution of the early placental mammal radiation using Bayesian phylogenetics.** *Science* 2001, **294(5550)**:2348-51.
14. Shimodaira H, Hasegawa M: **CONSEL: for assessing the confidence of phylogenetic tree selection.** *Bioinformatics* 2001, **17(12)**:1246-7.
15. Schmidt HA, *et al*: **TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing.** *Bioinformatics (Oxford, England)* 2002, **18(3)**:502-504.
16. Strimmer K, von Haeseler A: **Likelihood-mapping: a simple method to visualize phylogenetic content of a sequence alignment.** *Proceedings of the National Academy of Sciences of the United States of America* 1997, **94(13)**:6815-6819.
17. Loughran NB, *et al*: **The phylogeny of the mammalian heme peroxidases and the evolution of their diverse functions.** *BMC evolutionary biology* 2008, **8**:101.
18. Puigbo P, Garcia-Vallve S, McInerney JO: **TOPD/FMTS: a new software to compare phylogenetic trees.** *Bioinformatics (Oxford, England)* 2007, **23(12)**:1556-1558.
19. Behera MA, *et al*: **Thrombospondin-1 and thrombospondin-2 mRNA and TSP-1 and TSP-2 protein expression in uterine fibroids and correlation to the genes COL1A1 and COL3A1 and to the collagen cross-link hydroxyproline.** *Reproductive sciences (Thousand Oaks, Calif)* 2007, **14(8 Suppl)**:63-76.
20. Marini JC, *et al*: **Consortium for osteogenesis imperfecta mutations in the helical domain of type I collagen: regions rich in lethal mutations align with collagen binding sites for integrins and proteoglycans.** *Human mutation* 2007, **28(3)**:209-221.
21. Oyen O, *et al*: **Human testis cDNA for the regulatory subunit RII alpha of cAMP-dependent protein kinase encodes an alternate amino-terminal region.** *FEBS letters* 1989, **246(1-2)**:57-64.
22. Leclerc P, de Lamirande E, Gagnon C: **Cyclic adenosine 3',5'monophosphate-dependent regulation of protein tyrosine phosphorylation in relation to human sperm capacitation and motility.** *Biology of reproduction* 1996, **55(3)**:684-692.
23. Hunnicutt GR, Primakoff P, Myles DG: **Sperm surface protein PH-20 is bifunctional: one activity is a hyaluronidase and a second, distinct activity is required in secondary sperm-zona binding.** *Biology of reproduction* 1996, **55(1)**:80-86.
24. Swanson WJ, Nielsen R, Yang Q: **Pervasive adaptive evolution in mammalian fertilization proteins.** *Molecular biology and evolution* 2003, **20(1)**:18-20.
25. Arming S, *et al*: **In vitro mutagenesis of PH-20 hyaluronidase from human sperm.** *European journal of biochemistry/FEBS* 1997, **247(3)**:810-814.
26. Phelps BM, Myles DG: **The guinea pig sperm plasma membrane protein, PH-20, reaches the surface via two transport pathways and becomes localized to a domain after an initial uniform distribution.** *Developmental biology* 1987, **123(1)**:63-72.
27. Bleil JD, Wassarman PM: **Identification of a ZP3-binding protein on acrosome-intact mouse sperm by photoaffinity crosslinking.** *Proceedings of the National Academy of Sciences of the United States of America* 1990, **87(14)**:5563-5567.

28. Kim KS, Cha MC, Gerton GL: **Mouse sperm protein sp56 is a component of the acrosomal matrix.** *Biology of reproduction* 2001, **64(1)**:36-43.
29. Sherry ST, Ward M, Sirotkin K: **dbSNP-database for single nucleotide polymorphisms and other classes of minor genetic variation.** *Genome Res* 1999, **9(8)**:677-9.
30. Gupta SK, *et al*: **Structural and functional attributes of zona pellucida glycoproteins.** *Society of Reproduction and Fertility supplement* 2007, **63**:203-216.
31. Swanson WJ, *et al*: **Positive Darwinian selection drives the evolution of several female reproductive proteins in mammals.** *Proceedings of the National Academy of Sciences of the United States of America* 2001, **98(5)**:2509-2514.
32. Chakravarty S, *et al*: **Relevance of glycosylation of human zona pellucida glycoproteins for their binding to capacitated human spermatozoa and subsequent induction of acrosomal exocytosis.** *Molecular reproduction and development* 2008, **75(1)**:75-88.
33. Shabanowitz RB, O'Rand MG: **Characterization of the human zona pellucida from fertilized and unfertilized eggs.** *Journal of reproduction and fertility* 1988, **82(1)**:151-161.
34. Wassarman PM: **Mammalian fertilization: molecular aspects of gamete adhesion, exocytosis, and fusion.** *Cell* 1999, **96(2)**:175-183.
35. Patrat C, *et al*: **Zona pellucida from fertilised human oocytes induces a voltage-dependent calcium influx and the acrosome reaction in spermatozoa, but cannot be penetrated by sperm.** *BMC developmental biology* 2006, **6**:59.
36. Primakoff P, Hyatt H, Tredick-Kline J: **Identification and purification of a sperm surface protein with a potential role in sperm-egg membrane fusion.** *The Journal of cell biology* 1987, **104(1)**:141-149.
37. Evans JP: **The molecular basis of sperm-oocyte membrane interactions during mammalian fertilization.** *Human reproduction update* 2002, **8(4)**:297-311.
38. Wong GE, *et al*: **Analysis of fertilin alpha (ADAM1)-mediated sperm-egg cell adhesion during fertilization and identification of an adhesion-mediating sequence in the disintegrin-like domain.** *The Journal of biological chemistry* 2001, **276(27)**:24937-24945.
39. Iba K, *et al*: **The cysteine-rich domain of human ADAM 12 supports cell adhesion through syndecans and triggers signaling events that lead to beta1 integrin-dependent cell spreading.** *The Journal of cell biology* 2000, **149(5)**:1143-1156.
40. Carlson AE, *et al*: **CatSper1 required for evoked Ca2+ entry and control of flagellar function in sperm.** *Proceedings of the National Academy of Sciences of the United States of America* 2003, **100(25)**:14864-14868.
41. Podlaha O, Zhang J: **Positive selection on protein-length in the evolution of a primate sperm ion channel.** *Proceedings of the National Academy of Sciences of the United States of America* 2003, **100(21)**:12241-12246.
42. Avenarius MR, *et al*: **Human male infertility caused by mutations in the CATSPER1 channel protein.** *American Journal of Human Genetics* 2009, **84(4)**:505-510.
43. Podlaha O, *et al*: **Positive selection for indel substitutions in the rodent sperm protein catsper1.** *Molecular biology and evolution* 2005, **22(9)**:1845-1852.
44. Peter A, *et al*: **Semenogelin I and semenogelin II, the major gel-forming proteins in human semen, are substrates for transglutaminase.** *European journal of biochemistry/FEBS* 1998, **252(2)**:216-221.
45. Hurle B, *et al*: **Comparative sequence analyses reveal rapid and divergent evolutionary changes of the WFDC locus in the primate lineage.** *Genome research* 2007, **17(3)**:276-286.
46. Chamary JV, Hurst LD: **The price of silent mutations.** *Sci Am* 2009, **300(6)**:46-53.
47. Hurst LD, Pal C: **Evidence for purifying selection acting on silent sites in BRCA1.** *Trends Genet* 2001, **17(2)**:62-5.
48. iHOP: **The iHOP database.**.
49. Ensembl: **Ensembl.**http://www.ensembl.org, cited.
50. Chenna R, *et al*: **Multiple sequence alignment with the Clustal series of programs.** *Nucleic acids research* 2003, **31(13)**:3497-3500.
51. Rambaut A: **Se-AL Sequence alignment editor.** *Oxford* 1996.
52. Lanave C, *et al*: **A new method for calculating evolutionary substitution rates.** *J Mol Evol* 1984, **20(1)**:86-93.
53. Ronquist F, Huelsenbeck JP: **MrBayes 3: Bayesian phylogenetic inference under mixed models.** *Bioinformatics (Oxford, England)* 2003, **19(12)**:1572-1574.
54. Keane TM, *et al*: **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* 2006, **6**:29.
55. Brinkmann H, Philippe H: **Archaea sister group of Bacteria? Indications from tree reconstruction artifacts in ancient phylogenies.** *Mol Biol Evol* 1999, **16(6)**:817-25.
56. Ronquist F, Huelsenbeck JP: **MrBayes 3: Bayesian phylogenetic inference under mixed models.** *Bioinformatics* 2003, **19(12)**:1572-4.
57. Yang Z: **PAML: a program package for phylogenetic analysis by maximum likelihood.** *Computer applications in the biosciences: CABIOS* 1997, **13(5)**:555-556.
58. Yang ZW, Wong S, Nielsen R: **Bayes empirical bayes inference of amino acid sites under positive selection.** *Molecular biology and evolution* 2005, **22(4)**:1107-1118.

doi:10.1186/1471-2148-10-39
Cite this article as: Morgan *et al.*: Positive selection neighboring functionally essential sites and disease-implicated regions of mammalian reproductive proteins. *BMC Evolutionary Biology* 2010 **10**:39.

BMC Evolutionary Biology

BioMed Central
The Open Access Publisher

Colon cancer associated genes exhibit signatures of positive selection at functionally significant positions

BMC Evolutionary Biology 2012, **12**:114 doi:10.1186/1471-2148-12-114

Claire C Morgan (claire.morgan6@mail.dcu.ie)
Kabita Shakya (kabita.shakya1@gmail.com)
Andrew Webb (andrew.webb5@mail.dcu.ie)
Thomas A Walsh (thomas.walsh26@mail.dcu.ie)
Mark Lynch (mark.lynch10@gmail.com)
Christine E Loscher (Christine.Loscher@dcu.ie)
Heather J Ruskin (hruskin@computing.dcu.ie)
Mary J O'Connell (mary.oconnell@dcu.ie)

ISSN 1471-2148

Article type Research article

Submission date 1 February 2012

Acceptance date 22 June 2012

Publication date 12 July 2012

Article URL http://www.biomedcentral.com/1471-2148/12/114

Colon cancer associated genes exhibit signatures of positive selection at functionally significant positions

Claire C Morgan[1,2,†]
Email: claire.morgan6@mail.dcu.ie

Kabita Shakya[1,2,3,†]
Email: kabita.shakya1@gmail.com

Andrew Webb[1,2]
Email: andrew.webb5@mail.dcu.ie

Thomas A Walsh[1,2]
Email: thomas.walsh26@mail.dcu.ie

Mark Lynch[1,2,4]
Email: mark.lynch10@gmail.com

Christine E Loscher[4]
Email: Christine.Loscher@dcu.ie

Heather J Ruskin[2,3]
Email: hruskin@computing.dcu.ie

Mary J O'Connell[1,2,*]
Email: mary.oconnell@dcu.ie

[1] Bioinformatics and Molecular Evolution Group, School of Biotechnology, Dublin City University, Glasnevin, Dublin 9, Ireland

[2] Centre for Scientific Computing & Complex Systems Modelling (SCI-SYM), Dublin City University, Glasnevin, Dublin 9, Ireland

[3] School of Computing, Dublin City University, Glasnevin, Dublin 9, Ireland

[4] Immunomodulatory Research Group, School of Biotechnology, Dublin City University, Glasnevin, Dublin 9, Ireland

[*] Corresponding author. Centre for Scientific Computing & Complex Systems Modelling (SCI-SYM), Dublin City University, Glasnevin, Dublin 9, Ireland

[†] Equal contributors.

Abstract

Background

Cancer, much like most human disease, is routinely studied by utilizing model organisms. Of these model organisms, mice are often dominant. However, our assumptions of functional equivalence fail to consider the opportunity for divergence conferred by ~180 Million Years (MY) of independent evolution between these species. For a given set of human disease related genes, it is therefore important to determine if functional equivalency has been retained between species. In this study we test the hypothesis that cancer associated genes have different patterns of substitution akin to adaptive evolution in different mammal lineages.

Results

Our analysis of the current literature and colon cancer databases identified 22 genes exhibiting colon cancer associated germline mutations. We identified orthologs for these 22 genes across a set of high coverage (>6X) vertebrate genomes. Analysis of these orthologous datasets revealed significant levels of positive selection. Evidence of lineage-specific positive selection was identified in 14 genes in both ancestral and extant lineages. Lineage-specific positive selection was detected in the ancestral Euarchontoglires and Hominidae lineages for STK11, in the ancestral primate lineage for CDH1, in the ancestral Murinae lineage for both SDHC and MSH6 genes and the ancestral Muridae lineage for TSC1.

Conclusion

Identifying positive selection in the primate, Hominidae, Muridae and Murinae lineages suggest an ancestral functional shift in these genes between the rodent and primate lineages. Analyses such as this, combining evolutionary theory and predictions - along with medically relevant data, can thus provide us with important clues for modeling human diseases.

Keywords

Positive selection, Colon cancer, Adaptive evolution, Protein functional shift, Selective pressure, Evolutionary medicine

Background

Mouse models are currently used to research many human cancers including colon cancer. On a genome wide scale, mouse protein sequences share 78.5% sequence identity with human counterparts [1]. With such high levels of sequence identity it may seem reasonable to expect that many orthologs between mouse and human would have conserved functions. However, in the ~180 Million Years (MY) of independent evolution [2], it is possible that certain proteins have functionally diverged. One example of ortholog divergence between human and mouse is the *TDP1* gene, required in Topo1-DNA complex repair, protein sequence similarity of 81%. A point mutation from an adenine to a guanine at position 1478 in human *TDP1* is linked with a disorder known as SCAN1 that results in cerebellar atrophy and peripheral neuropathy. However, this mutation in mouse does not result in the same

condition/phenotype [3]. Specific mutations in any of the following genes in human result in disease: BCL10, PKLR and SGCA, but the same mutations in the mouse homologs do not result in phenotypic change to a disease state [4]. BRCA1 is heavily implicated in breast cancer in humans, with $BRCA1^{+/-}$ women having a 50% risk of developing breast cancer, while $BRCA^{+/-}$ mice do not exhibit increased susceptibility to cancer [5]. These observed differences in phenotype could potentially be the result of protein functional shifts in cancer-associated genes between human and mouse. While the analysis of the mouse lineage versus human is important from an evolutionary medicine perspective to determine/predict those specific cases where mouse may not effectively model the human disease phenotype, the analysis of all other lineages frames these results in the context of all mammals. Therefore, in this study we have not only examined the human and mouse lineages but all lineages leading to extant species in our dataset. This allows us to gain a greater understanding of the level of lineage-specific functional shift that has occurred in colon cancer associated genes.

Positive selection is the retention and spread of advantageous mutations throughout a population and has long been considered synonymous with protein functional shift. There are a number of driving forces for positive selection including external mechanisms such as adaptation to different ecological niches and response to disease and internal mechanisms such as co-evolution and compensatory mutations [6], all of which are relevant to the data and species we are analyzing. At the molecular level, the ratio of nonsynonymous substitutions per nonsynonymous site (*dN*) to synonymous substitutions per synonymous site (*dS*) is known as ω, and indicates the selective pressure at work in that sequence. If $\omega > 1$ it signifies positive selective pressure, $\omega = 1$ signifies neutral evolution, while $\omega < 1$ indicates purifying selective pressure. Previous work assessed the level of positive selection present in mammal genomes and estimated 5%-9% of genes in mammals are under positive selection under a Bayesian framework, and thus provides us with a reference or expected level of positive selection for our analysis [7,8].

Here we have applied a Maximum Likelihood method based on codon models of evolution to assess the selective pressures across our dataset [9]. These methods are far more robust than alternatives such as the sliding window approach [10], nonetheless they do suffer from limitations and have strict criteria in terms of dataset size for statistical robustness [11,12]. Another feature of sequence evolution that can negatively impact on a selective pressure analysis is recombination [13]. To evaluate the robustness of the Likelihood Ratio Tests (LRTs), simulations have been performed that show that type 1 error rates can be up to 90% with relatively high rates of recombination in protein coding sequences resulting in the misinterpretation of recombination as positive selection ([13] We have incorporated a test for recombination for all genes in the dataset prior to the ML selective pressure analysis. Recent studies using these codon models of evoluton in an ML framework have combined evolutionary predictions of positive selection with biochemical verification of functional affects of these substitutions [14-16], and thus support the link between positive selection and protein functional shift.

We have taken colon cancer as an example for our study given the large amount of mutation and epigenetic data available for this form of cancer [17]. Lineage-specific positive selection in genes associated with colon cancer is strongly suggestive of functional shift and could have serious implications in the use of certain lineages for modeling colon cancer.

Colorectal cancer (CRC) is the third most commonly diagnosed cancer in males and second in females and we have focused on this in our study [18]. CRC arises through the

accumulation of multiple genetic and epigenetic changes. Genetic changes consist of both somatic and germline (i.e. heritable) mutations, the genes in which there are germline mutations that are highly associated with the development of colon cancer are analyzed here (22 genes in total) and are referred to throughout this manuscript as "colon cancer associated genes". Colon cancer associated genes work in conjunction with other proteins and pathways and can be thought of as contributing to, rather than being the single cause of colon cancer (note: these genes also have other functions outside of their association that may contribute to selective pressure variation in different lineages). Epigenetic changes such as hypermethylation of certain genes, loss of imprinting and acetylation/phosphorylation/methylation of particular histones are also implicated in cancer. Detailed information on colon cancer epigenetics have been made available to the community through the StatEpigen biomedical resource [17]. Other events such as loss of heterozygosity, microsatellite instability and CpG island methylator phenotype can also play an important role.

Hereditary Non-Polyposis Colorectal Cancer (HNPCC) is a hereditary predisposition for the development of colorectal cancer, and accounts for 3% of all colon cancer cases [19]. The 22 genes we have analyzed were selected based on the presence of known germline mutations associated with colon cancer. What follows is a brief description of each gene in the study. The genes linked with HNPCC are: MLH1, PMS2, MSH2, MSH6, and PMS1, all of which are members of the MMR DNA repair pathway [19].

MLH1 (mutL homologue 1) is a mismatch repair gene and is commonly associated with HNPCC. Missense mutations in MLH1 occur in the C-terminal domain, which is responsible for constitutive dimerization with the mismatch repair endonuclease PMS2 [20]. Studies have also shown that microsatellite instability (MSI) is the molecular fingerprint of a deficient mismatch repair system. It is estimated that some 15% of colorectal cancers display MSI owing to the epigenetic silencing of MLH1, and/or germline mutation in any one of the following mismatch repair genes: PMS2, MLH1, MSH2, and MSH6 [21]. The mismatch repair endonuclease PMS2 is known to interact with MLH1 and is a component of the post-replicative DNA mismatch repair system (MMR). PMS2 is recruited to cleave damaged DNA this recruitment is triggered by the binding of MSH2 and MSH6 proteins to dsDNA mismatches followed by the recruitment of MLH1 (Figure 1). PMS1 is also involved in the repair of DNA mismatches, and it can form heterodimers with MLH1. Additional genes in our study include the tumor suppressor gene TP53, CDH1, MUTYH, and APC. TP53 is a hub protein in the cellular DNA damage response pathway known as the P53 signaling pathway [22], it is linked with colorectal cancer and many other cancers. The genes CDH1, MUTYH, and APC also interact with one another in addition to their involvement in the MMR pathway described above. For example, CDH1 and APC interact together as an ubiquitin ligase, which is involved in glycolysis regulation during the cell cycle [23]. In fact, most of the colon cancer associated genes in this study can be grouped into critical pathways, such as apoptosis, DNA damage control, and cell cycle signaling [24].

Figure 1 Phylogeny of animal species used in this study. The ancestral lineages tested in the analysis are labeled with their corresponding names as used throughout the text. Those lineages where positive selection was detected are labeled with filled circles, no evidence of positive selection is denoted with an empty circle

To assess if there is evidence for protein functional shift from the patterns of substitution in colon cancer associated genes we have carried out a selective pressure analysis using codon models of lineage-specific rate heterogeneity.

Methods

Sequence data assembled

The colon cancer gene dataset used in this study consists of 22 genes taken from the Cancer Gene Census at the Sanger Institute [25]. All 22 genes have reported cases of germline mutations that are associated with colon cancer (See Table 1 for summary of data, detail on the complete dataset is available in Additional file 1). Using the Compara data from Ensembl [26,27], single gene orthologs were identified for each gene across the vertebrate genomes chosen. The 21 species were selected based on the genome coverage. These included representatives from 3 of the 4 main lineages of Eutheria, namely Afrotheria, Euarchontoglires, and Laurasiatheria, as well as outgroup species such as platypus, zebrafish, and zebra finch (see Additional file 1).

Table 1 Colon Cancer Gene Set analyzed in this study

Gene (HGNC code)	Ensembl Identifier	Taxa Number[2]	Alignment Length[3]	Syndrome	Tumor Types Observed	Pathway(s)	References[4]
APC	ENSG00000134982	20	9177	Familial adenomatous polyposis (FAP)	Colon, thyroid, stomach, intestine	APC	[19,24]
ATM	ENSG00000149311	18	9189	Ataxia telangiectasia (A-T)	Leukemia, lymphoma, colorectal	CIN	[[24], http://statepigen.sci-sym.dcu.ie/]
BHD	ENSG00000154803	20	1737	Birt-Hogg-Dube syndrome	Renal, colon	AMPK, mTOR, STAT	[24]
BMPR1A	ENSG00000107779	19	1596	Juvenile polyposis	Gastrointestinal	SMAD	[24]
CDH1	ENSG00000039068	15	2649	Familial gastric carcinoma	Stomach	AP	[[24,28] (E-cadherin)]
MADH4	ENSG00000141646	16	1656	Juvenile polyposis	Gastrointestinal	SMAD	[[24] (SMAD4)]
MET	ENSG00000105976	21	4146	Hereditary papillary renal cell carcinoma (HPRCC)	Kidney, colorectal	RAS, PI3K, STAT, Beta-catenin, Notch	[24]
MLH1	ENSG00000076242	19	2274	Hereditary non-polyposis colon cancer (HNPCC)	Colon, uterus	MMR	[24]
MSH2	ENSG00000095002	18	2802	Hereditary non-polyposis colon cancer (HNPCC)	Colon, uterus	MMR	[24]
MSH6	ENSG00000116062	19	4101	Hereditary non-polyposis colon cancer (HNPCC)	Colon, uterus	MMR	[24]
MUTYH	ENSG00000132781	21	1569	Attenuated Polyposis	Colon	BER	[24]
NF1	ENSG00000196712	17	8523	Neurofibromatosis type I	Neurofibroma, colon	RTK	[24]
PMS1	ENSG00000064933	20	2799	Hereditary non-polyposis colon cancer (HNPCC)	Colon, uterus	MMR	
PMS2	ENSG00000122512	21	2592	Hereditary non-polyposis colon cancer (HNPCC)	Colon, uterus	MMR	[24]

PTEN	ENSG00000171862	18	1209	Cowden syndrome	Hamartoma, glioma, colorectum	PI3K	[[24], http://statepigen.sci-sym.dcu.ie/]
SDHB	ENSG00000117118	18	840	Hereditary paraganglioma, Carney–Stratakis	Paragangliomas, pheochromocytomas, gastrointestinal	HIF1	[24]
SDHC	ENSG00000143252	16	507	Hereditary paraganglioma, Carney–Stratakis	Paragangliomas, pheochromocytomas, gastrointestinal	HIF1	[24]
STK11	ENSG00000118046	18	1320	Peutz-Jeghers syndrome	Intestinal, ovarian, pancreatic, colorectal	PI3K	[24,29]
TP53	ENSG00000141510	16	1185	Li-Fraumeni syndrome/sarcoma	Breast, sarcoma, adrenal, brain, colorectal	p53	[24,29]
TSC1	ENSG00000165699	18	3495	Tuberous sclerosis	Hamartoma, kidney, colorectal	PI3K	[24,29]
TSC2	ENSG00000103197	19	5436	Tuberous sclerosis	Hamartoma, kidney, colorectal	PI3K	[24,29]
VHL	ENSG00000134086	18	639	Von Hippel-Lindau syndrom	Kidney, colorectal	HIF1	[24]

Each of the 22 genes analyzed in this study are detailed, including their HGNC approved gene symbols, and Ensembl gene IDs. The total number of species analyzed for each gene and the overall length of alignment in base pairs are also given. The syndrome, tumor type observed and pathway involved are detailed. References citing alternative gene names are identified using brackets

Multiple sequence alignment

The coding DNA sequences of the single gene orthologs were translated and the resulting amino acid sequences were aligned using the default parameters in ClustalW 2.0.12 [30,31]. Using in-house software, we mapped gaps from the amino acid multiple sequence alignment (MSA) to the corresponding nucleotide sequences to produce a nucleotide alignment. All alignments were reviewed for quality and any poorly aligned regions were manually edited using Se-Al [32]. All alignments are available in Additional file 2.

Alignment criteria for selective pressure analysis

It has been shown through computer simulations that sequence length has an impact on the power to infer positive selection [33]. Power was also found to increase with greater taxonomic representation and greater divergence, although extreme levels of divergence were found to cause a reduction in power. Simulations have shown that the presence of longer foreground branches also increased the power of the test statistic, but extremely long foreground branches reduce the power [34]. To increase the statistical power of the analysis performed here we have therefore only considered single gene families containing 6 or more taxa and lengths of greater than 500 amino acids.

Recombination analysis

Recombination events can result in the incorrect detection of positive selection. To reduce potential false positives in our analysis, we have implemented GENECONV (version 1.81a) [35] using default parameters. GENECONV detects gene conversion events between ancestors of sequences in the multiple sequence alignment. Default parameters where employed, 10,000 randomly permuted datasets were generated for each Single Gene Orthologous family and global inner fragments were listed if P-value was 0.05 or smaller.

Selective pressure analysis using codon models of evolution

Selective pressure analyses were performed using Codeml from PAML version 4.4 [36,37]. Because each gene family analyzed was composed of single gene orthologs, pruned species phylogenies were used as per previous publications [2,38]. Codeml implements a number of codon-based models in a Maximum Likelihood framework that can be used to test alternative and nested evolutionary hypotheses. Three different types of codon model were used in this study: (i) a homogeneous model (Model 0) - a single ω-value is estimated for the entire alignment; (ii) site-heterogeneous models - the sites of the gene sequence are grouped into two or more site classes, each with its own ω-value estimate; and (iii) lineage-specific heterogeneous models - a different ω parameter is estimated for different site classes in combination with different lineages [9,36,39].

Seven site-heterogeneous models were used, we have retained conventional annotations for these models: Model 1 (Neutral), Model 2 (Selection), Model 3 Discrete ($k = 2$), Model 3 Discrete ($k = 3$), Model 7, Model 8 and Model 8a. Two lineage-specific heterogeneous models were used: Model A and Model A Null. These models have been applied similarly elsewhere [40].

The goodness-of-fit of the different models was assessed statistically using a likelihood ratio test (LRT). The LRT compares the log-likelihoods of a null model with the alternative model. For hierarchically nested models, the test statistic of an LRT approximates the χ^2 distribution with degrees of freedom equal to the number of additional free parameters in the alternative model compared to the null model. Because of this, the critical value of the test statistic can be determined from standard statistical tables. If the p-value of the test statistic exceeds that critical value (i.e. if the alternative model fits the data significantly better than the null model), then the null model can be rejected. For example, if the test statistic of an LRT comparing Model 1 (Neutral) with Model 2 (Selection) is greater than the critical value determined from the χ^2 distribution, Model 1 can be rejected. If $\omega_1 > 1$ under Model 2, positive selection may be inferred. Additional file 3 shows the set of LRTs used for selection analysis.

In cases where positive selection is inferred, the posterior probability of a site belonging to the positively selected class is estimated using two calculations: Naïve Empirical Bayes (NEB) or Bayes Empirical Bayes (BEB). If both BEB and NEB are predicted, we will preferentially use the BEB results as have been shown to be more robust [37].

In-house software was designed to prepare all files for analysis and to process all output. PAML output files were parsed for parameter estimates and log likelihood values and LRTs were performed (see Additional file 3 for LRTs performed). Where positively selected sites were inferred under a given model, positively selected sites were mapped to the sequence (or sequences) of interest and included in the summary file (see Additional file 4). This software was used to reduce the scope for human error in PAML analyses and is available from the authors on request. Functional annotation of sites under positive selection for each protein was obtained from UniProt [41].

Human population analysis

Selective pressures within the present day human population were analyzed for those genes with evidence of lineage-specific positive selection in the human ancestral lineages. The online tool SNP@Evolution[2] and HapMap release II source data was used to look at variations within the East Asian (A), Northern and Western European (C), and African Yoruba (Y) populations. The "integrated haplotype score" or iHS, described first in [42], was employed here as a test for directional selection. The iHS is standardized using genome wide empirical distributions and has an approximate normal distribution allowing for direct comparisons of the score across genes, and it outperforms in comparison to other available approaches [42]. A derived allele that has been segregating in the population receives a large iHS ($> +2$) while a large negative iHS (<-2) indicates that the derived allele has increased in frequency.

Results and discussion

Starting with a dataset of 22 genes, we identified single gene orthologs across 21 complete vertebrate genomes. Ortholog identification resulted in families with between 15 and 21 taxa, and alignment lengths of between 507 and 9,189 base pairs thus satisfying the dataset requirements described in the materials and methods section. The test for recombination on all 22 genes is summarized in Additional file 5. The analysis revealed that only the TP53 protein showed significant levels of recombination, the regions where recombination was

present was noted and compared to regions where positive selection was detected. If these regions overlapped - the positive selection result was deemed a false positive.

To assess the selective pressure variation, we performed both site- and lineage-specific selective pressure analyses and subsequently assessed the statistical significance of all results via LRT analysis to ascertain the codon evolutionary model of best fit. In those cases where the ω value vastly exceeds 1, we have simply denoted them as $\omega >> 1$ throughout the manuscript, as there is no biological significance for these extremely large ω values (the precise numbers are shown in the Tables throughout). The lineage-specific analyses are more pertinent to the main focus of the paper - the identification of species-specific patterns of substitution in these colon cancer associated genes. Therefore the lineage-specific results have been described in detail in the following section. Site-specific results briefly summarized on a gene-by-gene basis. All positively selected sites were assessed using the functional information from the Uniprot database [41]. The model of best fit along with associated parameter estimates are described and a summary table for all estimates for each of the 22 genes is given in Additional file 4.

Lineage-specific selective pressure analyses

Lineage-specific models of codon evolution were assessed at multiple phylogenetic depths, (i) the extant lineages within the Euarchontoglires clade, and (ii), all ancestral lineages leading from the Euarchontoglires to modern mouse and human were also tested independently as depicted in Figure 1. Analysis of the extant human and mouse lineages did not yield evidence of positive selection. Conversely, analysis of the lineages within the Euarchontoglires clade resulted in significant evidence of lineage-specific positive selection, 6 genes in ancestral lineages and 12 in extant lineages, see Figure 1. These lineage-specific results include 6 ancestral lineages and 12 extant lineages with evidence of positive selection. The STK11 gene showed evidence of positive selection in the Euarchontoglires ancestral lineage and again in the Hominidae ancestral lineage. CDH1 showed patterns of substitution conducive with positive selection in the ancestral primate lineage. The ancestral Muridae lineage had evidence of positive selection acting on the TSC1 gene. The ancestral Murinae lineage showed evidence of positive selection for both MSH6 and SDHC, see Table 2 for summary.

Table 2 Summary of parameter estimates and likelihood scores for the model of best fit showing evidence of positive selection

Gene	Model	lnL	Parameter Estimates	Positive Selection	BEB Positively Selected Sites
Lineage-Specific Analyses					
Euarchontoglires Ancestral Branch					
STK11	modelA	−8602.921472	$p_0 = 0.93299$, $p_1 = 0.05633$, $p_2 = 0.01007$, $p_3 = 0.00061$ $\omega_0 = 0.03346$, $\omega_1 = 1.00000$, $\omega_2 = 197.90897$	Yes	3 > 0.50, 1 > 0.95, 0 > 0.99
Primate Ancestral Branch					
CDH1	modelA	−16658.03484	$p_0 = 0.75454$, $p_1 = 0.23453$, $p_2 = 0.00834$, $p_3 = 0.00259$ $\omega_0 = 0.05683$, $\omega_1 = 1.00000$, $\omega_2 = 10.20516$	Yes	9 > 0.50, 1 > 0.95, 0 > 0.99
Hominidae Ancestral Branch					
STK11	modelA	−8601.056009	$p_0 = 0.93574$, $p_1 = 0.05920$, $p_2 = 0.00476$, $p_3 = 0.00030$ $\omega_0 = 0.03323$, $\omega_1 = 1.00000$, $\omega_2 = 44.31709$	Yes	3 > 0.50, 2 > 0.95, 1 > 0.99
VHL	modelA	−4263.853291	$p_0 = 0.73748$, $p_1 = 0.25109$, $p_2 = 0.00853$, $p_3 = 0.00290$ $\omega_0 = 0.05985$, $\omega_1 = 1.00000$, $\omega_2 = 220.34533$	Yes	1 > 0.50, 0 > 0.95, 0 > 0.99
Chimpanzee Extant Branch					
TSC2	modelA	−42659.27711	$p_0 = 0.90352$, $p_1 = 0.09434$, $p_2 = 0.00194$, $p_3 = 0.00020$ $\omega_0 = 0.04404$, $\omega_1 = 1.00000$, $\omega_2 = 190.09480$	Yes	6 > 0.50, 2 > 0.95, 2 > 0.99
VHL	modelA	−4262.098043	$p_0 = 0.73571$, $p_1 = 0.25251$, $p_2 = 0.00877$, $p_3 = 0.00301$ $\omega_0 = 0.05976$, $\omega_1 = 1.00000$, $\omega_2 = 262.72662$	Yes	3 > 0.50, 0 > 0.95, 0 > 0.99
Gorilla Extant Branch					
MSH2	modelA	−19485.4338	$p_0 = 0.92233$, $p_1 = 0.06298$, $p_2 = 0.01375$, $p_3 = 0.00094$ $\omega_0 = 0.06427$, $\omega_1 = 1.00000$, $\omega_2 = 999.00000$	Yes	46 > 0.50, 34 > 0.95, 18 > 0.99
TSC2	modelA	−42569.22884	$p_0 = 0.89862$, $p_1 = 0.08796$, $p_2 = 0.01222$, $p_3 = 0.00120$ $\omega_0 = 0.04339$, $\omega_1 = 1.00000$, $\omega_2 = 999.00000$	Yes	27 > 0.50, 14 > 0.95, 12 > 0.99
MSH6	modelA	−34009.90221	$p_0 = 0.78382$, $p_1 = 0.18418$, $p_2 = 0.02591$, $p_3 = 0.00609$ $\omega_0 = 0.06974$, $\omega_1 = 1.00000$, $\omega_2 = 999.00000$	Yes	46 > 0.50, 34 > 0.95, 18 > 0.99
ATM	modelA	−69374.08393	$p_0 = 0.80673$, $p_1 = 0.17971$, $p_2 = 0.01109$, $p_3 = 0.00247$ $\omega_0 = 0.09745$, $\omega_1 = 1.00000$, $\omega_2 = 999.00000$	Yes	48 > 0.50, 23 > 0.95, 19 > 0.99
Orangutan Extant Branch					

TSC1	modelA	−24068.71106	$p_0 = 0.79963$, $p_1 = 0.18828$, $p_2 = 0.00978$, $p_3 = 0.00230$ $\omega_0 = 0.08020$, $\omega_1 = 1.00000$, $\omega_2 = 999.00000$	Yes	13 > 0.50, 6 > 0.95,5 > 0.99
TSC2	modelA	−42673.92339	$p_0 = 0.90414$, $p_1 = 0.09295$, $p_2 = 0.00263$, $p_3 = 0.00027$ $\omega_0 = 0.04433$, $\omega_1 = 1.00000$, $\omega_2 = 40.47366$	Yes	9 > 0.50, 0 > 0.95, 0 > 0.99
Marmoset Extant Branch					
TSC2	modelA	−42616.04524	$p_0 = 0.89841$, $p_1 = 0.09019$, $p_2 = 0.01035$, $p_3 = 0.00104$ $\omega_0 = 0.04325$, $\omega_1 = 1.00000$, $\omega_2 = 235.10448$	Yes	38 > 0.50, 9 > 0.95
MSH6	modelA	−34009.90221	$p_0 = 0.78382$, $p_1 = 0.18418$, $p_2 = 0.02591$, $p_3 = 0.00609$ $\omega_0 = 0.06974$, $\omega_1 = 1.00000$, $\omega_2 = 999.00000$	Yes	45 > 0.50, 16 > 0.95, 12 > 0.99
VHL	modelA	−4262.443441	$p_0 = 0.72045$, $p_1 = 0.22453$, $p_2 = 0.04195$, $p_3 = 0.01307$ $\omega_0 = 0.05886$, $\omega_1 = 1.00000$, $\omega_2 = 90.26952$	Yes	10 > 0.50, 0 > 0.95, 0 > 0.99
ATM	modelA	−69583.23068	$p_0 = 0.81640$, $p_1 = 0.18148$, $p_2 = 0.00173$, $p_3 = 0.00038$ $\omega_0 = 0.09939$, $\omega_1 = 1.00000$, $\omega_2 = 46.82466$	Yes	2 > 0.50, 0 > 0.95, 0 > 0.99
Muridae Ancestral Branch					
TSC1	modelA	−24126.17894	$p_0 = 0.80995$, $p_1 = 0.18416$, $p_2 = 0.00481$, $p_3 = 0.00109$ $\omega_0 = 0.08293$, $\omega_1 = 1.00000$, $\omega_2 = 999.00000$	Yes	1 > 0.59, 0 > 0.95, 0 > 0.99
Murinae Ancestral Branch					
SDHC	modelA	−3846.690164	$p_0 = 0.87666$, $p_1 = 0.08131$, $p_2 = 0.03846$, $p_3 = 0.00357$ $\omega_0 = 0.15340$, $\omega_1 = 1.00000$, $\omega_2 = 253.61375$	Yes	9 > 0.50, 2 > 0.95, 1 > 0.99
MSH6	modelA	−34190.13821	$p_0 = 0.79911$, $p_1 = 0.19671$, $p_2 = 0.00335$, $p_3 = 0.00082$ $\omega_0 = 0.07057$, $\omega_1 = 1.00000$, $\omega_2 = 126.22513$	Yes	3 > 0.50, 1 > 0.95, 0 > 0.99
Rat Extant Branch					
MADH4	modelA	−6092.186945	$p_0 = 0.93360$, $p_1 = 0.01536$, $p_2 = 0.05021$, $p_3 = 0.00083$ $\omega_0 = 0.01379$, $\omega_1 = 1.00000$, $\omega_2 = 102.33013$	Yes	24 > 0.50, 11 > 0.95, 10 > 0.99
NF1	modelA	−37750.29866	$p_0 = 0.96609$, $p_1 = 0.02476$, $p_2 = 0.00892$, $p_3 = 0.00023$ $\omega_0 = 0.02265$, $\omega_1 = 1.00000$, $\omega_2 = 999.00000$	Yes	39 > 0.50, 10 > 0.95, 10 > 0.99
Guinea pig Extant Branch					
TSC1	modelA	−24116.58577	$p_0 = 0.80206$, $p_1 = 0.18611$, $p_2 = 0.00961$, $p_3 = 0.00223$ $\omega_0 = 0.08093$, $\omega_1 = 1.00000$, $\omega_2 = 284.22603$	Yes	9 > 0.50, 4 > 0.95, 0 > 0.99

NF1	modelA	−37849.50819	$p_0 = 0.97375$, $p_1 = 0.02506$, $p_2 = 0.00116$, $p_3 = 0.00003$ $\omega_0 = 0.02414$, $\omega_1 = 1.00000$, $\omega_2 = 171.64068$	Yes	3 > 0.50, 1 > 0.95, 0 > 0.99
Rabbit Extant Branch					
MLH1	modelA	−19516.63525	$p_0 = 0.80595$, $p_1 = 0.18541$, $p_2 = 0.00703$, $p_3 = 0.00162$ $\omega_0 = 0.05262$, $\omega_1 = 1.00000$, $\omega_2 = 7.52747$	Yes	5 > 0.05, 3 > 0.95, 0 > 0.99
MUTYH	modelA	−15911.6175	$p_0 = 0.61027$, $p_1 = 0.37605$, $p_2 = 0.00846$, $p_3 = 0.00522$ $\omega_0 = 0.07703$, $\omega_1 = 1.00000$, $\omega_2 = 998.99697$	Yes	5 > 0.50, 4 > 0.95, 3 > 0.99
SDHC	modelA	−3822.683246	$p_0 = 0.57771$, $p_1 = 0.06636$, $p_2 = 0.31926$, $p_3 = 0.03667$ $\omega_0 = 0.12047$, $\omega_1 = 1.00000$, $\omega_2 = 3.59059$	Yes	51 > 0.50, 10 > 0.95, 8 > 0.99
ATM	modelA	−69582.95152	$p_0 = 0.81572$, $p_1 = 0.18045$, $p_2 = 0.00313$, $p_3 = 0.00069$ $\omega_0 = 0.09930$, $\omega_1 = 1.00000$, $\omega_2 = 7.41594$	Yes	6 > 0.50, 0 > 0.95, 0 > 0.99
BHD	modelA	−13523.51719	$p_0 = 0.90728$, $p_1 = 0.05930$, $p_2 = 0.03137$, $p_3 = 0.00205$ $\omega_0 = 0.02817$, $\omega_1 = 1.00000$, $\omega_2 = 6.50017$	Yes	10 > 0.50, 7 > 0.95, 1 > 0.99
Site-specific Analyses					
CDH1	m8	−16589.88768	$p = 0.21848$, $p_0 = 0.99291$, $p_1 = 0.00709$, $q = 0.80842$ ω=4.53766	Yes	15 > 0.5, 1 > 0.95, 0 > 0.99
PMS1	m8	−26480.39761	$p = 0.61337$, $p_0 = 0.93580$, $p_1 = 0.06420$, $q = 1.93110$ ω=1.32691	Yes	25 > 0.50, 1 > 0.95, 0 > 0.99
PMS2	m8	−27449.3651	$p = 0.29104$, $p_0 = 0.91064$, $p_1 = 0.08936$, $q = 1.31619$ ω=1.28855	Yes	37 > 0.50, 1 > 0.95, 0 > 0.99
MUTYH	m8	−15797.6226	$p = 0.37255$, $p_0 = 0.97242$, $p_1 = 0.02758$, $q = 1.00900$ ω=2.44412	Yes	18 > 0.5, 1 > 0.95, 0 > 0.99
TP53	m8	−8688.19126	$p = 0.40362$, $p_0 = 0.94645$, $p_1 = 0.05355$, $q = 1.77507$ ω=1.97385	Yes	13 > 0.5, 3 > 0.95, 0 > 0.99

The model of best fit is summarized below for those genes with evidence of positive selection. The lineage-specific results for each lineage tested from the Euarchontoglires ancestor to modern lineages are shown in the top panel and the site-specific results are shown in the bottom panel. The model abbreviations are as per main text. P refers to the number of free parameters estimated in that model. BEB = Bayes Empirical Bayes estimations. The number of positively selected sites identified can be found the final column, sites are separated by the posterior probability cutoffs of 0.50, 0.95, and 0.99

In the following section, we have analyzed the positively selected sites for those genes with evidence of lineage-specific positive selection in the context of their potential functional relevance for those genes. This was carried out for all genes where functional sites and/or domains have been elucidated. All sites described were calculated via Bayes Empirical Bayes (BEB) analysis (unless otherwise specified). In all cases we are assessing the potential functional importance of residues based on their sequence position. There are instances where we identify stretches of protein sequence under positive selection - there is a possibility that these regions may have very different functions despite their sequence position. For a total 16 of the 22 genes there were partial or complete 3D structures available. However, many of the positively selected sites identified were located in regions that were not yet fully resolved at the structural level, and so only the 3D model for STK11 is given. Corresponding alignments are available in Additional file 2. The complete set of model estimates for the entire dataset are available in Additional file 4.

Positive selection in the Euarchontoglires Ancestral branch

The most ancestral branch tested was the Euarchontoglires ancestral branch, i.e. the ancestor of the Primate, Rodent and Glires clade as depicted in Figure 1. The STK11 alignment consists of 18 taxa and was the only gene that showed evidence of positive selection in this lineage. STK11 (Serine/Threonine-protein kinase 11) plays an essential role in G1 cell cycle arrest and acts as a tumor suppressor. It phosphorylates and activates members of the AMPK-related subfamily of protein kinases (Baas [43],; Boudeau [44],). Mutations in STK11 cause Peutz-Jeghers syndrome (PJS), a rare autosomal dominant disorder characterized by multiple gastrointestinal hamartomatous polyps and an increased risk of various neoplasms including gastrointestinal cancer [45,46]. From the literature we currently know of 17 sites across this gene that when mutated are associated with colon-cancer. The Euarchontoglires ancestral lineage has 1.1% of sites under positive selection ($\omega > > 1$). The positively selected residues were located on the 3D structure of this enzyme (See Figure 1 inset). Position 206 with a PP = 0.889 is a hydrophobic Alanine or Valine in Euarchontoglires species or a negatively charged Glutamic acid or positively charged Lysine in non-Euarchontoglires species. This residue also lies in close proximity to sporadic cancer site A205T and colorectal cancer site D208N in Human [47]. Positively selected position 301 in Euarchontoglires (P = 0.885) is present in Euarchontoglires species as an Arginine residue and all non-Euarchontoglires as an uncharged Glutamine residue. Site 301 is close to R297K and region 303–306 both of which have been implicated in PJS [48].

Positive selection in the Primate Ancestral branch

The branch leading from the Euarchontoglires ancestor towards the primates was analyzed, we have termed this branch the ancestral Primate branch as depicted in Figure 1. The CDH1 dataset consists of 15 taxa and following LRT analysis showed evidence of lineage-specific analysis identified positive selection in 1.1% of sites in the Primate Ancestor (ω=10.21). Positively selected sites were compared to human Swiss-Prot entry (P12830) and it was found that position 604, with a Posterior Probability (PP) of 0.549, falls in close proximity to gastric cancer variant R598Q [49]. At position 604 Primates have a negatively charged Glutamic acid while non-primates have a polar uncharged Glutamine.

Positive selection in the Hominidae Ancestral branch

The next branch in the primate clade is that leading to modern great apes, i.e. Hominidae, as depicted in Figure 1. This lineage also showed evidence of positive selection again in the STK11 gene in 0.51% of sites, or 3 positions, with $\omega >> 1$. See Figure 2(a) and Table 2. These positions were then compared to the human Swiss-Prot sequence (Q15831). Position 347 represents a radical substitution, as the Hominidae code for an Alanine (a small hydrophobic residue) whereas the Murinae lineage encode an Arginine at this position (a basic, hydrophilic, and positively charged residue). For positively selected site 378, the ancestral Hominidae lineage encodes the polar residue Serine, while the closely related species studied encode the small amphiphilic Glycine. The functions of these specific sites have not been reported thus far in the literature but are likely to be of considerable interest as they mark adaptations unique to the ancestral Hominidae.

Figure 2 Positive selection analysis for 4 genes: (a) STK11, (b) CDH1, (c) MUTYH, and (d) TP53. The x-axis depicts the gene from start to end of alignment. The Y-axis is the posterior probability. The vertical red bars on each graph represent the known cancer causing variants from human populations. The black dots on each graph represent the positively selected sites identified in this study

A second gene showing evidence of positive selection in the Hominidae ancestral branch is the VHL dataset consisting of 18 taxa. The VHL gene encodes Von Hippel-Lindaue tumour suppressor protein. Mutations occurring in this gene can result in von Hippel-Lindau disease (VHDL) - a dominantly inherited familial cancer syndrome [50]. VHL exhibited weak evidence of positive selection with 1.1% of sites in the ancestral Hominidae lineage under positive selection. There was one amino acid that had low coverage in the alignment (present only in 6/18 species), as this is a very weak results we have not expanded upon it any further.

Positive selection in the Extant Primate branches

Analysis of modern non-human primates also identified positive selection in a number of genes. In VHL positive selection was detected in the Chimpanzee lineage with 1.2% of sites with $\omega >> 1$, and also in the Marmoset lineage with 5.5% sites with $\omega >> 1$. Sites under positive selection were compared against human Swiss-Prot entry (P40337), however the region (1–60) was only represented by 11/18 species in the alignment and therefore we do not have sufficient confidence in these positions to explore these sites in more detail.

The MSH6 gene dataset contained 19 taxa and showed evidence of positive selection in both the Gorilla and Marmoset lineages each displaying 3.2% of sites with $\omega >> 1$. Gorilla and Marmoset extant lineages were compared against human (P52701) Swiss-Prot entry. No relevant functional information could be extracted from positively selected sites in Gorilla, however 2/45 positively selected sites in Marmoset fall in close proximity to cancer variants. Marmoset positively selected site 803 ($PP = 0.551$) coincides with colorectal cancer variants D803G [51] and V800A [52] in Human. Position 803 in Marmoset is a negatively charged Glutamic acid while in all other mammals it is a small negatively charged Aspartic acid. Positively selected site 1099 in Marmoset ($PP = 0.614$) is located between human colorectal cancer variants R1095H [53] and T1110C [54].

MSH2 alignment consists of 18 taxa. The function of the MSH2 protein is in post-replicative DNA mismatch repair system (MMR). Mutations in MSH2 result in hereditary non-polyposis

colorectal cancer type 1 (HNPCC1) [55]. Lineage-specific positive selection was identified in 1.5% of sites within the extant Gorilla lineage with $\omega >> 1$. Positively selected sites were compared against human Swiss-Prot sequence (P43246). All 15 of the BEB identified sites occur between amino acid position 124–142 which overlaps with the region containing variants N127S, N139S and I145M associated with HNPCC1 [55].

Tuberous sclerosis 2 protein (TSC2) interacts with TSC1 protein and mutations in this gene can cause tuberous sclerosis type 2 [56]. The alignment of TSC2 consisted of 19 taxa. Lineage-specific positive selection was identified in the following extant lineages, the percentage of sites under positive selection in each lineage is shown in brackets, in all cases $\omega >> 1$: Chimpanzee lineage (0.2%), Gorilla (1.3%), Orangutan (0.29%), and, Marmoset (1.1%). Positively selected sites were compared against human Swiss-Prot sequence (P49815) however the functional information was not available to contextualize these results.

ATM acts as a DNA checkpoint sensor by activating checkpoint signaling upon double strand breaks [57]. The alignment of ATM consisted of 18 taxa and positive selection was detected in the following lineages (again the percentage of the alignment under positive selection is shown in brackets): Gorilla (1.4%, $\omega >> 1$), Marmoset (0.21%, $\omega >> 1$), and Rabbit (0.38%, $\omega = 7.42$). BEB significant sites were compared to human (Q13315) and mouse (Q62388) Swiss-Prot entries to determine the functional relevance of selected sites. In the Gorilla lineage positively selected site 2067 (PP = 0.787), where in humans a substitution of Alanine to Aspartate in this same position can result in Ataxia telangiectasia (AT) which is a severe disease that causes weakened immune function and higher disposition to cancer [57]. No other functionally relevant information was found upon comparison of Swiss-Prot information against either Marmoset or Rabbit.

The extant Orangutan lineage also showed evidence of positive selection in the TSC1 gene for 1.2% of its alignment $\omega >> 1$. Positively selected sites were compared against human Swiss-Prot sequence (Q92574) and mouse Swiss-Prot sequence (Q9EP53) however there was insufficient information to extrapolate potential functional impacts of these sites.

Human population level analysis using HapMap data

Genes displaying evidence of positive selection in lineages leading to *Homo sapiens*, i.e. the primate and Hominidae lineages (STK11, CDH1 and VHL), were further analyzed to determine if there is evidence for ongoing positive directional selection in modern day human populations. The integrated haplotype score, his [42], was calculated for each SNP in STK11, CDH1 and VHL genes across African Yorubu (Y), East Asian (A) and European (C) populations. One intronic SNP in the SDK11 gene, had an iHS score of +2.0385 in European populations. In the CDH1 gene, two intronic SNPs with iHS scores of +2.0433 and +2.5838 respectively were identified in the East Asian populations. The iHS scores of greater than +2 indicate that these alleles are segregating at a significant rate within their given populations. No population level directional selection was identified in the VHL gene in modern humans.

Positive selection in the Ancestral Muridae branch

The ancestral Muridae branch marks the most recent common ancestor of modern mouse, rat and guinea pig species and is depicted in Figure 1. Tuberous sclerosis 1 protein (TSC1) interacts with TSC2 and acts as a tumour suppressor gene [56]. Defects in TSC1 cause tuberous sclerosis type 1 which is an autosomal dominant multi-system disorder. There were

a total of 18 taxa analysed in for the TSC1 gene and 0.59% of sites in the Muridae ancestral lineage were identified with $\omega >> 1$. As before for TSC1: positively selected sites were compared against human Swiss-Prot sequence (Q92574) and mouse Swiss-Prot sequence (Q9EP53) however there was insufficient information to extrapolate potential functional impacts of these sites.

Positive selection in the Ancestral Murinae branch

Tha ancestral Murinae branch defines the most recent common ancestor of mouse and rat. In total there were two genes identified as being under positive selection in the Murinae lineage. The first is the MSH6 gene that acts as a DNA mismatch repair protein and is a component of the post–replicative DNA mismatch repair system [58]. MSH6 also heterodimerizes with MSH2 to form MutS-alpha, a protein complex that functions by binding to DNA mismatches and initiating DNA repair [59]. Mutations in MSH6 have been reported to cause HNPCC type 5 [60], atypical HNPCC, and familial colorectal cancers (suspected or incomplete HNPCC) [61]. The MSH6 dataset consists of 19 taxa. Lineage-specific analysis of the ancestral Murinae lineage revealed 0.42% of the sites (3 residues) in MSH6 under positive selection, $\omega >> 1$ (see Table 2). The corresponding Swiss-Prot sequence (P54276) lacked functional details for these positions, therefore, potential functional effects remain unknown. However, examination of the alignment at this position revealed the subsitution of residues with unrelated biochemical properties at these positions. At positively selected site 374 (numbered as per Swiss-Prot entry), the Murinae lineage has a Proline whereas remaining species tested encode either Glutamic acid, Aspartic acid, or Lysine. As Proline produces "kinks" in the α-helical regions of proteins, such a substitution could alter the protein structure substantially. Positively selected site 759 is a Leucine in the Murinae, all other non-outgroup species encode aliphatic residues (Isoleucine or Valine). The ancestral Murinae has a Cysteine at Swiss-Prot position 1259 while all other species have an Alanine at this position. These residues are of specific interest for further *in vitro* functional assaying given their uniqueness to the rodent clade and their retention in all modern rodents tested.

The second gene with evidence of positive selection on the ancestral Murinae lineage is the SDHC (Succinate dehydrogenase cytochrome b 560 subunit, mitochondrial) gene. The SDHC function is to act as a membrane-anchoring subunit for the SDH protein. Defects in this protein are reported in paragangliomas and gastric stromal sarcomas [62]. The dataset for the SDHC consisted of 16 taxa. Lineage-specific positive selection was detected in the ancestral Murinae lineage with 4.2% of sites (9 residues) in this protein with $\omega >> 1$ (Table 2). Comparison with the human sequence from Swiss-Prot (Q99643) and mouse sequence (Q9CZB0) placed 8 of these sites either in transmembrane or topological domains across the gene, with the additional positively selected residue (position 128) neighboring a metal binding site at position 127.

Positive selection in the Extant Rabbit branch

The SDHC gene again showed evidence of positive selection, this time in the extant Rabbit lineage with 35.59% of sites under positive selection ($\omega = 3.59$). 15/51 positively selected sites were identified as occurring within 10 amino acid positions of metal binding site at position 127 that is also mentioned in the ancestral Murinae analysis. While there are extremely high levels of positive selection identified in the rabbit lineage, no other relevant functional information could be gleaned from the databases at this point.

The MUTYH alignment consisted of 21 taxa and showed evidence of lineage-specific analysis identified positive selection in 1.4% of sites in the extant Rabbit lineage ($\omega >> 1$). Positively selected sites were compared to human (Q9UIF7) and mouse (Q99P21) Swiss-Prot entries, however no relevant functional information could be extrapolated. Radical substitutions occurred in all 5 BEB sites in the extant Rabbit lineage, three of which are at positions 485–487 in the Nudix hydrolase domain.

The MLH1 gene codes for a critical protein involved with the post-replicative DNA mismatch repair system. Defects in this gene result in hereditary non-polyposis colorectal cancer type 2 (HNPCC2) [63]. The alignment of MLH1 consists of 19 taxa and again positive selection was detected in the extant rabbit lineage in 0.87% of sites ($\omega = 7.53$). Positively selected sites were compared against human Swiss-Prot sequence (P40692) and mouse Swiss-Prot sequence (Q9JK91). At amino acid position 120, Rabbit has a polar uncharged Serine residue while all other species tested have a hydrophobic Alanine residue. This positively selected site falls in a region dense with HNPCC2 variants at positions A111V, T116K, T117M, Y126N, A128P [63-65]. Positively selected residues in Rabbit: 209, 478 and 514, each fall within 8 amino acid positions of HNPCC2 variants: V213M, R474Q and V506A [66]. And position 478 identified as under positive selection also lies in close proximity to a colorectal cancer variant R472I (Kim [67],).

Finally, the BHD gene showed evidence of positive selection in the extant Rabbit lineage. The function of the BHD gene is still largely unknown, however it is thought that it may be a tumour suppressor and it may be involved in colorectal tumorigenesis [68]. The alignment consisted of 20 taxa and positive selection was detected in 3.34% of sites ($\omega = 6.5$), again unique to the Rabbit lineage. BEB significant sites were compared to human (Q8NFG4) and mouse (Q8QZS3) Swiss-Prot entries to determine their functional relevance. All 10 of the positively selected sites in Rabbit occur in a small region from position 61–83 and border a known human cancer variant at position 79 that when mutated from Serine to Tryptophan results in sporadic colorectal carcinoma.

Positive selection in the Extant Rodent and Guinea Pig branches

MADH4 is the co-activator and mediator of signal transduction by TGF-beta. Defects in MADH4 result in pancreatic, colorectal, juvenile polyposis syndrome, juvenile intestinal polyposis and primary pulmonary hypertension [69,70]. The Rat lineage was identified as being under lineage-specific positive selection in the MADH4 gene where 5.1% of sites are evolving with $\omega >> 1$ (number of taxa = 16). Positively selected sites were compared to human (Q13485) and mouse (P97471) Swiss-Prot entries. The majority of positively selected residues in this protein are sequential with 18/24 sites under positive selection in the rat lineage within 10 amino acid positions of the natural human variant 493. When position 493 is mutated from Aspartate to Histidine pancreatic carcinoma is induced [71].

NF1 is thought to be a regulator of RAS activity [72]. Defects in NF1 can cause colorectal carcinoma and breast cancer [70]. The NF1 dataset consists of 17 taxa. Lineage-specific positive selection was identified in 0.92% of sites in Rat with $\omega >> 1$ and 0.12% of sites in guinea pig with $\omega >> 1$. BEB significant sites were compared to human (P21359) and mouse (Q04690) Swiss-Prot sequences, however there was no functionally relevant information available.

TSC1 also shows evidence of positive selection in the extant guinea pig lineage with 1.2% of the sites with $\omega >> 1$. As before, the positively selected sites were compared against human Swiss-Prot sequence (Q92574) and mouse Swiss-Prot sequence (Q9EP53) however there was insufficient information to extrapolate potential functional impacts of these sites.

Results of site-specific selective pressure analyses

The site-specific results may be beneficial to those working on rational mutagenesis and/or the identification of functionally important regions in these colon cancer associated genes and so these results have been summarized. We have identified five genes that have signatures of site-specific positive selection, namely: CDH1, MUTYH, PMS1, PMS2 and TP53, representing ~23% of the dataset. For each of these five genes, the model of best fit was the site-heterogeneous model "model 8", see Table 2 for summary.

Defects in the CDH1 member of the Cadherin family are linked to hereditary diffuse gastric cancer [24,28]. The CDH1 alignment contained 15 taxa and site-specific analysis identified 0.71% sites evolving under strong positive selection, $\omega = 4.54$, see Table 2. We compared these sites to the human Swiss-Prot entry (P12830) to obtain relevant functional information, see Figure 2(b). The vast majority of positively selected sites (12 sites) in the protein are found within the extracellular topological domain (positions 155–709). Many of these positively selected are in close proximity to natural cancer variants. For example, position 421 is under position selection and resides within a region (418–423) known to be missing in gastric carcinoma samples [73]. Positions 457, 465, and 467 are under positive selection and map in close proximity to natural variant E463Q found in gastric carcinoma samples [49]. Position 700 resides within the metalloproteinase cleavage site (700–701) of CDH1. Position 735 is in close proximity to a gamma-secretase/PS1 cleavage site (731–732) [74], and position 553 is in close proximity to a glycosylation site (558), essential for the posttranslational modification of proteins [75]. In the CDH1 gene, the majority of species tested (8/15) have hydrophobic residues (Isoleucine, Valine, Leucine) at position 553, the glires group (mouse, rat, guinea pig and rabbit) have small residues (Alanine, Serine, Threonine), but human, gorilla, and dog have large aromatic residues (Phenylalanine) that could significantly alter the protein structure and may affect binding at the glycosylation site at position 558.

The MUTYH dataset consisted of 21 taxa and site-specific analysis identified 18 sites under positive selection ($\omega = 2.44$), representing 2.8% of the MUTYH protein (Table 2). A total of 10 unique sites are reported as natural cancer variants in human (Q9UIF7), see Figure 2(c). Positively selected sites 406 and 412 are in close proximity to natural cancer variants at positions 402 and 411 respectively. Positively selected sites 521, 528 and 538 also map in close proximity to natural variants, 526 and 531 respectively. Also of note are the replacement substitutions observed at Swiss-Prot positions 406 and 412, these are radical with potential effects on protein structure. At position 406 there is a large aromatic Trytophan in Primates, and a hydrophobic Leucine and Valine present in the Glires. At position 412 there is an hydrophobic Leucine in Primates and a positively charged Histidine in the Glires.

PMS1 (postmeiotic segregation increased 1) encodes a DNA mismatch repair protein and this dataset consisted of 20 taxa. Defects in PMS1 are reported to cause hereditary non-polyposis colorectal cancer type 3 (HNPCC3) [76]. Analysis of PMS1 identified site-specific model of codon evolution model 8 as best fit, estimating 25 positively selected sites (6.4% of the alignment) with $\omega = 1.33$ (Table 2). We compared these sites against human Swiss-Prot

sequence P54277. Positively selected site 387 resides in close proximity to position 394 - a natural variant (M394T) reported in incomplete HNPCC and HNPCC3 [77]. Due to limited functional data it was unfeasible to study the remaining 24 sites. However, due to PMS1 function in DNA mismatch repair, these positively selected sites could prove as ideal candidates for mutagenesis studies in the future.

Mismatch repair endonuclease PMS2 (postmeiotic segregation increased 2) is a component of the post-replicative DNA mismatch repair system [78]. Defects in PMS2 are reported in HNPCC [76]. The PMS2 dataset contained 21 taxa and site-specific analysis identified 8.9% of sites under positive selection in this PMS2 protein, $\omega = 1.29$ (Table 2). Functional relevance of these sites was determined by comparison to Human Swiss-Prot sequence (P54278). The vast majority of sites (32) reside within the 430–645 region of the alignment. This region of the alignment is highly variable and could not be not improved manually. Functional characterization for this region is also lacking and therefore we could not assess functional relevance. Outside this region, two positively selected sites, 402 and 406 (PP = 0.632 and 0.728 respectively) flank a phosphoserine modification site (403) [79]. Both substitutions are radical and could affect the function at position 403.

TP53 (cellular tumor antigen p53) acts as a tumor suppressor by inducing apoptosis or arresting growth depending on the physiological circumstances and cell type [80]. The TP53 protein (P04637) is 393 residues in length with 343 of these sites reported as natural variants that cause/lead to cancer including but not limited to colorectal and gastric cancers [41,81,82]. In our analysis of TP53 we have 16 taxa. Mutations in this gene radically affect function and therefore we would expect to find evidence of strong purifying selection across sites and lineages. However, results indicate that site specific positive selection is at work with 13 sites are under positive selection, $\omega = 1.97$. See Figure 2(d) and Table 2 for detailed analysis. On inspection of these 14 sites, we determine that 11 are located within the first region of the protein (positions 1–83), a region responsible for interaction with the methyltransferase HRMT1L2 and the recruiting of promoters to the TP53 gene [83]. We identified a cluster of positively selected sites, namely positions 46 and 47, along with an additional 7 sites within ten residues 39, 52, 53, 54, 55, 56, and 59 (see Additional file 4). Mutation of position 46 can abolish phosphorylation by HIPK2 and acetylation of K-382 by CREBBP [84]. Region 66–110 of TP53 is involved in interaction with WWOX protein and we have identified two sites (Swiss-Prot positions: 72 and 81), under positive selection within this region. Positively selected position 129, is located within a region reported to interact with HIPK1 (100–370) and AXIN1 (116–292), and in addition is also located within a region (positions 113–236) that is required for interaction with FBX042. Positively selected residue 355 is located within the CARM1 interaction region (300–393), the HIPK2 interacting region (319–360), and the oligomerization region (325–356).

Conclusion

The results we have presented are indicative of selective pressures acting in a lineage-specific manner. The positively selected sites we have identified in this study frequently reside in regions of functional importance, such as glycosylation sites, protease cleavage sites, and sites known to interact with proteins involved in DNA damage repair pathways. Also of note, positively selected residues are frequently located at-, or in close proximity to-, known cancer associated sites although the statistical significance of these coincidences cannot be concluded with such a small sample sizes. Larger sample sizes and more complete functional

information will be hugely beneficial in resolving whether these positively selected residues are most likely positioned to or at variants associated with cancer.

In using the mouse as a model organism for colon cancer, we are making an assumption that the orthologs in both species are functioning in precisely the same way despite ~ 180 million years of independent evolution. We found no evidence of functional divergence in the extant human and mouse lineages for the genes analyzed. However, upon testing the lineages leading from the MRCA of mouse and human, i.e. Euarchontoglires, positive selection has occurred on certain branches and in specific lineages. In the ancestral lineages from the divergence of primates, rodents and glires there is evidence of positive selection in 6 of the 22 genes tested (this includes the VHL result but as from Table 2 it is clear that this is a weak result). In total, considering all lineages analyzed including extant lineages, we have detected lineage-specific positive selection in 64% of the genes analyzed (i.e. 14/22 genes). Studies on the levels of polymorphism observed in *Drosophila* species indicate that positive selection is pervasive in this species with positive selection present in ~25% of the genes [85]. Previous studies on the levels of positive selection in primates compared to rodents and in the Hominidae reveal much lower levels of positive selection in the range of 5-9% of genes in the genome [7,8]. If these previous analyses were to act as a measurement of expectation then we should have identified only 1 gene under positive selection in this dataset that is comprised of mammals for the most part (taking the Drosophila data as the upper bound we would expect in the region of 6 genes with evidence of positive selection).

On grouping the cancer associated genes according to their involvement in functional pathways we determined that the MMR DNA damage response pathway has evidence of positive selection in 3 components of the pathway – 2 of which are site-specific and one of which is specific to the ancestral Murinae lineage suggesting a specific selective pressure in this clade for this process. The site-specific analyses identified a total of 5 genes that are positively selected: CDH1, MUTYH, PMS1, PMS2 and TP53. These results are important for contributing to our understanding of fundamental functions of these proteins and have provided potential targets for rational mutagenesis.

Overall, these results indicate that the function of certain proteins associated with colon cancer display distinct lineage-specific patterns of substitution indicative of positive selection in the ancestral human and mouse lineages. There are a number of selective pressures on any given protein that can contribute to patterns of substitution that are "*falsely"* indicative of positive selection. The necessity to continue to interact with protein partners may be a strong driving the evolution of the proteins in this study as many form functional complexes with one another or other proteins [86]. Compensatory mutations may also contribute to elevated levels of ω [87]. The effective population size (N_e) of the species tested vary enormously, with estimations for modern human populations in the range of $N_e = 7{,}500$ to 3,100 [88], while estimations for modern mouse populations range from $N_e = 58{,}000$ to 25,000 [89] and this large difference in N_e may also contribute to detection of false positives. We have also detected weak evidence for ongoing selective pressure in the human genome on the STK11 and CDH1, but these signals of selection may be artifacts of the very small effective population size of modern humans. Smaller N_e values are associated with increased fixation of slightly deleterious substitutions and subsequent elevated ω values [90]. Such slightly deleterious mutations in turn can lead to additional compensatory substitutions that become fixed. Teasing apart substitutions that have become fixed due to positive selection from slightly deleterious substitutions fixed due to small N_e *[*91] will aid in a more complete understanding of protein evolution in the future.

Abbreviations

BEB, Bayes empirical bayes; *dN*, Nonsynonymous substitutions per nonsynonymous site; *dS*, Synonymous substitutions per synonymous sites; LRT, Likelihood ratio test; ML, Maximum likelihood; MY, Millions of years; N_e, Effective population size; NEB, Naïve empirical bayes; PP, Posterior probability

Competing interests

The authors declare no conflict of interest.

Authors' contributions

CCM and KS carried out all data assembly. KS, CCM and AEW carried out all homolog identification and MSAs. CCM carried out all data quality and phylogeny analyses. CCM, KS, AEW, and TAW carried out all selective pressure analyses and designed the necessary software. ML carried out all structural analyses. All authors participated in drafting the manuscript. MJO'C conceived of the study, its design and coordination and drafted the manuscript. All authors read and approved the final manuscript.

Acknowledgements

We would like to thank the Irish Research Council for Science, Engineering and Technology (Embark Initiative Postgraduate Scholarship CCM) and DCU O'Hare scholarship (KS) for financial support. CCM is funded by the Irish Research Council for Science, Engineering and Technology (Embark Initiative Postgraduate scholarship RS2000172 to CCM). TAW is funded by School of Biotechnology and Pierse Trust Scholarships at DCU. MJO'C and AEW are funded by Science Foundation Ireland RFP: EOB2673. We would like to thank the SFI/HEA Irish Centre for High-End Computing (ICHEC) for processor time and technical support for both phylogeny reconstruction and selective pressure analyses. We would like to thank the SCI-SYM centre at DCU for processor time.

References

1. Waterston RH, Lindblad-Toh K, *et al*: **Initial sequencing and comparative analysis of the mouse genome.** *Nature* 2002, **420**(6915):520–562.

2. Benton MJ, Donoghue PC: **Paleontological evidence to date the tree of life.** *Mol Biol Evol* 2007, **24**(1):26–53.

3. Hirano R, Interthal H, *et al*: **Spinocerebellar ataxia with axonal neuropathy: consequence of a Tdp1 recessive neomorphic mutation?** *EMBO J* 2007, **26**(22):4732–4743.

4. Gao L, Zhang J: **Why are some human disease-associated mutations fixed in mice?** *Trends Genet* 2003, **19**(12):678–681.

5. Hakem R, de la Pompa JL, *et al*: **The tumor suppressor gene Brca1 is required for embryonic cellular proliferation in the mouse.** *Cell* 1996, **85**(7):1009–1023.

6. MacColl AD: **The ecological causes of evolution.** *Trends Ecol Evol* 2011, **26**(10):514–522.

7. Arbiza L, Dopazo J, *et al*: **Positive selection, relaxation, and acceleration in the evolution of the human and chimp genome.** *PLoS Comput Biol* 2006, **2**(4):e38.

8. Kosiol C, Vinar T, *et al*: **Patterns of positive selection in six Mammalian genomes.** *PLoS Genet* 2008, **4**(8):e1000144.

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

10. Schmid K, Yang Z: **The trouble with sliding windows and the selective pressure in BRCA1.** *PLoS One* 2008, **3**(11):e3746.

11. Bush RM: **Predicting adaptive evolution.** *Nat Rev Genet* 2001, **2**(5):387–392.

12. Wong WS, Yang Z, *et al*: **Accuracy and power of statistical methods for detecting adaptive evolution in protein coding sequences and for identifying positively selected sites.** *Genetics* 2004, **168**(2):1041–1051.

13. Anisimova M, Nielsen R, *et al*: **Effect of recombination on the accuracy of the likelihood method for detecting positive selection at amino acid sites.** *Genetics* 2003, **164**(3):1229–1236.

14. Levasseur A, Gouret P, *et al*: **Tracking the connection between evolutionary and functional shifts using the fungal lipase/feruloyl esterase A family.** *BMC Evol Biol* 2006, **6:**92.

15. Moury B, Simon V: **dN/dS-based methods detect positive selection linked to trade-offs between different fitness traits in the coat protein of potato virus Y.** *Mol Biol Evol* 2011, **28**(9):2707–2717. (Published advance access March 28th 2012, page numbers not currently available): doi:10.1093/molbev/mss073.

16. Loughran NB, Hinde S, *et al*: **Functional consequence of positive selection revealed through rational mutagenesis of human myeloperoxidase.** *Mol Biol Evol* 2012.

17. Barat A, Ruskin HJ: **A manually curated novel knowledge management system for genetic and epigenetic molecular determinants of colon cancer.** *Open Colorectal Cancer J* 2010, **3:**36–46.

18. Ferlay JSH, Bray F, Forman D, Mathers C, Parkin DM: *GLOBOCAN 2008 v1.2, Cancer Incidence and Mortality Worldwide: IARC CancerBase No. 10 [Internet]*; 2008.

19. Strate LL, Syngal S: **Hereditary colorectal cancer syndromes.** *Cancer Causes Control* 2005, **16**(3):201–213.

20. Kosinski J, Hinrichsen I, *et al*: **Identification of Lynch syndrome mutations in the MLH1-PMS2 interface that disturb dimerization and mismatch repair.** *Hum Mutat* 2010, **31**(8):975–982.

21. Vilar E, Gruber SB: **Microsatellite instability in colorectal cancer-the stable evidence.** *Nat Rev Clin Oncol* 2010, **7**(3):153–162.

22. Kulesz-Martin M, Liu Y: **p53 protein at the hub of cellular DNA damage response pathways through sequence-specific and non-sequence-specific DNA binding.** *Oxford J* 2000, **22**(6):9.

23. Tudzarova S, Colombo SL, *et al*: **Two ubiquitin ligases, APC/C-Cdh1 and SKP1-CUL1-F (SCF)-beta-TrCP, sequentially regulate glycolysis during the cell cycle.** *Proc Natl Acad Sci U S A* 2011, **108**(13):5278–5283.

24. Vogelstein B, Kinzler KW: **Cancer genes and the pathways they control.** *Nature Medicine* 2004, **10**(8):789–799.

25. Futreal PA, Coin L, *et al*: **A census of human cancer genes.** *Nat Rev Cancer* 2004, **4**(3):177–183.

26. Hubbard T, Barker D, *et al*: **The Ensembl genome database project.** *Nucleic Acids Res* 2002, **30**(1):38–41.

27. Hubbard T, Andrews D, *et al*: **Ensembl 2005.** *Nucleic Acids Res* 2005, **33:**D447–D453.

28. Yoon KA, Ku JL, *et al*: **Germline mutations of E-cadherin gene in Korean familial gastric cancer patients.** *J Human Genet* 1999, **44**(3):177–180.

29. Lyon, France: International Agency for Research on Cancer; 2010. Available from: http://globocan.iarc.fr.

30. Chenna R, Sugawara H, *et al*: **Multiple sequence alignment with the Clustal series of programs.** *Nucleic Acids Res* 2003, **31**(13):3497–3500.

31. Larkin MA, Blackshields G, *et al*: **Clustal W and Clustal X version 2.0.** *Bioinformatics* 2007, **23**(21):2947–2948.

32. Rambaut A: *Se-AL Sequence alignment editor*. Oxford: Software package; 1996.

33. Anisimova M, Bielawski JP, *et al*: **Accuracy and power of the likelihood ratio test in detecting adaptive molecular evolution.** *Mol Biol Evol* 2001, **18**(8):1585–1592.

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

35. Sawyer S: **Statistical tests for detecting gene conversion.** *Mol Biol Evol* 1989, **6**(5):526–538.

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

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

38. Murphy WJ, Eizirik E, *et al*: **Resolution of the early placental mammal radiation using Bayesian phylogenetics.** *Science* 2001, **294**(5550):2348–2351.

39. Nielsen R, Yang Z: **Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene.** *Genetics* 1998, **148**(3):929–936.

40. Loughran NB, O'Connor B, *et al*: **The phylogeny of the mammalian heme peroxidases and the evolution of their diverse functions.** *BMC Evol Biol* 2008, **8:**101.

41. UniProt: **Ongoing and future developments at the Universal Protein Resource.** *Nucleic Acids Res* 2011, **39**(Database issue):D214–219.

42. Voight BF, Kudaravalli S, *et al*: **A map of recent positive selection in the human genome.** *PLoS Biol* 2006, **4**(3):e72.

43. Boudeau J, Baas AF, *et al*: **MO25alpha/beta interact with STRADalpha/beta enhancing their ability to bind, activate and localize LKB1 in the cytoplasm.** *EMBO J* 2003, **22**(19):5102–5114.

44. Baas AF, Boudeau J, *et al*: **Activation of the tumour suppressor kinase LKB1 by the STE20-like pseudokinase STRAD.** *EMBO J* 2003, **22**(12):3062–3072.

45. Hemminki A, Markie D, *et al*: **A serine/threonine kinase gene defective in Peutz-Jeghers syndrome.** *Nature* 1998, **391**(6663):184–187.

46. Nakagawa H, Koyama K, *et al*: **Nine novel germline mutations of STK11 in ten families with Peutz-Jeghers syndrome.** *Hum Genet* 1998, **103**(2):168–172.

47. Dong SM, Kim KM, *et al*: **Frequent somatic mutations in serine/threonine kinase 11/Peutz-Jeghers syndrome gene in left-sided colon cancer.** *Cancer Res* 1998, **58**(17):3787–3790.

48. Westerman AM, Entius MM, *et al*: **Novel mutations in the LKB1/STK11 gene in Dutch Peutz-Jeghers families.** *Hum Mutat* 1999, **13**(6):476–481.

49. Berx G, Becker KF, *et al*: **Mutations of the human E-cadherin (CDH1) gene.** *Hum Mutat* 1998, **12**(4):226–237.

50. Latif F, Tory K, *et al*: **Identification of the von Hippel-Lindau disease tumor suppressor gene.** *Science* 1993, **260**(5112):1317–1320.

51. Kolodner RD, Tytell JD, *et al*: **Germ-line msh6 mutations in colorectal cancer families.** *Cancer Res* 1999, **59**(20):5068–5074.

52. Ohmiya N, Matsumoto S, *et al*: **Germline and somatic mutations in hMSH6 and hMSH3 in gastrointestinal cancers of the microsatellite mutator phenotype.** *Gene* 2001, **272**(1–2):301–313.

53. Kariola R, Otway R, *et al*: **Two mismatch repair gene mutations found in a colon cancer patient–which one is pathogenic?** *Hum Genet* 2003, **112**(2):105–109.

54. Berends MJ, Wu Y, *et al*: **Molecular and clinical characteristics of MSH6 variants: an analysis of 25 index carriers of a germline variant.** *Am J Human Genet* 2002, **70**(1):26–37.

55. Ollila S, Dermadi Bebek D, *et al*: **Mechanisms of pathogenicity in human MSH2 missense mutants.** *Hum Mutat* 2008, **29**(11):1355–1363.

56. Tee AR, Fingar DC, *et al*: **Tuberous sclerosis complex-1 and −2 gene products function together to inhibit mammalian target of rapamycin (mTOR)-mediated downstream signaling.** *Proc Natl Acad Sci U S A* 2002, **99**(21):13571–13576.

57. Kishi S, Zhou XZ, *et al*: **Telomeric protein Pin2/TRF1 as an important ATM target in response to double strand DNA breaks.** *J Biol Chem* 2001, **276**(31):29282–29291.

58. Blackwell LJ, Bjornson KP, *et al*: **DNA-dependent activation of the hMutSalpha ATPase.** *J Biol Chem* 1998, **273**(48):32049–32054.

59. Blackwell LJ, Martik D, *et al*: **Nucleotide-promoted release of hMutSalpha from heteroduplex DNA is consistent with an ATP-dependent translocation mechanism.** *J Biol Chem* 1998, **273**(48):32055–32062.

60. Wu Y, Berends MJ, *et al*: **A role for MLH3 in hereditary nonpolyposis colorectal cancer.** *Nat Genet* 2001, **29**(2):137–138.

61. Plaschke J, Kruger S, *et al*: **Eight novel MSH6 germline mutations in patients with familial and nonfamilial colorectal cancer selected by loss of protein expression in tumor tissue.** *Hum Mutat* 2004, **23**(3):285.

62. Niemann S, Muller U: **Mutations in SDHC cause autosomal dominant paraganglioma, type 3.** *Nat Genet* 2000, **26**(3):268–270.

63. Bronner CE, Baker SM, *et al*: **Mutation in the DNA mismatch repair gene homologue hMLH1 is associated with hereditary non-polyposis colon cancer.** *Nature* 1994, **368**(6468):258–261.

64. Pensotti V, Radice P, *et al*: **Mean age of tumor onset in hereditary nonpolyposis colorectal cancer (HNPCC) families correlates with the presence of mutations in DNA mismatch repair genes.** *Genes Chromosomes Cancer* 1997, **19**(3):135–142.

65. Kurzawski G, Suchy J, *et al*: **Germline MSH2 and MLH1 mutational spectrum including large rearrangements in HNPCC families from Poland (update study).** *Clin Genet* 2006, **69**(1):40–47.

66. Tournier I, Vezain M, *et al*: **A large fraction of unclassified variants of the mismatch repair genes MLH1 and MSH2 is associated with splicing defects.** *Hum Mutat* 2008, **29**(12):1412–1424.

67. Kim JC, Kim HC, *et al*: **hMLH1 and hMSH2 mutations in families with familial clustering of gastric cancer and hereditary non-polyposis colorectal cancer.** *Cancer Detect Prev* 2001, **25**(6):503–510.

68. Nickerson ML, Warren MB, *et al*: **Mutations in a novel gene lead to kidney tumors, lung wall defects, and benign tumors of the hair follicle in patients with the Birt-Hogg-Dube syndrome.** *Cancer Cell* 2002, **2**(2):157–164.

69. Sayed MG, Ahmed AF, *et al*: **Germline SMAD4 or BMPR1A mutations and phenotype of juvenile polyposis.** *Ann Surg Oncol* 2002, **9**(9):901–906.

70. Sjoblom T, Jones S, *et al*: **The consensus coding sequences of human breast and colorectal cancers.** *Science* 2006, **314**(5797):268–274.

71. Hahn SA, Schutte M, *et al*: **DPC4, a candidate tumor suppressor gene at human chromosome 18q21.1.** *Science* 1996, **271**(5247):350–353.

72. Ballester R, Marchuk D, *et al*: **The NF1 locus encodes a protein functionally related to mammalian GAP and yeast IRA proteins.** *Cell* 1990, **63**(4):851–859.

73. Tamura G, Sakata K, *et al*: **Inactivation of the E-cadherin gene in primary gastric carcinomas and gastric carcinoma cell lines.** *Jpn J Cancer Res* 1996, **87**(11):1153–1159.

74. Marambaud P, Shioi J, *et al*: **A presenilin-1/gamma-secretase cleavage releases the E-cadherin intracellular domain and regulates disassembly of adherens junctions.** *EMBO J* 2002, **21**(8):1948–1956.

75. Zhou F, Su J, *et al*: **Unglycosylation at Asn-633 made extracellular domain of E-cadherin folded incorrectly and arrested in endoplasmic reticulum, then sequentially degraded by ERAD.** *Glycoconj J* 2008, **25**(8):727–740.

76. Nicolaides NC, Papadopoulos N, *et al*: **Mutations of two PMS homologues in hereditary nonpolyposis colon cancer.** *Nature* 1994, **371**(6492):75–80.

77. Wang Q, Lasset C, *et al*: **Prevalence of germline mutations of hMLH1, hMSH2, hPMS1, hPMS2, and hMSH6 genes in 75 French kindreds with nonpolyposis colorectal cancer.** *Hum Genet* 1999, **105**(1–2):79–85.

78. Sacho EJ, Kadyrov FA, *et al*: **Direct visualization of asymmetric adenine-nucleotide-induced conformational changes in MutL alpha.** *Mol Cell* 2008, **29**(1):112–121.

79. Beausoleil SA, Villen J, *et al*: **A probability-based approach for high-throughput protein phosphorylation analysis and site localization.** *Nat Biotechnol* 2006, **24**(10):1285–1292.

80. Guo A, Salomoni P, *et al*: **The function of PML in p53-dependent apoptosis.** *Nat Cell Biol* 2000, **2**(10):730–736.

81. Varley JM, McGown G, *et al*: **An extended Li-Fraumeni kindred with gastric carcinoma and a codon 175 mutation in TP53.** *J Med Genet* 1995, **32**(12):942–945.

82. Guran S, Tunca Y, *et al*: **Hereditary TP53 codon 292 and somatic P16INK4A codon 94 mutations in a Li-Fraumeni syndrome family.** *Cancer Genet Cytogenet* 1999, **113**(2):145–151.

83. An W, Kim J, *et al*: **Ordered cooperative functions of PRMT1, p300, and CARM1 in transcriptional activation by p53.** *Cell* 2004, **117**(6):735–748.

84. Hofmann TG, Moller A, *et al*: **Regulation of p53 activity by its interaction with homeodomain-interacting protein kinase-2.** *Nat Cell Biol* 2002, **4**(1):1–10.

85. Bierne N, Eyre-Walker A: **The genomic rate of adaptive amino acid substitution in Drosophila.** *Mol Biol Evol* 2004, **21**(7):1350–1360.

86. Fraser HB, Hirsh AE, *et al*: **Evolutionary rate in the protein interaction network.** *Science* 2002, **296**(5568):750–752.

87. Lunzer M, Golding GB, *et al*: **Pervasive cryptic epistasis in molecular evolution.** *PLoS Genet* 2010, **6**(10):e1001162.

88. Tenesa A, Navarro P, *et al*: **Recent human effective population size estimated from linkage disequilibrium.** *Genome Res* 2007, **17**(4):520–526.

89. Salcedo T, Geraldes A, *et al*: **Nucleotide variation in wild and inbred mice.** *Genetics* 2007, **177**(4):2277–2291.

90. Eyre-Walker A, Keightley PD, *et al*: **Quantifying the slightly deleterious mutation model of molecular evolution.** *Mol Biol Evol* 2002, **19**(12):2142–2149.

91. Eyre-Walker A, Keightley PD: **Estimating the rate of adaptive molecular evolution in the presence of slightly deleterious mutations and population size change.** *Mol Biol Evol* 2009, **26**(9):2097–2108.

Additional files

Additional_file_1 as DOC

Additional file 1 Details of the data used in the analysis, the 21 species and their genome coverage. Orthologs that were not found by the Ensembl genome browser are labeled in black, orthologs identified are shown in white.

Additional_file_2 as DOC

Additional file 2 Complete set of all multiple sequence alignments used in the analysis. The data is presented on a gene-by-gene basis in nexus format.

Additional_file_3 as DOC

Additional file 3 Likelihood ratio tests performed and their associated significance values. (DOC 31 kb)

Additional_file_4 as DOC

Additional file 4 Full set of models, associated likelihood scores and parameter estimates for all genes in the colon cancer gene dataset. This information is given alphabetically on a gene-by-gene basis. All estimated parameters, Likelihood values and BEB or NEB sites are listed.

Additional_file_5 as DOC

Additional file 5 Full set of recombination test results on a per gene and per species basis. The value highlighted in yellow for TP53 represents a region where recombination was detected with reasonable confidence that also coincided with a positively selected residue (i.e. false positive).

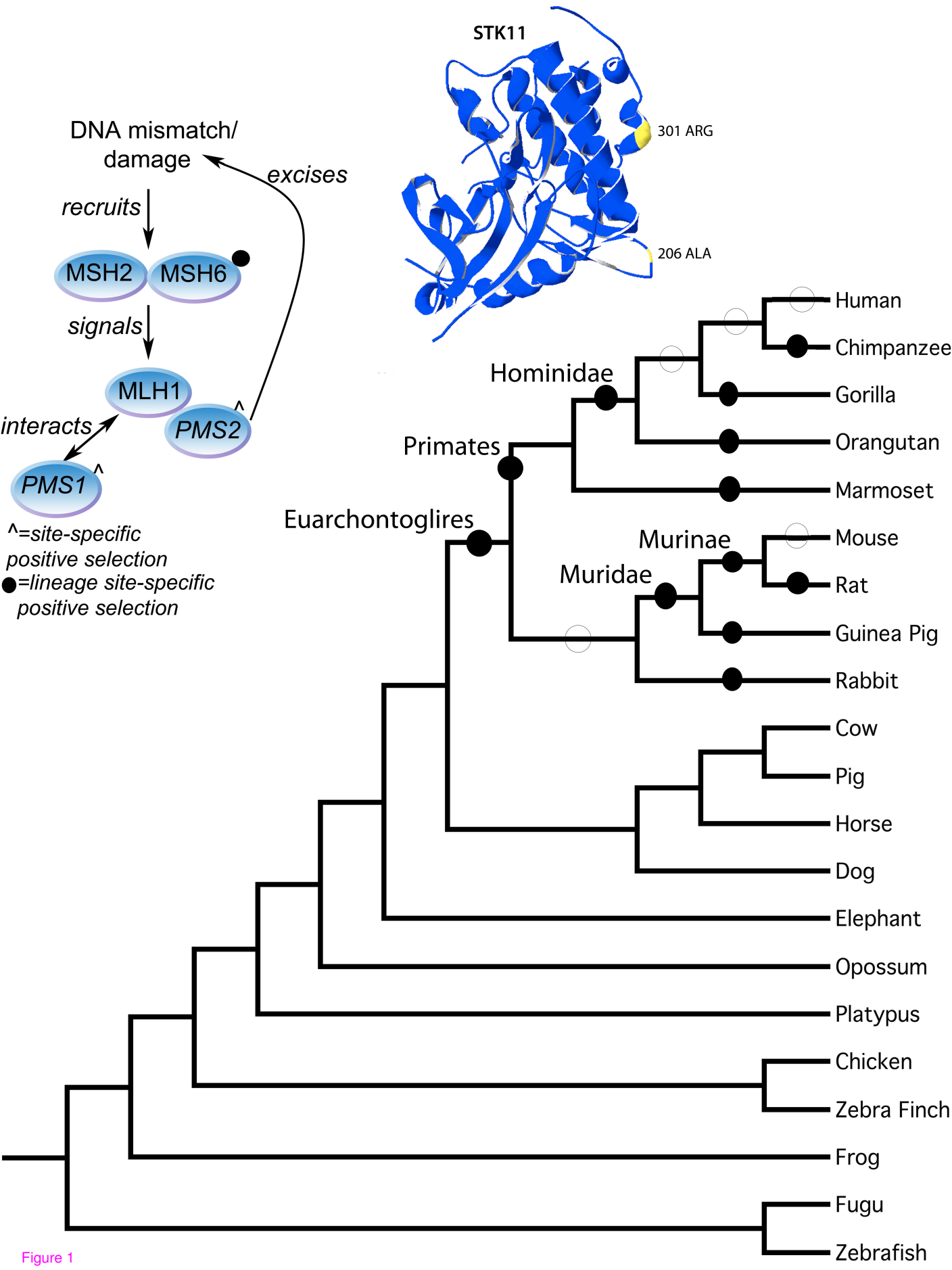

STK11
301 ARG
206 ALA
DNA mismatch/
damage
excises
recruits
MSH2
MSH6
signals
MLH1
PMS2
interacts
PMS1
^=site-specific
positive selection
●=lineage site-specific
positive selection
Human
Chimpanzee
Hominidae
Gorilla
Orangutan
Primates
Marmoset
Euarchontoglires
Murinae
Mouse
Muridae
Rat
Guinea Pig
Rabbit
Cow
Pig
Horse
Dog
Elephant
Opossum
Platypus
Chicken
Zebra Finch
Frog
Fugu
Zebrafish

Figure 1

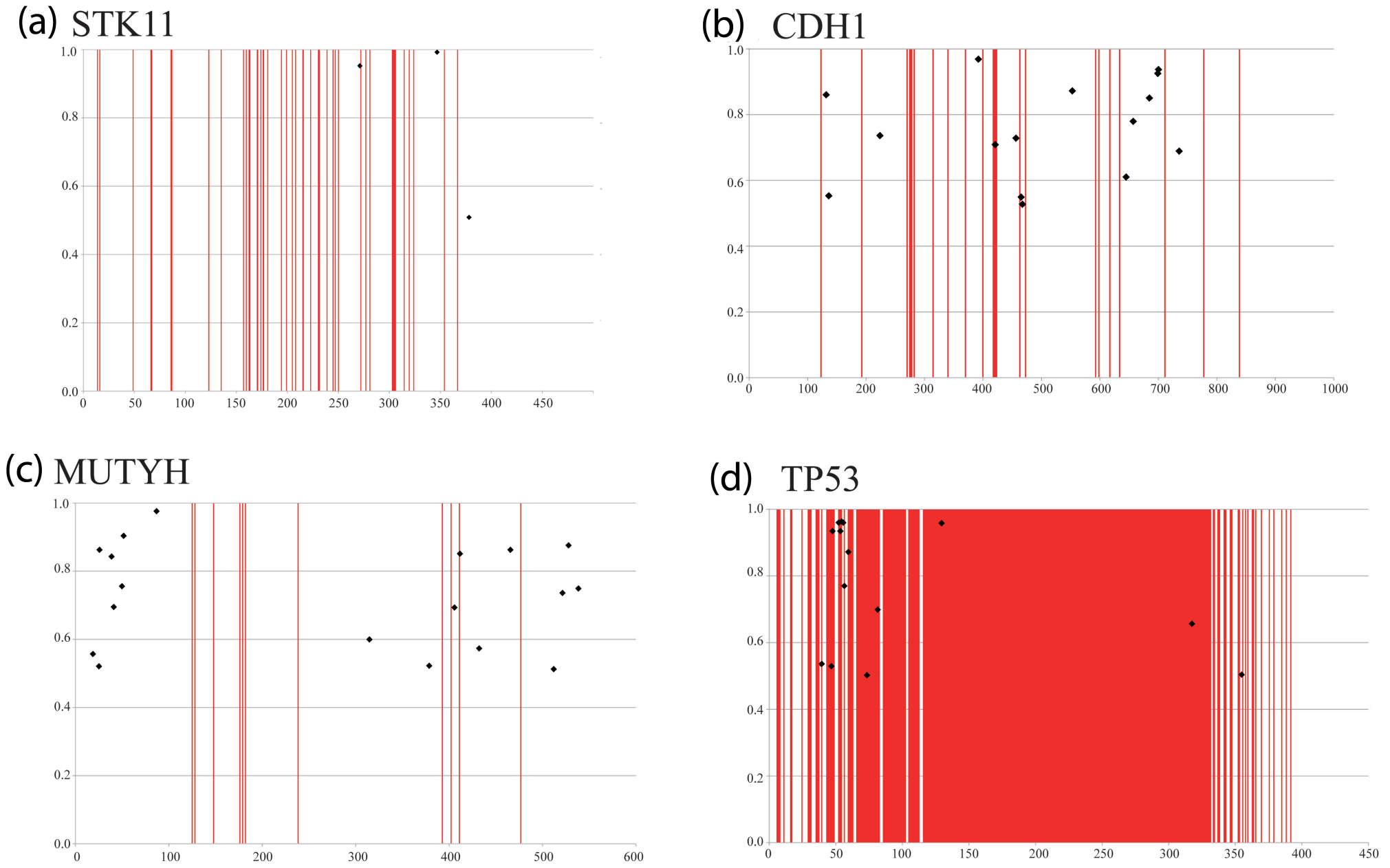

(a) STK11
1.0
0.8
0.6
0.4
0.2
0.0
0
50
100
150
200
250
300
350
400
450
(b) CDH1
1.0
0.8
0.6
0.4
0.2
0.0
0
100
200
300
400
500
600
700
800
900
1000
(c) MUTYH
1.0
0.8
0.6
0.4
0.2
0.0
0
100
200
300
400
500
600
(d) TP53
1.0
0.8
0.6
0.4
0.2
0.0
0
50
100
150
200
250
300
350
400
450

Figure 2

Additional files provided with this submission:

Additional file 1: Supplementary_File1.doc, 100K
http://www.biomedcentral.com/imedia/1626758213670556/supp1.doc
Additional file 2: Supplementary_File2.doc, 4124K
http://www.biomedcentral.com/imedia/5600110446705566/supp2.doc
Additional file 3: Supplementary_File3.doc, 31K
http://www.biomedcentral.com/imedia/8265080006705566/supp3.doc
Additional file 4: Supplementary_File4.doc, 1417K
http://www.biomedcentral.com/imedia/1531824470721031/supp4.doc
Additional file 5: Supplementary_File5.doc, 162K
http://www.biomedcentral.com/imedia/1247687687721032/supp5.doc