The evolution of the mammal placenta — a computational approach to the identification and analysis of placenta-specific genes and microRNAs.

by

Thomas A. Walsh, M.Sc. Bioinformatics,
B.Sc. (Hons) Computer Games Technology

Dublin City University

School of Biotechnology

A thesis presented to Dublin City University for the Degree of Doctor of Philosophy

Supervisor:
Dr Mary J. O'Connell

June 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: ____________________ (Thomas Walsh)
ID No.: ____________________
Date: ____________________

For all my parents

Acknowledgements

First and foremost, I would like to thank Dr Mary O'Connell for her enthusiasm, guidance and patience throughout my time in the BME lab. Thank you for hanging on even by the thinnest of threads to see things through.

This research was done with the support of the School of Biotechnology Pierse Trust Scholarship, for which I am very appreciative. I am also thankful to Westmeath County Council for financial assistance during the first three years of my PhD.

I am grateful to all in the BME lab for the exchange of banter and ideas, and for the occasional moment of inspiration. My heartfelt appreciation goes out to all those in the wider community of the School of Biotechnology and of DCU who have always made me feel welcome over the years. One research group in particular deserves an honourable mention, but I must refrain, as it would make the others feel ba-a-a-a-ad.

I hold eternal gratitude for the help and support of family, friends and the many people in my life who have shown great generosity of spirit. I am especially thankful to those who encouraged me to complete my formal education at a time when I had other ideas. You will be happy to know that it is finally complete.

Table of Contents

List of Figures

List of Tables

List of Equations

List of Abbreviations

A (amino acid)	Alanine
A (nucleotide)	Adenine
ADAM12	ADAM metallopeptidase domain 12
Adh	Alcohol dehydrogenase
ADM	Adrenomedullin
Ago	Argonaute
Ago2	Argonaute-2
AIC	Akaike information criterion
ALB	Albumin
AMOSA	A simulated annealing based multi-objective optimization algorithm
Arg	Arginine
AUC	Area Under the ROC Curve
BBH	Bidirectional Best Hit
BEB	Bayes empirical Bayes
BIC	Bayesian information criterion
BLAST	Basic Local Alignment Search Tool
BLOSUM	BLOcks of Amino Acid SUbstitution Matrix
C (amino acid)	Cysteine
C (nucleotide)	Cytosine
C19MC	Chromosome 19 microRNA cluster
CAR	Correlation Adjusted marginal coRrelation
CCBP2	Chemokine binding protein 2
CCNE1	Cyclin E1

CCNT2	Cyclin T2
CDS	Coding DNA sequence
CSF3R	Colony stimulating factor 3 receptor
ΔAUC	Difference in area under the ROC curve
ΔG	Differential of Gibbs free energy
D	Aspartic acid
DCS	Duplication Consistency Score
DGCR8	DiGeorge syndrome critical region gene 8
DIC	Deviance information criterion
DiF	Dinucleotide frequency
Dn	Rate of non-synonymous substitution per non-synonymous site
DNA	Deoxyribonucleic Acid
Ds	Rate of synonymous substitution per synonymous site
E	Glutamic acid
ERF	Ets2 repressor factor
ESX1	ESX homeobox 1
Exp5	Exportin-5
F	Phenylalanine
F81 model	Felsenstein 1981 model
FBN2	Fibrillin 2
FBXO31	F-box only protein 31
FPR	False-positive rate
G (amino acid)	Glycine
G (nucleotide)	Guanine

GC3	GC content in the third codon position
Gly	Glycine
GO	Gene ontology
GTF	General Transcription Factor
GTR model	General Time Reversible model
H	Histidine
HGNC	HUGO Gene Nomenclature Committee
HITS-CLIP	High-throughput sequencing of RNA isolated by crosslinking immunoprecipitation
HKY85	Hasegawa, Kishino and Yano model
hLRT	Hierarchical likelihood ratio test
HoT	Heads or Tails method
HPC	High-Performance Computing
HQ	Hannan-Quinn information criterion
I	Isoleucine
ICHEC	Irish Centre for High-End Computing
INSL4	Insulin-like 4
ISM2	Isthmin 2
JC69 model	Jukes and Cantor model
JTT	Jones, Taylor and Thornton model
K	Lysine
K2P model	Kimura 2-Parameter model
K3P model	Kimura 3-Parameter model
KISS1	Kisspeptin
L	Leucine

LBA	Long branch attraction
Ldbr	Lariat debranching
LEP	Leptin
LG	Le and Gascuel model
LM	Likelihood Mapping
lnL	Log-likelihood
LRT	Likelihood Ratio Test
μ	Mutation rate
M	Methionine
MAICE	Minimum Akaike Information Criterion Estimate
MCMC	Markov chain Monte Carlo
MET (gene)	MET proto-oncogene
Met (amino acid)	Methionine
MFE	Minimum Free Energy
miRISC	microRNA-induced silencing complex
miRNA	microRNA
miTI	microRNA-Target Interaction
miTP	microRNA-Target Prediction
ML	Maximum Likelihood
MLE	Maximum Likelihood Estimation
MRCA	Most Recent Common Ancestor
MRE	microRNA Recognition Element
mRNA	Messenger RNA
MSA	Multiple Sequence Alignment

MSH	Melanocyte-stimulating hormone
MYA	Millions of Years Ago
ν	Degrees of freedom
N (amino acid)	Asparagine
N (nucleotide)	Unspecified nucleotide
NC	Non-chromosomal
NCBI	National Center for Biotechnology Information
NCOA6	Nuclear receptor coactivator 6
N_e	Effective population size
NEB	Naïve empirical Bayes
OMIM	Online Mendelian Inheritance in Man
ORF	Open Reading Frame
OTU	Operational Taxonomic Unit
P (amino acid)	Proline
P (statistic)	P-value
PAGE4	P antigen family member 4
PAML	Phylogenetic Analysis by Maximum Likelihood
PAPPA	Pregnancy-associated plasma protein A, pappalysin 1
PAPS	3′-Phosphoadenosine-5′-phosphosulfate
PCR	Polymerase Chain Reaction
PEG10	Paternally expressed 10
PEM	Preferential Expression Measure
PHLDA2	Pleckstrin homology-like domain, family A, member 2
PLAC1	Placenta-specific 1

PP	Posterior Probability
pre-miRNA	Precursor microRNA
pri-miRNA	Primary microRNA
PROCR	Protein C receptor, endothelial
PSG	Pregnancy specific beta-1-glycoprotein
pSILAC	Pulsed SILAC
PSS	Positively selected site
Q	Glutamine
QC	Quality control
R	Arginine
RISC	RNA-induced silencing complex
RNA	Ribonucleic Acid
ROC	Receiver Operating Characteristic
RSD	Reciprocal Smallest Distance
RTL1	Retrotransposon-like 1
S	Serine
s	Selection coefficient
Ser	Serine
SILAC	Stable isotope labelling with amino acids in cell culture
siRNA	Short interfering RNA
SPM	Tissue-specificity Measure
SPS	Seed-Pairing Stability
SULT1E1	Sulfotransferase family 1E, estrogen-preferring, member 1
SVM	Support Vector Machine

T (amino acid)	Threonine
T (nucleotide)	Thymine
TA	Target Abundance
TAC3	Tachykinin 3
Trp	Tryptophan
TPR	True-positive rate
TSV	Tab-Separated Variable
UTR	Untranslated region
v	Version
V	Valine
Val	Valine
VT	Müller and Vingron model
ω	Ratio of non-synonymous substitutions per non-synonymous site to synonymous substitutions per synonymous site
W	Tryptophan
WAG	Whelan and Goldman model
WC	Watson-Crick complementarity
WEKA	Waikato Environment for Knowledge Analysis
WGD	Whole-Genome Duplication
X	Unspecified amino acid
XAGE2	X antigen family, member 2
XAGE3	X antigen family, member 3
Y	Tyrosine
ZFP36L1	ZFP36 ring finger protein-like 1

"Why a four-year-old child could understand this report.
Run out and find me a four-year-old child."
— Groucho Marx

"In all scientific research, the researcher may or may not find what he is looking for — indeed, his hypothesis may be demolished — but he is certain to learn something new...which may be and often is more important than what he had hoped to learn."
— Robert A. Heinlein

Abstract

The presence of a placenta is an important synapomorphy that defines the mammal clade. From the fossil record we know that the first placental mammal lived approximately 125 million years ago, with the chorioallantoic placenta evolving not long after. In this thesis a set of 22 complete genomes from Eutherian, non-Eutherian and outgroup species are compared, the aim being to identify protein-coding and regulatory alterations that are likely to be implicated in the emergence of mammal placenta in the fossil record. To this end we have examined the roles played by positive selection and miRNA regulation in the evolution of the placenta. We have identified those genes that underwent functional shift uniquely in the ancestral placental mammal lineage and that are also heavily implicated in disorders of the placenta. Carrying out a thorough analysis of non-coding regions of the 22 genomes included in the study we identified a cohort of miRNAs that exist only in placental mammals. Many of the placenta related genes described above have multiple predicted "placenta-specific" miRNA binding sites. Together these results indicate a role for both adaptation in protein-coding regions and emergence of novel non-coding regulators in the origin and evolution of mammal placentation.

Chapter 1: Introduction

1.1 Thesis Overview

This thesis describes two parallel and complementary strands of investigation into the origin and evolution of the placenta in mammals: a study of the role played in the evolution of placental genes by positive selection for gene sequence change resulting in a functional shift, and an analysis of the extent of regulatory innovation due to novel miRNA-target interactions specific to the choriallantoic placenta of Eutherian mammals. Before addressing the latter, it was necessary to perform an ancillary investigation into the relative performance of available biological sequence based methods for identifying miRNA-target interactions, so as to identify the miRNA-target prediction method best able to recover known miRNA-target interactions and to distinguish targets from non-targets.

Chapter 2 describes the analysis of functional shift and selective pressure on placental genes in the ancestral Eutherian lineage, while Chapter 4 reports on the study of regulatory miRNA innovation in the Eutherian clade. Chapter 3 outlines the comparison of miRNA-target predictions from different methods, which guided the choice of miRNA-target prediction method used in Chapter 4.

Chapter 1 offers a general background treatment of the field of molecular evolution, while laying conceptual foundations for the analyses conducted in Chapters 2 to 4. Sections 1.2 through to 1.7 provide an overview of concepts underlying the methods and analyses throughout the thesis. However, Sections 1.6 and 1.7 — covering homology and phylogeny, respectively — are particularly important for understanding how sequences sharing a common ancestor can be, firstly, identified and compared, and secondly, incorporated into a dendritic model of descent with modification from a common ancestor.

Section 1.8 focuses on the analysis of selective pressure at the molecular level, and as such, underlies much of the work done in Chapter 2. Section 1.9 explores mechanisms of regulatory innovation, particularly the biogenesis and modes of action of miRNAs, the evolution of miRNAs and their targets, and methods used to predict miRNA targets. This section outlines key concepts underlying the work in Chapters 3 and 4 on comparing the accuracy of miRNA-target prediction methods and on miRNA-target interactions predicted to occur between placental miRNAs and placental genes.

Section 1.10 describes the evolutionary history, key features and observed variety of the choriallantoic placenta, providing some biological context to the work described throughout the thesis. Finally, Section 1.11 recapitulates the overall aim and structure of the thesis, leading into Chapters 2 to 4.

Chapters 5 and 6 explore the implications of the results obtained in Chapters 2 to 4, discuss the conclusions drawn from these results and outline areas that future work could build on the work in this thesis. The electronic appendix contained in the enclosed disc contains an electronic copy of this thesis, along with an electronic journal recording the progress of analyses, bioinformatics pipelines devised and implemented, scripts and modules developed, and output produced during the course of the work described herein.

1.2 Evolutionary Theory

Evolution is a process by which biological species arise and change, mediated by the change in heritable traits of the members of such species as they are passed from parents to offspring, or from one organism to another in the case of horizontal gene transfer (Syvanen 1985). For a given heritable trait that can take different forms in different individual organisms, the differential reproductive success of individuals in a population which possess that trait, determines the frequency of the given trait (Darwin 1859).

In eukaryotes such as those focused on in this thesis, the vast majority of heritable information is stored in molecules of deoxyribonucleic acid (DNA) in the nucleus of each cell. The unit of inheritance is the gene, which corresponds to a region of DNA that, when expressed, is transcribed to a ribonucleic acid (RNA) molecule in a sequence-specific manner. This RNA may then be translated into a protein that serves a function in the organism, or may be functional in its own right.

Over time, traits that are deleterious to an organism will tend to decrease in frequency and be removed from the population, while traits that are advantageous to an organism will tend to increase in frequency and become fixed in the population, in a process known as **natural selection** (Darwin 1859). A trait may also change in frequency due to random sampling effects by a process known as **genetic drift**, such that neutral mutations may become fixed or removed (Kimura 1968). This process plays a dominant role in evolution at the molecular level, according to the neutral theory of evolution (Kimura 1968). Depending on the effective population size of a species, this process can also result in a slightly advantageous trait being removed or a slightly deleterious trait being fixed in a process known as the nearly neutral theory (Ohta 1973, Ohta 1992). The process of fixation is therefore affected not only by the benefits or costs conferred on the organism by the trait itself, but also by other factors such as the size of the population within which the organism lives (see Section 1.4).

It is useful to draw a distinction between an organism's genotype (i.e. the organism's genetic information), and its phenotype (i.e. the observable traits of an organism) (Johannsen 1911). While heritable information is passed from one organism to another in the form of a genotype, natural selection acts on the phenotype of the organism, and only indirectly affects the frequency of a given gene in the population.

If a mutant allele arises that causes a change in its corresponding protein or functional RNA molecule, this can be considered to have changed the phenotype of the organism. However, if the phenotypic effect is too small to appreciably affect the survival rate of organisms carrying the mutant trait (i.e. the organisms' fitness), then it is effectively invisible to natural selection. The mutant is therefore selectively neutral, but may be fixed at random in accordance with the neutral and nearly neutral theories (Kimura 1968, Ohta 1973, Ohta 1992).Where a mutant trait affects the fitness of an organism, this effect is usually deleterious, and is selected against. On the rare occasion that a mutant trait has a beneficial effect on an organism's chance of survival, natural selection favours this form of the trait and it will tend to become fixed in the population (Darwin 1859). For example, Figure 1.1 shows the distribution of fitness effects in diploid yeast lines in which mutations were induced by exposure to the mutagenic agent ethylmethane sulphonate (Wloch *et al.* 2001). About 40% of the mutations shown in Figure 1.1 were lethal, with most of the remainder being either neutral or deleterious.

Mutation is therefore the mechanism by which novel traits are introduced to a species, while the fixation of selectively neutral and favoured mutations is the process by which a species adapts to its environment and those novel traits become characteristic of the species as a whole as they are swept through the population (Huxley 1942).

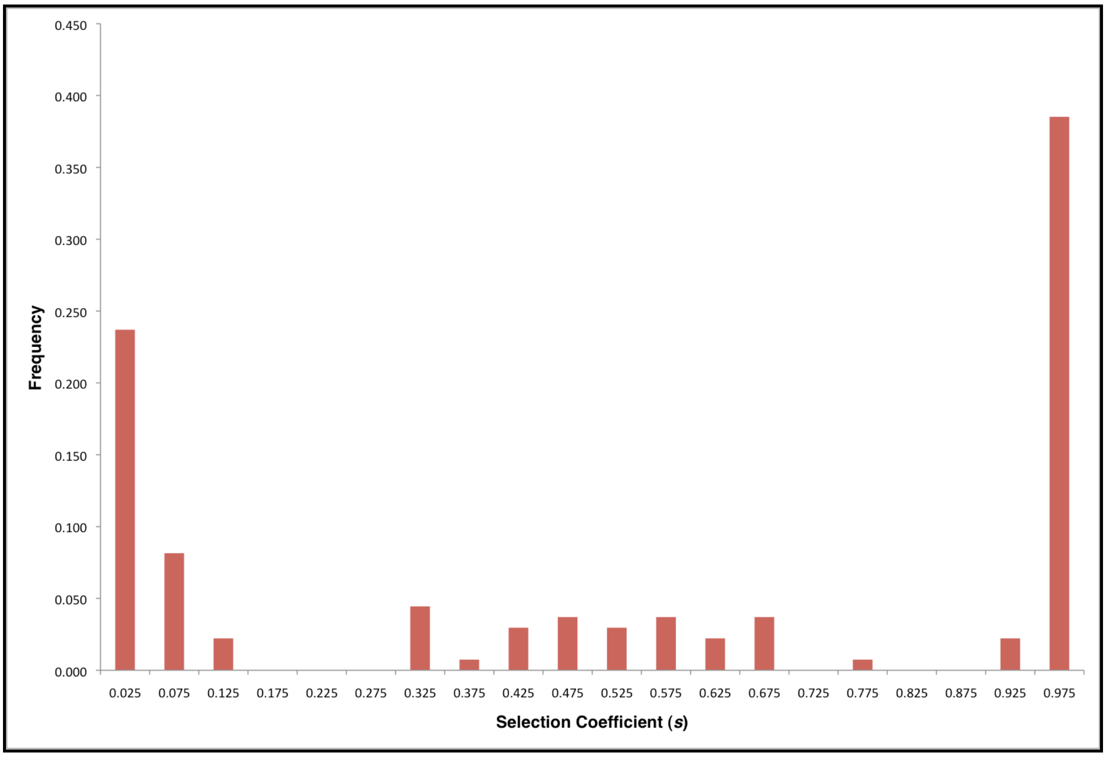


Figure 1.1: Distribution of fitness effects in yeast.

In a study by Wloch *et al.* (2001), yeast lines were exposed to a mutagen (i.e. ethylmethane sulphonate) and the fitness effects of the consequent mutations were observed. The horizontal axis shows different levels of fitness effect in terms of the selection coefficient *s* relative to the wild type, such that yeast lines in the leftmost category underwent mutations that were neutral or nearly so (i.e. $s \approx 0$), while all of those in the rightmost category suffered a lethal mutation (i.e. $s = 1$). Fitness was measured by relative growth of each yeast culture. Figure created using data from Wloch *et al.* (2001).

1.3 Mutation

Mutation of a genotype and the consequent alteration of its phenotype can occur in a number of ways — leading to an alteration in the protein-coding or regulatory features of the genome (Wray 2007). Mutational mechanisms include mutations affecting a small number of nucleotides within a gene, such as point mutations, insertions and deletions; mutations affecting entire genes or chromosomal segments, such as gene duplication and deletion, as well as gene fusion and fission; and whole-genome duplications, in which an organism's entire genome is duplicated (Eyre-Walker and Keightley 2007, Kaessmann 2010)

1.3.1 Point Mutation

A number of mechanisms can bring about a point mutation, and these mechanisms can be grouped into two general types: spontaneous and induced mutations (Patthy 2008). Spontaneous mutations may occur, for example, due to tautomerisation of amino and keto groups of bases, causing the formation of nonstandard base pairs, or through the deamination of cytosine to form uracil (eventually replaced by thymine) (Patthy 2008). An induced mutation occurs in the presence of a mutagenic agent. For example, nitrous acid can convert cytosine to uracil, which then pairs with adenine instead of guanine, or it can deaminate adenine to form hypoxanthine, which pairs with cytosine instead of thymine (Patthy 2008). In mammals, there is a phenomenon known as biased gene conversion, whereby the DNA repair machinery is biased towards the insertion of G and C in damaged DNA, thereby increasing the GC content in that region (Foster and Hickey 1999). Most point mutations are identified and corrected by the ‘proofreading’ activity of DNA polymerases, or by mismatch repair enzymes (Patthy 2008). However, some mutations are not identified by these correction mechanisms and persist in the organism (Patthy 2008). Point mutations may also introduce a new stop codon when a nonsense mutation occurs (Lewin 2008). A **transition** occurs when a purine mutates to another purine (i.e. an adenine mutates to a guanine, or *vice versa*), or when a pyrimidine mutates to another pyrimidine (i.e. a cytosine mutates to a thymine, or *vice versa*); while a **transversion** occurs when a purine (i.e. adenine or guanine) mutates to a pyrimidine (i.e. cytosine or thymine), or *vice versa* (Patthy 2008).

Not all DNA substitutions in protein-coding DNA bring about change at the amino acid level. Because of the degeneracy of the genetic code, most amino acids are coded for by multiple codons. A **synonymous mutation** is a change in the coding DNA sequence (CDS) of a gene that does not result in a changed amino acid sequence, whereas a **non-synonymous mutation** is a change in the coding sequence of a gene that does bring about a change at peptide level. However, not all synonymous mutations are without effect, as has been evidenced by bias in codon usage that constrains the set of codons used for a given amino acid (Akashi 1994, DeBry and Marzluff 1994, Kimchi-Sarfaty *et al.* 2007). It is believed that this codon usage bias is as a result of selection for more efficient transcription (Xia 1996). Other regions of the genome contain specific codons because they are splice sites and it is thought that codon usage biases in these regions may play a role in the mediation of alternative splicing (Abril *et al.* 2005). There are many examples in the literature of point mutations in protein-coding regions that result in altered phenotypes, a classic example being the melanocyte-stimulating hormone (MSH) receptor. Point mutations in this receptor result in variations in coat colour of mammals (Robbins *et al.* 1993). There is also a large body of evidence that point mutations in regulatory (non-coding) elements have important functional effects on phenotype. These included point mutations in the cis-regulatory regions of specific genes that have been shown to alter, for example, paternal care behaviour in rodents (Hammock and Young 2005), bristle patterning on the legs of adult fruit flies (Stern 1998) and obsessive compulsive behaviour in humans (Walitza *et al.* 2002), as reviewed recently by Wray *et al.* (2007).

1.3.2 Insertion and Deletion Mutation

Replication errors can also bring about an insertion or deletion mutation — also known as an **indel**. For example, replication slippage in the vicinity of short repeat sequences can result in the deletion or insertion of one or more nucleotides (Patthy 2008). Replication slippage in the 5′ to 3′ direction results in a deletion, while slippage in the 3′ to 5′ direction causes an insertion of the repeat sequence motif (Patthy 2008). The presence of an intercalating agent can also bring about an insertion mutation by increasing the distance between adjacent nucleotides in a DNA sequence (Brown 2002). In mammals, deletions are somewhat more common than insertions, and the majority of indels are short, involving the insertion or deletion of three codons or less (Taylor *et al.* 2004). For example, given a wild type nucleotide sequence 'GATTACACA', an insertion after the first codon of the nucleotides 'GAA' would result in the mutant sequence 'GATGAATACACA', while a deletion of the codon 'TAC' from the original sequence would result in 'GATACA'.

When the size of the insertion or deletion is a multiple of three bases, the indel is far less likely to disrupt the open reading frame of a gene sequence and correspondingly more likely to be accepted. For example, the wild type nucleotide sequence above would be translated as the amino acid sequence 'DYT', the example insertion mutant sequence would be translated into amino acid as 'DEYT', and the deletion mutant would be translated as 'DT'. The C-terminal threonine (T) and downstream amino acids are still translated from both mutant nucleotide sequences, because the reading frame is unaffected by the mutations. On the other hand, an indel that disrupts the reading frame — known as a frameshift mutation — will certainly change the downstream amino acid sequence, and is likely to bring a new stop codon into phase, which usually renders a protein non-functional. An example would be the deletion of the first cytosine (C) in the wild type example sequence, which would produce the sequence 'GATTAAACA', in turn resulting in the truncation of the protein product due to the newly introduced stop codon (shown in red). Because of this, such mutations are less likely to be accepted than in-frame indels, though it is possible for frameshift mutations to be accepted, as with a Flavobacterial oligomer hydrolase whose origin lies with a frameshift mutation in an existing gene (Ohno 1984).

1.3.3 Gene Duplication and Loss

Gene duplication is a significant source of evolutionary novelty (Ohno 1970) and can occur as a result of a number of different events, as depicted in Figure 1.2: (a) DNA-mediated duplication can take place as a result of unequal crossing over during recombination, and may cause the duplication of one or more genes (Kaessmann 2010); (b) RNA-mediated duplication may be caused by reverse transcription of a messenger RNA molecule, that results in a gene duplicate lacking introns and, in many cases, associated regulatory regions (Kaessmann 2010).

Gene duplication is estimated to occur in eukaryotes at a rate of up to 0.01% per gene per million years (Lynch and Conery 2000). Upon duplication, a number of distinct fates are possible for the duplicates, including: (i) dosage selection (retention of both copies due to dosage effects), (ii) escape from adaptive conflict (where the ancestral protein had two functions in adaptive conflict and each duplicate retains one of each of the conflicting functions — this is a form of subfunctionalisation), (iii) neofunctionalisation (acquisition of a new function by one duplicate), (iv) subfunctionalisation (division of the original gene function into two subfunctions, each performed by one duplicate), and (v), non-functionalisation (loss of function in one duplicate through pseudogenisation). A number of examples of gene duplication and associated functional divergence have been identified, including the co-option of an RNase gene for the purpose of extracting nutrients from bacteria in a leaf-eating monkey (Zhang *et al.* 2002), the evolution of a duplicate crystallin in zebrafish (Smith *et al.* 2006), the evolution of a galactokinase (GAL1) and co-inducer (GAL3) in yeast species *Saccharomyces cerevisiae* from a single gene with both functions (Hittinger and Carroll 2007), and the expansion and divergence of the opsin gene repertoire among ray-finned fish (Rennison *et al.* 2012).

Non-functionalisation is the most common fate for a duplicate gene, often with pseudogenisation of the duplicate (Lynch and Conery 2000). Using survivorship rates of gene duplicates in *C. elegans*, *D. melanogaster* and *S. cerevisiae*, Lynch and Conery (2000) estimated that the average half-life of a gene duplicate is about 4 million years. Based on this estimate, a comfortable majority of duplicated genes can be expected to have become a pseudogene within 10 million years.

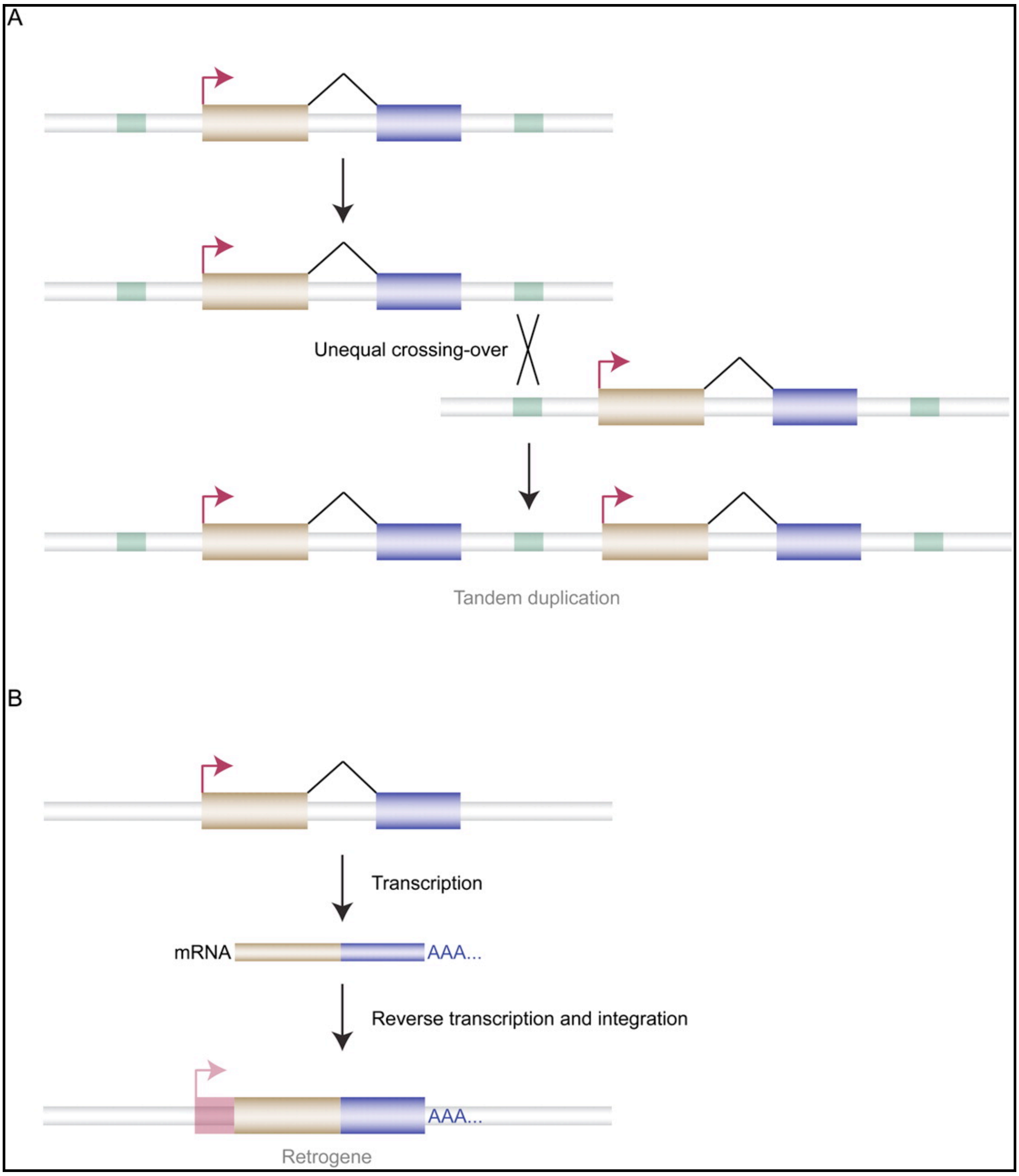


Figure 1.2: Mechanisms of gene duplication.

Mechanisms of gene duplication include: (A) DNA-mediated duplication as a result of unequal crossing-over during recombination; (B) RNA-mediated duplication due to reverse transcription and integration of mRNA. Figure taken with permission from (Kaessmann 2010).

1.3.4 Gene Fusion and Fission

Partial duplication of a coding gene may sometimes result in a gene fission/fusion process: where the original gene undergoes ‘fission’ into two new genes, or part of the gene undergoes ‘fusion’ with a second gene to form a chimeric gene (Snel *et al.* 2000). A classic example of a fusion event is offered by the Jingwei gene in *Drosophila*, whose origin was first studied by Long and Langley (1993). Jingwei was found to have formed from part of the alcohol dehydrogenase (Adh) gene and another, unrelated gene in the common ancestor of *Drosophila yakuba* and *Drosophila teissieri* (Long and Langley 1993). Again in *Drosophila*, specifically in the *melanogaster* subgroup, a gene fission event occurred that involved the monkey king gene; in this case the fission was mediated by gene duplication and subsequent partial degeneration (Wang *et al.* 2004).

1.3.5 Whole-Genome Duplication

In addition to duplication at the level of the gene, DNA sequences can be duplicated at the chromosomal scale, or in rare instances, at the scale of the genome — a whole-genome duplication (WGD) (Kaessmann 2010). Ohno (1970) first suggested the "2R hypothesis" that two rounds of whole-genome duplication have occurred in the ancestral vertebrate; this has been supported by more recent genomic analysis (Hokamp *et al.* 2003, Kasahara 2007, Huminiecki and Heldin 2010).

Genes in particular functional categories tend to be duplicated either in a gene duplication or a whole-genome duplication in a mutually exclusive manner (Maere *et al.* 2005). Genes duplicated by WGD tend to retain more interacting partner proteins than those duplicated in a smaller scale event (Guan *et al.* 2007). This may be due to dosage-balance effects, such that the relative expression levels of a set of genes affects their function (Conant and Wolfe 2008). In any case, as with gene duplication, many of the duplicate genes are lost over time following a whole-genome duplication (Scannell *et al.* 2007).

1.4 Fixation

Whatever the nature of a mutation, its emergence as characteristic of a species as a whole depends on whether it increases in frequency in the population until it reaches fixation: a situation in which the frequency of an allele in the population reaches 100% (Patthy 2008). As outlined above, this can happen by two general mechanisms: genetic drift, in which a mutation may become fixed by chance, and natural selection, in which a mutation may become fixed because it has a beneficial effect on the phenotype (**positive selection**) or become lost from the population because it has a deleterious effect (**purifying** or **negative selection**).

Most mutations within a gene sequence are neutral or nearly neutral, and have a negligible effect at the phenotypic level; if such mutations become fixed, it is due to genetic drift. Whether advantageous or deleterious, such nearly neutral mutations are fixed more readily in populations with a small effective population size (N_e).

Mutations that do impact on the phenotype of an organism (at the macroevolutionary level), or the function of a gene/protein (at the molecular evolutionary level), may also be affected by genetic drift, but natural selection plays a more significant role in determining their fate (Hurst 2009). Most function-altering mutations are deleterious and natural selection tends to remove these from a population through a process known as negative or purifying selection. However, a small number of mutations have a beneficial effect on the function and natural selection will tend to positively select these beneficial mutations and bring them to fixation, through a process known as positive selection or adaptive evolution.

The relative effect of natural selection as opposed to genetic drift is directly proportional to the relative fitness of the mutant compared to the wild type. The relative impact of natural selection is also directly proportional to the size of a population: larger populations tend to be subject to stronger natural selection, while genetic drift plays a larger role in smaller populations (Ohta 1973). In a larger population, genetic drift is less likely to bring an allele to fixation before natural selection tends to dominate; furthermore, in such cases the sensitivity of an allele to the effects of natural selection is increased as the fitness of that allele deviates from neutrality (see Figure 1.3). Conversely, smaller populations tend to have less genetic variation, which in turn limits the effectiveness of natural selection and the ability of a species to adapt to changing circumstances (Willi *et al.* 2006).

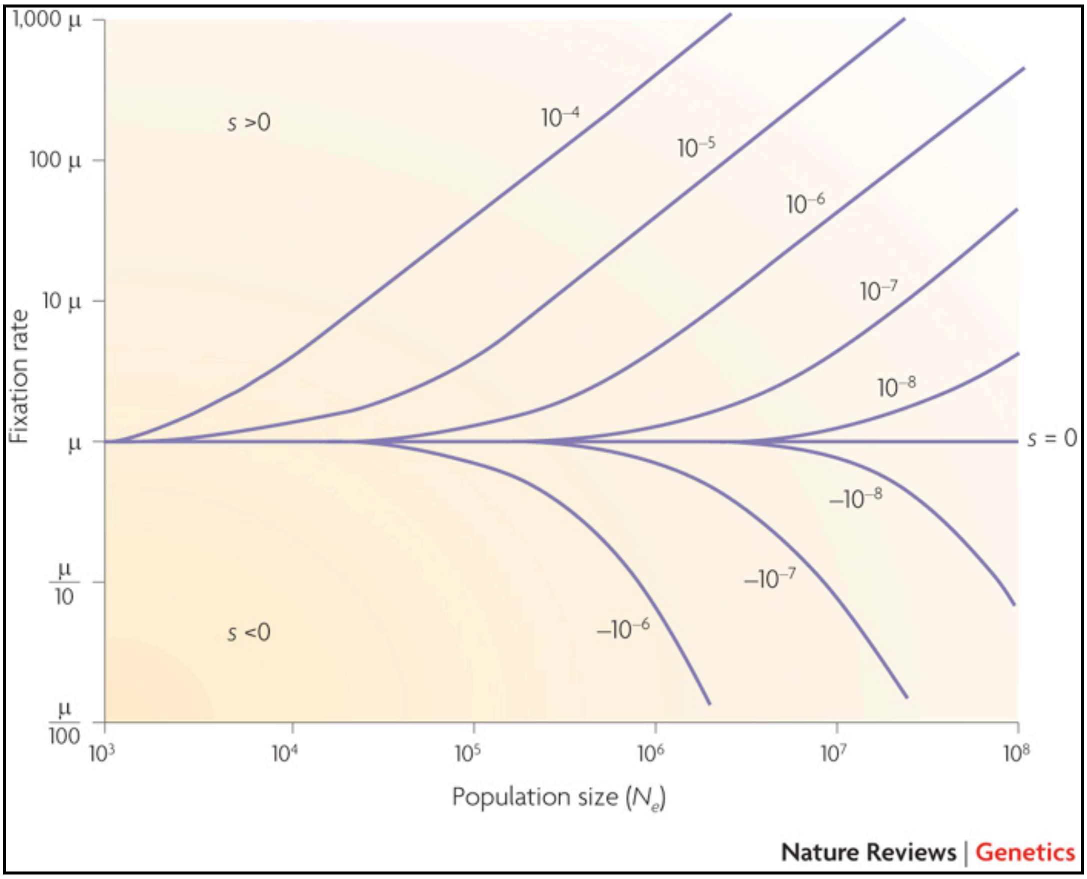


Figure 1.3: Fixation rate and the effective population size.

This graph shows the expected fixation rate for a novel mutation as a function of the effective population size (N_e). The horizontal axis shows the effective population size, while the vertical axis shows the fixation rate in terms of the mutation rate (μ). Different curves are shown for different values of the selection coefficient (s), which indicates the fitness of the mutation relative to the wild type. Mutations with a selection coefficient of zero are strictly neutral, while those with a positive or negative selection coefficient are advantageous or deleterious, respectively. Figure taken from (Hurst 2009). Permission obtained via the Copyright Clearance Center.

1.5 Speciation

Foreshadowed by Poulton (1903), Ernst Mayr in 1970 formulated the biological species concept: a species is a group of actually or potentially interbreeding populations that are reproductively isolated from other groups (Mayr 1970). This definition implicitly excludes organisms that reproduce asexually and is confounded by hybridisation between species (Mallet 2005) and horizontal gene transfer (Koonin *et al.* 2001, Keeling and Palmer 2008). Nevertheless, in animal species the vertical transfer of genes from parents to offspring predominates. More recent formulations of the biological species concept account for this by describing the reproductive isolation of species as substantial, if not complete (Coyne and Orr 2004). Such reproductive isolation can occur due to barriers in geography in allopatric speciation[1] (Mayr 1970) or in the absence of such barriers in sympatric speciation (Poulton 1903) due to, for example, ecological or behavioural factors (Butlin *et al.* 2008).

Once a speciation event has occurred, meaningful comparison of the descendant species is made possible by comparison of homologous traits, whether these be morphological or molecular. **Homolog** was defined by Richard Owen (1843) as "the same organ in different species under every variety of form and function". More recent definitions encompass both morphological and molecular homology in an evolutionary context, with the descent of two entities from a common ancestor constituting homology (Koonin 2005).

[1] Mayr (1970) defines allopatric speciation as "geographic speciation", and sympatric speciation as "speciation without geographic isolation by the acquisition of isolating mechanisms within a deme".

1.6 Homology

A homologous trait is a characteristic present in two or more species that is derived from a common ancestor. This is distinct from a homoplastic or analogous character, a trait present in two or more species that evolved separately and was not present in the common ancestor. The specialised beaks of the various species of Galapagos finch are an example of morphological homology. By comparing the different beak shapes (in addition to other homologous characters), it should be possible to reconstruct the evolutionary history of the Galapagos finches.

However, the identification of homologous traits is not always so straightforward. For example, the homology of a horse's hoof and the human middle finger might not be so immediately clear. The wings of bat and bird might appear to the naïve observer to be homologous, but this is not the case. Though these structures may be homologous as pectoral limbs, as wings they are homoplastic, having arisen by convergent evolution (McGhee 2011). In any case, it is clear from these examples that correct identification of homologous relationships is crucial if any inferences are to be made on the basis of those homologies.

1.6.1 Molecular Homology

As is the case with morphological homology, so it is also true of molecular homology. A prerequisite for reconstructing the evolutionary history of a gene is the correct identification of its homologs. Pevsner (2003) outlines a number of different cues that are used to infer homology between two or more genes. These are as follows:

- sequence similarity
- bidirectional best hit (see below)
- common sequence motifs
- similar biological function
- conserved **microsynteny**: conservation of gene order in local genomic regions
- similarity of protein tertiary structure

Cues other than sequence similarity are not always available for every putative homologous pair and sequence similarity searches remain the predominant tool used for identification of homologous groups in molecular evolutionary analyses (Fuellen 2008). The pairwise alignment software BLAST is commonly used to identify homologs of a query gene in a database of sequences (Camacho *et al.* 2009, Altschul *et al.* 1997, Altschul *et al.* 1990). BLAST searches may be used in a reciprocal manner, such that a putative homolog from an initial search is used as a query against the genome containing the original query gene, to obtain the bidirectional best hit (BBH). It is frequently used in comparative genomic analyses where the aim is to identify orthologous families of sequences. (Overbeek *et al.* 1999).

The identification of homology between sequences *per se* is not always a sufficient basis to draw conclusions on the nature of their relationship from a common ancestor; this is because of the occurrence of gene duplication resulting in mixtures of orthologs and paralogs — homologs whose common ancestor diverged in a speciation or gene duplication event, respectively — in many gene family datasets. For example, Fitch and Margoliash (1967) posit a scenario in which a set of haemoglobin homologs are taken from a number of species — such that each species is represented by a single gene — and their amino acid sequences are compared and used to infer the evolutionary history of the haemoglobin gene family. In this scenario, half of the homologs are α-globins and half are β-globins. Such a scenario would likely produce a nonsensical result, with α-globins and β-globins clustering together irrespective of their species of origin (Fitch and Margoliash 1967). Indeed, such a situation arose — and was highlighted as an example of this issue — in the study by Martin and Burg (2002) of heat shock 70 (HSP70) genes in sharks.

This example illustrates the necessity of ensuring that the set of homologs in a given gene family is as complete as possible. The advent of genome sequencing projects has made this a practical prospect. With complete genomes of high quality, it is much more plausible that all homologs for a given gene will be identified in a given genome, thus allowing for reconstruction of an accurate evolutionary history of the gene family. This example also underlines the importance of distinguishing between homologous genes whose most recent common ancestor (MRCA) diverged in a speciation event from those whose common ancestry diverged in a gene duplication.

I. Types of Homologous Genes

Fitch (1970) proposed that homologs should be divided into two subtypes. A homolog in a different species that evolved from a common ancestor with a given gene by a speciation event, such that their evolutionary history mirrors that of their respective species, is known as an **ortholog**. Where a gene duplication event occurs and descendants of the duplicate genes remain, each such gene is a **paralog**.

Figure 1.4 (taken from Jensen 2001) depicts an evolutionary tree for three species A, B and C, with one, two and three genes in each species, respectively. All the genes in Figure 1.4 are homologs of each other, since they all share a common ancestor. The ancestral node of any pair of genes indicates whether those genes are orthologous or paralogous. For example, gene B2 is a paralog of gene C1, since their MRCA split in two in a gene duplication, while B2 is an ortholog of both C2 and C3, since their MRCA diverged into two lineages in a speciation event. Note also that genes C2 and C3 are paralogs of each other, while being co-orthologs of B2 (Jensen 2001).

Sonnhammer and Koonin (2002) proposed a further division of paralogs into two subtypes: in-paralogs and out-paralogs. An **in-paralog** is a paralog formed by a duplication that occurred after a particular speciation event. In Figure 1.4, genes C2 and C3 are in-paralogs with respect to Speciation 2, since they arose in Gene duplication 2 after Speciation 2. An **out-paralog**, on the other hand, is a paralog formed by a duplication that occurred before a particular speciation event. In Figure 1.4, B2 and C1 are out-paralogs with respect to Speciation 2, because they diverged in Gene duplication 1 before Speciation 2.

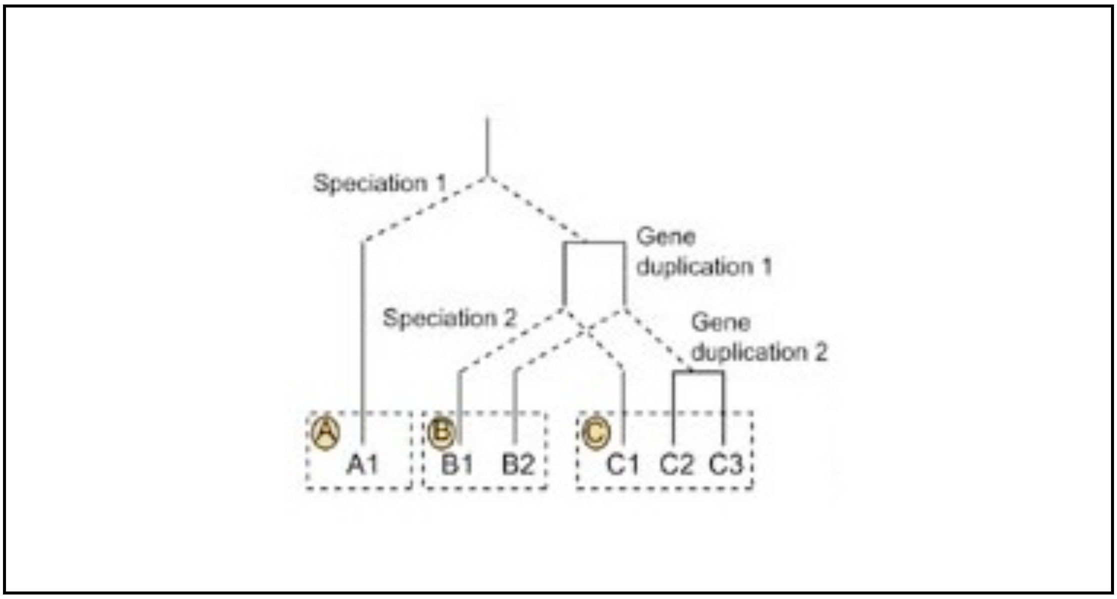


Figure 1.4: Types of homologous genes.

The phylogeny shows a simplified evolutionary history for genes in species A, B and C. Speciation events are denoted by forking lines, while gene duplications are denoted by horizontal lines. The genes A1, B1, B2, C1, C2 and C3 are all homologs of each other, but the exact relationship of any pair of homologs differs depending on the circumstances in which their most recent common ancestor (MRCA) diverged. Genes whose MRCA diverged in a speciation event are orthologs, while genes whose MRCA diverged in a duplication event are paralogs. Figure taken with permission from Jensen (2001).

II. The Ortholog Conjecture

The hypothesis that orthologs are more functionally conserved than paralogs of similar sequence divergence is known as the **ortholog conjecture** (Koonin 2005, Tatusov *et al.* 1997). The validity of this hypothesis underlies much of the functional annotation of newly sequenced genomes in the absence of experimental validation of function (Koonin 2005).

Some confusion has arisen from the ortholog conjecture, with some authors defining orthology in terms of conservation of function (Fàbrega *et al.* 2001, Gerlt and Babbitt 2000). Jensen (2001) argued against the incorporation of function into the definition of orthology (since it is possible for orthologs to diverge functionally and for paralogs to retain the same function), proposing that new nomenclature be used to classify homologs in terms of functional conservation, if necessary. In an accompanying comment, Gerlt and Babbitt acknowledged that orthology and paralogy were originally defined in terms of speciation and gene duplication events, respectively, but pointed out that in practice such events are not always inferred with certainty (Jensen 2001). Gerlt and Babbitt further proposed that the terms isofunctional and heterofunctional be used to denote homologs with similar and differing functions, respectively (Jensen 2001).

Citing a discussion of the issue by Fitch (2000), Kuzniar *et al.* (2008) defined orthologs explicitly in terms of an evolutionary relationship — "homologous sequences derived by a speciation event from a single ancestral sequence in the last common ancestor of the species being compared" — but then went on to restate the ortholog conjecture as part of the same definition, illustrating the pervasiveness of the concept of the ortholog conjecture. Studer and Robinson-Rechavi (2009) highlighted the need to test the ortholog conjecture more rigorously, rather than simply assuming it to be true. A number of subsequent studies have found evidence in support of the ortholog conjecture, in terms of protein structure (Peterson *et al.* 2009b), intron position (Henricson *et al.* 2010), domain architecture (Forslund *et al.* 2011) and expression patterns in mammals (Huerta-Cepas *et al.* 2011). In contrast, Qian and Zhang (2009) studied the evolution of sub-cellular localisation and found no difference between orthologs and paralogs. Furthermore, Nehrt *et al.* (2011) found that paralogs tend to have higher functional conservation than orthologs, in a study that compared functional annotation from the Gene Ontology (GO) database (Ashburner *et al.* 2000). However, in a 2012 paper, members of the Gene Ontology Consortium cautioned against the use of GO annotation to test the ortholog conjecture, highlighting the incompleteness of GO annotation and in particular the species bias of GO terms, since different organisms are used to study different aspects of biology (Thomas *et al.* 2012). The authors further stated that this species bias in GO term frequency is likely to give the impression of conservation between paralogs within the same species. Altenhoff *et al.* (2012) also performed a test of the ortholog conjecture using GO annotation, controlling for authorship bias in GO term frequency, species bias in GO term frequency, variation in similarity between pairs of species and propagated GO annotation bias. Controlling for these biases, Altenhoff *et al.* (2012) found weak but significant support for the ortholog conjecture[2]: in a comparison of orthologs and paralogs from 13 species, the excess similarity of GO terms for orthologs over paralogs ranged from 0.028 for homologs with 50-60% sequence identity (t-test, $P < 2.8 \times 10^{-9}$), to 0.136 for homologs with more than 90% sequence identity (t-test, $P < 8.1 \times 10^{-91}$). Therefore, orthologs do tend to be more conserved than paralogs, but not overwhelmingly so.

III. Inference of Orthology and Paralogy

A gene family in which no gene duplications or losses had occurred would be composed entirely of orthologs and would have an evolutionary history mirroring that of the species. In a gene family without gene loss, a gene duplication would have the effect of duplicating the gene family history; in the absence of further duplications, both duplicates would again mirror the evolution of species after the initial duplication event.

[2] One exception to this trend occurred in homologs with 10-20% identity; in this case, paralogs were found to have an excess similarity in GO terms of 0.025 relative to orthologs (t-test, $p < 2.2 \times 10^{-4}$). However, as noted by the authors, identifying homologs can be challenging and error prone in the "twilight zone" of sequence similarity; distinguishing orthologs from paralogs even more so.

Tree reconciliation — comparison of the gene family history with that of the species — can facilitate inference of duplication events and distinguish orthologs from paralogs within the gene family. This classic method of identifying orthologs and paralogs is known as tree reconciliation because the phylogenetic trees of species and gene family are 'reconciled' such that — notwithstanding the effects of gene loss, horizontal gene transfer and incomplete lineage sorting — monophyletic groups of orthologs on the gene tree are congruent with the topology of the species phylogeny (Eulenstein *et al.* 1998, Page and Charleston 1997, Maddison 1997, Mirkin *et al.* 1995). (See Section 1.7 for a brief introduction to phylogenetic trees.)

In practice this process is complicated by a number of factors. Errors in inference of either the gene tree or species tree can be propagated to the reconciliation process (Koonin 2005). In addition, gene loss can make it difficult to infer duplication events correctly; in extreme cases of reciprocal loss in two daughter lineages, paralogs may be misidentified as orthologs. Furthermore, tree reconciliation is computationally expensive and difficult to "scale-up" for large numbers of gene families (Koonin 2005), with only one orthology database (i.e. Ensembl Compara) inferring orthology by tree reconciliation across multiple genomes (Vilella *et al.* 2009).

Heuristic alternatives to tree reconciliation include the bidirectional best hit (BBH), in which two genes are considered orthologs if each is the best aligning gene to the other in their respective genomes (Overbeek *et al.* 1999); and the reciprocal smallest distance (RSD), where two genes are considered orthologs if each has the shortest estimated evolutionary distance to the other in their respective genomes (Wall *et al.* 2003). A disadvantage of both these methods is that for a given set of species, they are only able to identify **one-to-one orthologs**: a set of orthologs such that each species has no more than one ortholog. This can be a considerable limitation if a gene family contains **one-to-many orthologs** — a set of orthologs such that some species have multiple genes co-orthologous to a single gene in other species — or **many-to-many orthologs** — a set of orthologs such that some species have multiple genes co-orthologous to multiple genes in other species.

More sophisticated orthology prediction methods use sequence similarity to cluster groups of putative orthologs and in-paralogs across pairs of genomes, such as Inparanoid (Ostlund *et al.* 2009, Remm *et al.* 2001); or across multiple genomes, such as EggNOG (Muller *et al.* 2009, Jensen *et al.* 2008), or OrthoMCL (Li *et al.* 2003).

1.6.2 Site-wise Molecular Homology

Fundamentally, the identification of homologous biological sequences involves identifying homologous sites in those sequences. Typically, homologs are identified by sequence similarity to a gene of interest in a pairwise manner, such that many of the sites in the two genes are identical or similar, and by implication homologous. When a set of homologs have been identified, these are "aligned" in preparation for more detailed comparison. Alignment is essentially the identification of homologous positions within these homologous sequences. In a multiple sequence alignment (MSA), each sequence occupies its own row, and each column represents a site that is homologous across the set of aligned sequences. An MSA is effectively a hypothesis about sitewise homology, and maximising the alignment of truly homologous positions is critical to ensuring the validity of downstream analyses (Anisimova *et al.* 2010, Wong *et al.* 2008, Kumar and Filipski 2007). There are two main aspects to the process of correctly aligning homologous sequences: handling of (I) substitutions and (II) indels.

I. Substitutions

When a point mutation increases in frequency in a population such that every member of that species has the novel form (i.e. the point mutation is fixed in the population), this is known as a point substitution. Comparison of character states at homologous sites requires that point substitutions be accounted for, and this is typically done with a substitution model.

Such a **substitution model** is a matrix with a row and column for each possible character state, with the expected frequency of a substitution from character state R to character state C reflected by the value of the element in row R and column C. With increasingly similar sequences — and correspondingly fewer substitutions — such a substitution model tends toward the identity matrix.

Nucleotide Substitution Models

Perhaps because of the smaller number of possible states, substitution models were developed for DNA sequences before peptide sequences. The first such model was the JC69 model developed by Jukes and Cantor (1969). The JC69 model assumes equal nucleotide base frequencies: $\pi_A = \pi_C = \pi_G = \pi_T = 0.25$, where π_N represents the frequency of base *N*. The JC69 model also assumes that all substitutions are equiprobable, so this can be modelled by a single parameter, μ (the overall substitution rate). Felsenstein (1981) introduced the F81 model, in which base frequencies are allowed to vary: $\pi_A \neq \pi_C \neq \pi_G \neq \pi_T$; and the substitution probability for each base is determined by the base frequency. At about this time it was becoming clear that transition and transversion rates could not always be assumed equal (Brown *et al.* 1982, Gojobori *et al.* 1982, Curtis and Clegg 1984); this necessitated the development of models that could account for different transition and transversion rates. Kimura (1980) devised the K2P model, which assumes base frequencies to be equal, but allows for different rates of transition and tranversion; this was followed by the K3P model (Kimura 1981), which models one transition rate and two transversion rates. Hasegawa *et al.* (1985) developed the HKY85 model, which has both variable base frequencies and one rate for transition and transversion. The general time reversible (GTR) model accounts for variable base frequencies and all possible nucleotide substitutions using a symmetrical matrix (Lanave *et al.* 1984, Tavaré 1986, Rodríguez *et al.* 1990). Development of nucleotide substitution models has continued (Tamura 1992, Tamura and Nei 1993, Zharkikh 1994), but the GTR is arguably the most complex nucleotide substitution model in common usage (Posada and Crandall 2001). In any case, the increased complexity of a model does not, in itself, ensure that the model is a significantly better fit to the data than a simpler model (see below).

Amino Acid Substitution Models

The first amino acid substitution model was that devised by Dayhoff *et al.* (1978). This was empirically derived from estimated mutation rates among a set of 34 protein superfamilies with at least 85% sequence similarity. The high sequence similarity used was necessary to avoid counting successive substitutions at the same site. Henikoff and Henikoff (1992) developed the BLOSUM amino acid substitution matrices from their BLOCKS database, which contained a larger number of aligned proteins at a range of evolutionary distances. This enabled a marked improvement in performance on the Dayhoff model, and the Blosum62 matrix (derived from sequences with 62% identity) was found to be the most generally useful (Henikoff and Henikoff 1992). In the same year, Jones *et al.* (1992) developed the JTT substitution matrix using a similar method to that of Dayhoff *et al.* (1978), but with the much larger set of protein sequences in the nascent Swiss-Prot database (Bairoch and Boeckmann 1992).

General amino acid substitution matrices continue to be developed. The VT model (Müller and Vingron 2000) was developed using a novel resolvent method — capable of estimating amino acid substitution rates from alignments of varying levels of divergence — with the SYSTERS database of aligned protein families (Krause and Vingron 1998). The WAG model (Whelan and Goldman 2001) used an approximate likelihood method to estimate amino acid substitution rates from nearly 4,000 amino acid sequences in the 180 alignments comprising the BRKALN database (D.T. Jones, unpublished). Like WAG, the LG model (Le and Gascuel 2008) was developed using a likelihood method, in this case implemented in the software XRATE (Holmes and Rubin 2002, Klosterman *et al.* 2006). However, for estimation of amino acid substitution rates, Le and Gascuel (2008) used a much larger database of amino acid alignments — Pfam (Bateman *et al.* 2002) — comprising nearly 4,000 alignments of approximately 50,000 protein sequences.

Much work has also been done to try and capture the complexities of biological sequence evolution, such as invariant sites (Reeves 1992), among-site rate variation (Yang 1996), variation in sequence composition between lineages (Foster 2004) and variation in substitution rates across lineages (Lartillot 2004, Lartillot *et al.* 2007).

Selection of the Substitution Model

The selection of the substitution model in itself is not trivial and can have a critical impact on downstream phylogenetic analyses (Felsenstein 1978, Bruno and Halpern 1999, Keane *et al.* 2006). For nested models (e.g. JC is a special case of K2P, which is in turn a special case of GTR), the likelihood ratio test (LRT) (Hoel 1962) may be used to select the most appropriate model to use (Goldman 1993). Given a series of models, the model of best fit may be selected using hierarchical likelihood ratio tests (hLRTs) (Huelsenbeck and Crandall 1997). However, the applicability of hLRTs to model selection has been questioned (Burnham and Anderson 2002), and hLRTs can not be used to compare non-nested models or to assess model selection uncertainty (Posada and Buckley 2004). Alternative methods of model comparison include:

- **Akaike information criterion** (AIC) (Akaike 1973, Akaike 1974), given in Equation 1.1.

  $$AIC = -2 \cdot \log(\ell) + 2 \cdot k$$

  Equation 1.1: Akaike information criterion.
  The *AIC* is given in terms of the log-likelihood of a model (i.e. log(ℓ)) and a penalty term 2*k*, such that *k* is the number of free parameters in the model.

  The AIC reflects the amount of information lost when using a given model to approximate the underlying process.

- **Bayesian information criterion** (BIC) (Schwarz 1978), given in Equation 1.2.

$$BIC = -2 \cdot \log(\ell) + k \cdot \log(n)$$

Equation 1.2: Bayesian information criterion.

The BIC is given in terms of log-likelihood of the model (i.e. $\log(\ell)$) and a penalty term $k \cdot \log(n)$, where k is the number of free parameters in the model and n is the number of data points.

The BIC is equivalent to choosing the model with the maximum posterior probability, assuming the models have equal prior probabilities (Posada and Buckley 2004).

- **Deviance information criterion** (DIC) (Spiegelhalter *et al.* 2002), given in Equation 1.3.

$$DIC = \overline{D} + p_D$$

Equation 1.3: Deviance information criterion.

The DIC is given in terms of the posterior mean deviance ($\overline{D}$) and a penalty term (p_D) defined (with respect to the relevant parameters) as the difference between the posterior mean of the deviance and the deviance of the posterior means.

The DIC can be seen as a generalisation of AIC and BIC within a Bayesian framework; the principal drawback of the BIC is that it necessitates that model fitting be performed using Markov chain Monte Carlo methods (Burnham and Anderson 2002).

- **Hannan-Quinn information criterion** (HQ) (Hannan and Quinn 1979), given in Equation 1.4.

$$HQ = -2 \cdot \log(\ell) + 2k \cdot \log(\log(n))$$

Equation 1.4: Hannan-Quinn information criterion.

The HQ information criterion is given in terms of the log-likelihood of a model (i.e. $\log(\ell)$) and a penalty term $2k \cdot \log(\log(n))$, such that k is the number of free parameters in the model and n is the number of data points.

The Hannan-Quinn information criterion has been described by Burnham and Anderson (2002) as widely cited but little used. Perhaps this is partly because the effect of the penalty term — proportional to $\log(\log(n))$ — is low even for large n (Claeskens and Hjort 2008).

Posada and Buckley (2004) suggest that the BIC should be used if the aim is to identify the true model, while the AIC should be used if the aim is to approximate reality. An advantage of the BIC is that it strongly penalises parameter-rich models; both the AIC and hLRT tend to favour parameter-rich models in large datasets (Schwarz 1978). While not explicitly conceived in terms of parsimony, it has been argued that the AIC achieves a parsimonious solution as a by-product, trading off a greater fit of the model to the data against the introduction of too many parameters, which could lead in turn to model overfitting (Anderson 2008).

Amino acid sequences are considered to be better than nucleotide sequences for reconstructing substitution events in protein-coding genes. Synonymous nucleotide substitutions in particular tend to occur relatively frequently, so multiple substitutions and reversions introduce conflict into the phylogenetic signal at each site. The smaller number of possible nucleotide states — four as opposed to twenty for amino acids — compound this effect, since the same site is more likely to have the same character state by chance in distantly related taxa (Felsenstein 1996a, Russo *et al.* 1996).

Indels

Insertions and deletions (indels) in molecular sequences pose a considerable challenge to the identification of homologous sites, especially at greater evolutionary distances. Indels are most commonly represented by a gap character (e.g. hyphen) in sequence alignments, and are typically modelled using an affine gap penalty, which penalises the insertion and extension of a gap differently, since in practice many indels involve multiple sites. For example, ClustalW incorporates a gap opening penalty in addition to a gap extension penalty (Larkin *et al.* 2007, Thompson *et al.* 1994). Gap penalties may also be informed by biochemical knowledge; for example, ClustalW locally reduces gap penalties in peptide regions containing hydrophilic amino acids, so as to reward new gaps in hydrophilic loop regions where indels are expected to have less impact on protein structure, while discouraging their introduction in regions corresponding to regular secondary structure. For the same reason, MUSCLE (Edgar 2004a) locally increases gap penalties in hydrophobic regions. MUSCLE also incorporates an affine gap penalty, but includes an additional gap closing penalty, to ensure that the treatment of gaps is symmetrical (Edgar 2004a). Gaps are handled in an evolution-aware manner by PRANK (Löytynoja and Goldman 2008), which attempts to explicitly model insertion and deletion events over the evolutionary history of the sequences being aligned, such that aligned sites are presented as homologous. In doing so, PRANK eliminates a systematic bias present in previously developed methods of multiple sequence alignment, in which an inability to separate distinct, nearby insertions can lead to their spurious alignment, which can in turn cause such misaligned insertions to be misidentified as mutation hotspots (Löytynoja and Goldman 2008).

1.7 Phylogeny and Phylogenetic Reconstruction

The **phylogenetic tree** provides a model to describe the evolution of many extant species from a common ancestor through a series of speciation events. Darwin's Origin of Species, published in 1859, contained one of the earliest examples of a phylogenetic tree (Darwin 1859). The conceptual framework of the phylogeny has endured, and remains widely applicable today, although not universally so (Doolittle 1999).

A rooted phylogenetic tree for the species used in this study is depicted in Figure 1.6. In this phylogeny, each of the rightmost nodes – the **operational taxonomic units** (OTUs) or leaves – represents an extant species, such as human or mouse. Each node to the left of the leaves – termed an internal node – represents the most recent common ancestor (MRCA) of all the species to the right of that node, and the forked branches leading from each internal node represent a speciation event resulting in two divergent lineages. The leftmost node – the root – represents the most recent common ancestor of all the species in the tree. The bars to the right of the phylogeny highlight key clades.

Phylogenies are commonly represented in Newick format (Felsenstein 1996b), a text representation of the relationships within a phylogeny, which lists the children of a node in comma-separated lists enclosed in nested sets of parentheses. See Figure 1.5 for an example.

A phylogenetic relationship can also be represented in terms of a **split**: a partition (i.e. splitting) of the set of taxa into two groups, so that, for example, the species in one partition are more closely related to each other than to the species in the other partition. Any compatible set of splits can be used to construct a unique phylogenetic tree, and *vice versa* (Buneman 1971).

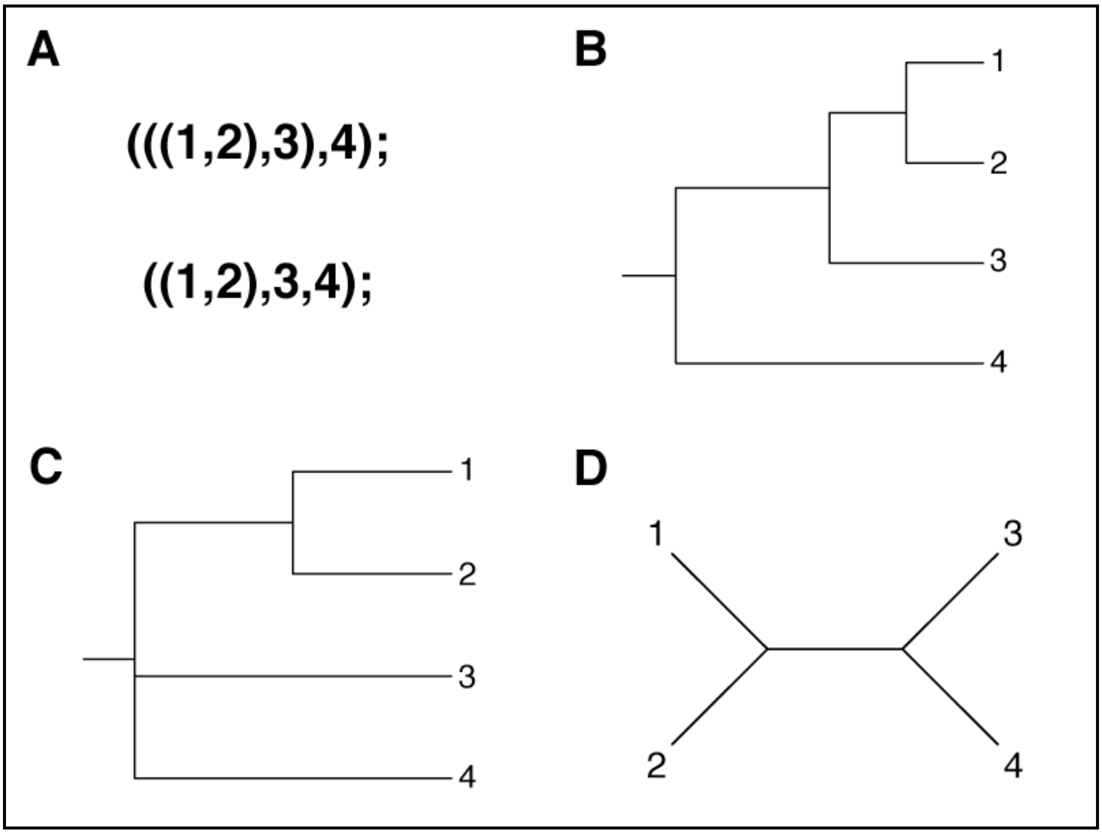


Figure 1.5: Newick format trees.

Shown is an illustration of the Newick tree format, a text file format commonly used for representing phylogenies in text. A Newick tree comprises a set of taxa where siblings are listed in a comma-separated list, while internal nodes are represented by nested parentheses enclosing those taxa that descend from the given node. For example, the upper Newick text in (A) corresponds to the rooted phylogeny depicted in (B), with taxon 1 grouped with 2, taxa 1 and 2 grouped with 3, and taxa 1, 2, and 3 grouped with taxon 4. Due to the nature of the Newick format, all Newick trees have an implied root. A workaround used in some contexts is to treat Newick trees with a multifurcation at the 'root' as unrooted phylogenies. For example, the lower Newick text in (A) corresponds to the phylogeny depicted in (C). Strictly speaking, this phylogeny is rooted, with a trifurcation at its root. However, some phylogenetic software (e.g. PAML, see Section 1.8.3) interprets such a Newick phylogeny as being an unrooted phylogeny like that shown in (D).

Figure 1.6 Legend:

The phylogeny shown on page 35 overleaf depicts the evolutionary relationships of the 22 species studied in this thesis; these species relationships are uncontroversial (Benton and Donoghue 2007). The species shown are (from top to bottom): Human (*Homo sapiens*), Chimp (*Pan troglodytes*), Gorilla (*Gorilla gorilla*), Orangutan (*Pongo abelii*), Macaque (*Macaca mulatta*), Marmoset (*Callithrix jacchus*), Mouse (*Mus musculus*), Rat (*Rattus norvegicus*), Guinea Pig (*Cavia porcellus*), Dog (*Canis familiaris*), Horse (*Equus caballus*), Cow (*Bos taurus*), Bat (*Myotis lucifugus*), Elephant (*Loxondonta africana*), Opossum (*Monodelphis domestica*), Platypus (*Ornithorhynchus anatinus*), Lizard (*Anolis carolinensis*), Chicken (*Gallus gallus*), Zebra Finch (*Taeniopygia guttata*), Frog (*Xenopus tropicalis*), Fugu (*Takifugu rubripes*) and Zebrafish (*Danio rerio*). The clade bars to the right show the extent of clades of interest: every species within the edges of each bar is within that clade, and the ancestral node for each clade is marked on the phylogeny by a small circle of the same colour as the corresponding clade bar. Credit for all animal silhouettes goes to Dr Mary J. O'Connell, except for those of the bat and lizard. The bat silhouette is a public domain image sourced from Wikimedia Commons (http://commons.wikimedia.org), while the lizard silhouette was derived from an image in Fritts and Leasman-Tanner (2001). Permission to use the lizard image was obtained from the US Geological Survey.

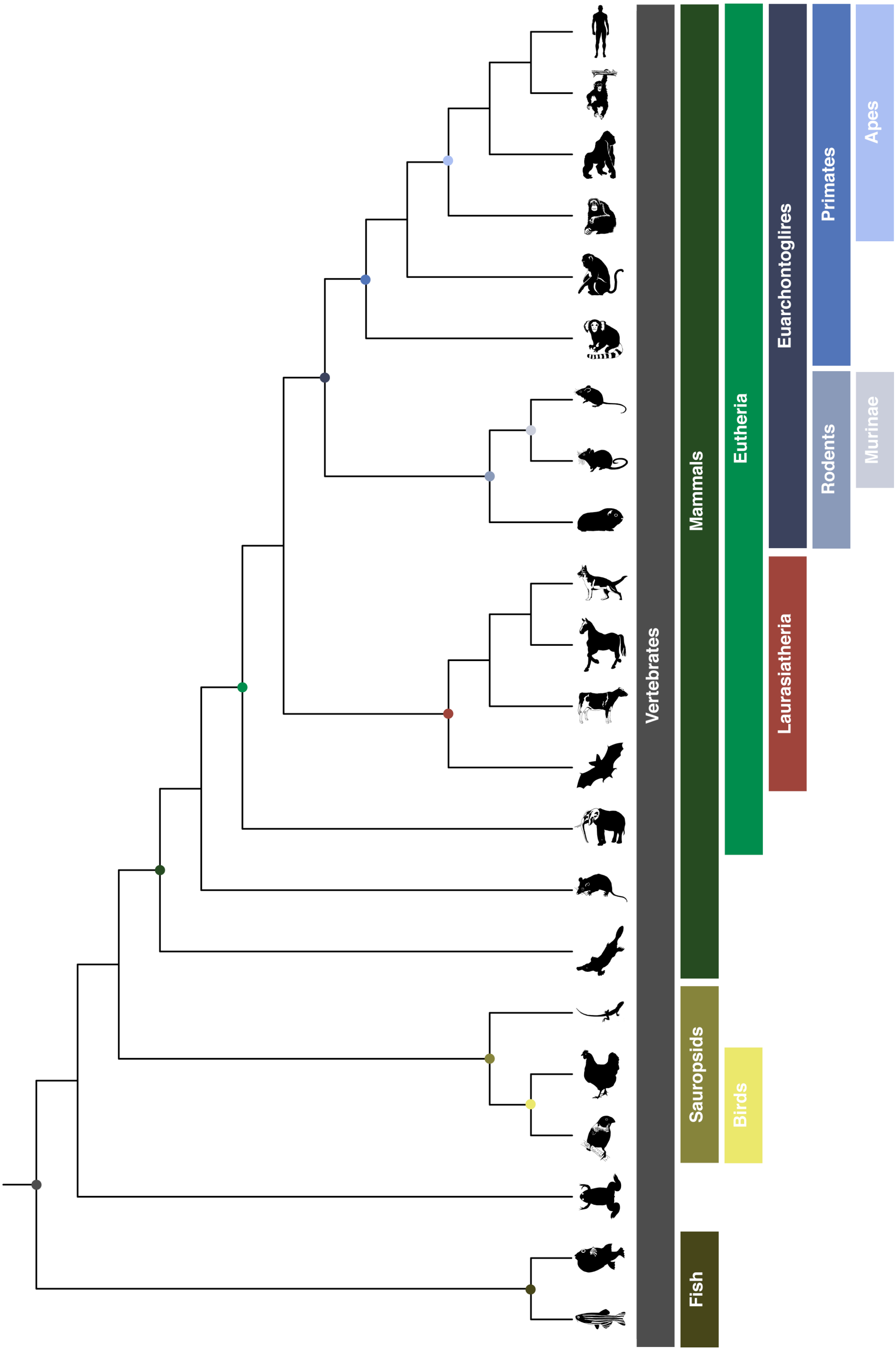


Figure 1.6: Phylogeny of species studied in this thesis.

The relationships between the species used in this analysis are largely uncontroversial and the divergence times are well calibrated and understood (Murphy *et al.* 2001, Benton and Donoghue 2007). There has been some uncertainty over the Eutherian root and the placement of Chiroptera within Laurasiatheria (Asher *et al.* 2009), although Song *et al.* (2012) have suggested that this uncertainty can be resolved by using a multispecies coalescent model in conjunction with phylogenomic data. In any case, with respect to the species in this study, the phylogeny of Benton and Donoghue (2007) is congruent with that of Song *et al.* (2012).

1.7.1 Molecular Phylogeny

Given a set of homologous genes or proteins, the gene family phylogeny can be inferred by comparison of homologous sites. There are a range of different methods available for this, falling into four main groups: (I) maximum parsimony, (II) distance methods, (III) maximum likelihood and (IV) Bayesian inference. These are described briefly in the following sections.

Irrespective of the method used, the inference of phylogeny can be confounded by a number of different factors. These include horizontal gene transfer (see Section 1.5), gene duplication followed by differential patterns of loss (see Section 1.6.1) and compositional bias brought about by biased gene conversion (see Section 1.3.1). Other confounding factors in phylogeny reconstruction include variation of substitution rates among sites (Yang 1996) and over time (Lockhart *et al.* 2006), both of which violate the assumptions of homogeneous substitution models in common use; and incomplete lineage sorting, in which the evolutionary history of a gene differs from that of species because two or more speciation events took place in a shorter time period than it took for the given gene to reach fixation (Degnan and Rosenberg 2009, Philippe *et al.* 2011).

I. Maximum Parsimony

Felsenstein (2004) describes a maximum parsimony tree as "that phylogeny on which, when we reconstruct the evolutionary events leading to our data, there are as few events as possible." Fitch (1971) developed a maximum parsimony algorithm to identify the most parsimonious tree for a set of input sequences.

Maximum parsimony can not make use of all sites in the input sequences; only those that are parsimony-informative. For a site to be parsimony-informative, there must be at least two character states for that site, each of which must be present in at least two sequences. For example, given a phylogeny with four taxa — human, mouse, platypus and frog — the presence of a chorioallantoic placenta is a parsimony-informative feature, since it is present in human and mouse and absent in platypus and frog. This supports a phylogeny grouping human with mouse and platypus with frog, at the expense of the other two possible phylogenies for these 4 species.

On the other hand, the presence of hair is not parsimony-informative in this case: human, mouse and platypus possess hair, while the frog does not. None of the 3 possible phylogenies of these 4 species are favoured by maximum parsimony with respect to the presence of hair. In a similar manner, the homologous character states at a sequence alignment site must favour one or more phylogenies as maximally parsimonious in order to be considered parsimony-informative.

The advantages of parsimony include its simplicity and low computational requirements. A significant disadvantage is its lack of an explicit evolutionary model. With more divergent sequences, inability to account for multiple substitutions at the same site causes it to underestimate sequence divergence and can lead to long branch attraction (LBA) (Yang and Rannala 2012).

Long branch attraction is a phenomenon that has the effect of causing species with higher rates of substitution — which on a molecular phylogeny, tend to reside at the end of long branches — to attract each other in an inferred phylogeny, and occurs when the sequence identity is so low as to enter the Felsenstein zone (Huelsenbeck and Hillis 1993). Maximum parsimony is particularly vulnerable to LBA, and in the Felsenstein zone will converge on an incorrect topology even with the addition of more data (Felsenstein 2004). This contrasts with maximum likelihood (ML): while ML phylogenies were found to be poorly resolved in the Felsenstein zone, ML was at least statistically consistent, whereas maximum parsimony methods were found to be more vulnerable to long branch attraction artefacts, offering greater support to an incorrect phylogeny (Swofford *et al.* 2001).

II. Distance Matrix Methods

Distance matrix methods include the least squares method (Cavalli-Sforza and Edwards 1967, Fitch and Margoliash 1967), the minimum evolution method (Rzhetsky and Nei 1994, Desper and Gascuel 2002), and neighbour-joining (Saitou and Nei 1987). These methods involve the construction of a matrix showing pairwise evolutionary distances between homologous sequences, which quantify the divergence between sequences. Pairwise distances can be obtained by, for example, a simple estimate of identity between sequences, although greater accuracy can be achieved using evolutionary models. For example, the formula for estimating the evolutionary distance per site between two sequences using the K2P model (Kimura 1980) is shown in Equation 1.5.

$$D_{K2P} = -\frac{1}{2}\ln(1-2p-q) - \frac{1}{4}\ln(1-2q)$$

Equation 1.5: Evolutionary distance under K2P model.

Shown is the evolutionary distance per site between two sequences under the K2P model (D_{K2P}), in terms of the proportion of sites with transition substitutions (p) and the proportion of sites with transversion substitutions (q).

Distance methods are more computationally efficient than other methods and can incorporate explicit models of evolution, but information is lost in the process of converting sequences to distances, and these methods are best suited to relatively similar sequences (San Mauro and Agorreta 2010).

III. Maximum Likelihood

Maximum likelihood estimation (MLE) is a common statistical method (Edwards 1972) that has been applied successfully to the field of phylogenetics (Felsenstein 1981, Felsenstein 2004, Yang 2006). The likelihood of a model with respect to a set of data is defined as the probability of the data, given the model. Maximum likelihood estimates the model parameters that maximise the likelihood of the given model. In phylogenetics, the data set is composed of the aligned gene family sequences, while the model comprises a substitution model and phylogenetic tree.

Advantages of maximum likelihood include the ability to incorporate explicit models of evolution, allowing these to be tested and improved over time, and the capability of testing hypotheses relating to the molecular clock or positive selective pressure, which can allow the nature of the evolution of a gene family to be analysed. Disadvantages of using maximum likelihood in phylogenetic inference include its computational complexity and its sensitivity to the use of a mis-specified evolutionary model (Yang and Rannala 2012). Of these two disadvantages, the former has been ameliorated by increases in available computing power in recent years, while the latter can be avoided by careful choice of evolutionary model.

IV. Bayesian Inference

Perhaps because of its relatively greater computational requirements, Bayesian inference was applied to phylogenetic inference more recently (Rannala and Yang 1996, Mau *et al.* 1999, Larget and Simon 1999). Fundamental to Bayesian inference is Bayes' Theorem (see Equation 1.6 below), which acts as a framework for updating a probability in light of new evidence (Press 2007). The posterior probability is obtained in terms of the prior probability and the likelihood of the evolutionary model (i.e. substitution model and phylogeny).

$$P(A \mid B) = \frac{P(A) \cdot P(B \mid A)}{P(B)}$$

Equation 1.6: Bayes' Theorem.

This relates the probability of A given B (i.e. *P(A|B)*) to the probability of A (i.e. *P(A)*), the probability of B given A (i.e. *P(B|A)*) and the probability of B (i.e. *P(B)*). The terms *P(A|B)*, *P(B|A)* and *P(A)* are called the posterior probability, likelihood and prior probability, respectively. Phylogenetic inference using maximum likelihood focuses on the likelihood of a phylogeny, while Bayesian inference of phylogeny takes account of the prior probability, to estimate the posterior probability of one or more phylogenies.

An exhaustive Bayesian inference of phylogenies with more than a modest number of species is not practical even with the recent increases in available computing power, since the 'tree space' of possible topologies expands factorially, so that a tree with 50 taxa has a number of possible topologies comparable with the number of protons in the visible universe (Felsenstein 2004).

In practice, most methods performing Bayesian inference of phylogeny use a sampling technique known as Markov chain Monte Carlo (MCMC) (Metropolis *et al.* 1953, Hastings 1970, Green 1995). In this approach, tree space is explored by a series of stochastic perturbations to phylogenies, in which the parameters of a given phylogeny are altered so as to sample its 'neighbours', thereby exploring the regions of tree space with more optimal phylogenies. A key property of MCMC is that the posterior probability of a phylogeny can be approximated by the proportion of time that the phylogeny is sampled from tree space (Tierney 1994).

Thus, while other methods infer a single phylogeny, the output of Bayesian phylogenetic methods is a distribution of trees sampled from tree space. These are typically sampled from a phase in the MCMC process at which the sampled phylogenies have converged towards an optimal region in tree space, with trees from the initial "burn-in" period before convergence being ignored (Felsenstein 2004). A maximum *a posteriori* probability tree may be obtained from this distribution (Rannala and Yang 1996). An arguably more powerful approach expounded by Huelsenbeck *et al.* (2002) is to summarise the sampled trees in a majority rule consensus tree (Larget and Simon 1999, Huelsenbeck *et al.* 2001). This allows the estimation, for each clade in the phylogeny, of the posterior probability that the given clade is true (Huelsenbeck *et al.* 2002).

Bayesian methods are similar to maximum likelihood estimation in that both make use of the likelihood function; hence, they share important statistical estimator properties like efficiency and consistency. Trees inferred by Bayesian methods have associated posterior probabilities that indicate in an intuitive manner how confident one can be in the given phylogeny. The prior probability can incorporate information that is known about the tree or model parameters (Yang and Rannala 2012). On the other hand, poor choice of prior can adversely impact the results of a Bayesian analysis (Rannala *et al.* 2012).

In addition, the underlying distribution in tree space is typically unknown, so it can be difficult to determine when a set of MCMC chains has achieved convergence on an optimal region in the parameter space. One **convergence diagnostic** — used by MrBayes (Altekar *et al.* 2004, Ronquist and Huelsenbeck 2003, Huelsenbeck and Ronquist 2001) — is the standard deviation of split frequencies, where split frequencies are estimated from the set of samples from tree space, and their standard deviation reflects the divergence between sampled trees. (See Section 1.7 for a definition of 'split' in this context.) Therefore, a lower value of the standard deviation of split frequencies indicates that the sampled phylogenies are becoming more similar, and are approaching convergence on the optimal region of tree space (Ronquist *et al.* 2007).

A further issue is the question of whether tree space has been sampled sufficiently, which is a function of the MCMC mixing behaviour: the extent to which optimal regions of tree space are explored in proportion to their posterior probability (Beaumont and Rannala 2004, Ronquist *et al.* 2007). As with convergence, our ignorance of the underlying distribution in tree space can pose a challenge in assessing whether that tree space has been sampled sufficiently, or whether, for example, the MCMC process has been 'trapped' in a region of tree space that is locally optimal, but not globally so (i.e. better than other trees in its neighbourhood, but not best overall). The use of Metropolis-coupled MCMC (Geyer 1991) can improve mixing behaviour and ameliorate the problem of local optima, and was successfully applied to Bayesian phylogenetic inference by Huelsenbeck *et al.* (2001).

1.8 Detecting Positive Selection at the Molecular Level

While neutral evolution dominates at the molecular level, natural selection can act both to restrict and promote change, in processes known as negative and positive selection, respectively. Numerous studies have found evidence for a link between positive selection at the molecular level and functional shift (Hughes and Nei 1988, Tanaka and Nei 1989, Yokoyama 1996, Messier and Stewart 1997, Levasseur *et al.* 2006, Loughran *et al.* 2012). Freely available statistical packages such as PAML (Yang 2007, Yang 1997) and the more recently developed HyPhy (Pond *et al.* 2005) include software to estimate levels of positive selection and thereby infer functional shift. This has led to an explosion in the number of positively selected genes identified by these methods (Kosiol *et al.* 2008, Hou *et al.* 2009, Metzger and Thomas 2010, Huang *et al.* 2012, Loughran *et al.* 2012). However, the statistical methods employed are not without their critics (Hughes and Friedman 2010), and inferences of positive selection using these methods are not always borne out by follow-up experiments. For example, in a study of fish rhodopsins by Yokoyama *et al.* (2008), eight amino acid sites were inferred by be under positive selection by PAML, none of which were found to affect the wavelength of maximum absorption of the rhodopsins in site-directed mutagenesis experiments.

1.8.1 Principles of Selective Pressure Analysis

Most positive selection inference methods are fundamentally based on the comparison of synonymous and non-synonymous substitutions between homologous protein-coding genes in different species (Miyata and Yasunaga 1980). With a sufficient number of aligned homologous positions, it becomes possible to estimate the rates of non-synonymous substitutions per non-synonymous site (Dn) and of synonymous substitutions per synonymous site (Ds) across the gene sequence. These values can then be used to estimate ω — the key statistic in codon-based selective pressure analysis (see Equation 1.7).

$$\omega = \frac{Dn}{Ds}$$

Equation 1.7: Estimation of *ω* value.

The value of *ω* is equal to the ratio of non-synonymous substitutions per non-synonymous site (*Dn*) to synonymous substitutions per synonymous site (*Ds*).

The value of *ω* can be used to infer the selective pressure on a given gene, region of a gene, or indeed on a lineage containing that homologous sequence, in the following ways:

- $\omega < 1$: indicates that the sequence is likely to be subject to negative (or purifying) selection; substitutions in this region that alter the peptide sequence are selected against, and the functional constraints on that region are high.
- $\omega \approx 1$: indicates that the sequence is most likely undergoing neutral evolution; both non-synonymous and synonymous substitutions can accumulate in approximately equal amounts without having a positive or negative impact on the function of the protein.
- $\omega > 1$: indicates that positive selection has occurred; this happens when beneficial mutations are selectively retained by natural selection.

See Figure 1.7 for an illustration of non-synonymous and synonymous substitutions and the way in which they are used to infer selective pressure.

1.8.2 Methods of Selective Pressure Analysis

Early methods of selective pressure analysis were predominantly distance-based (Miyata and Yasunaga 1980, Li *et al.* 1985, Nei and Gojobori 1986, Li 1993, Comeron 1995). These methods generally involved counting non-synonymous and synonymous differences between two sequences, while correcting for multiple substitutions at the same site and site type, (whether this be synonymous or non-synonymous).

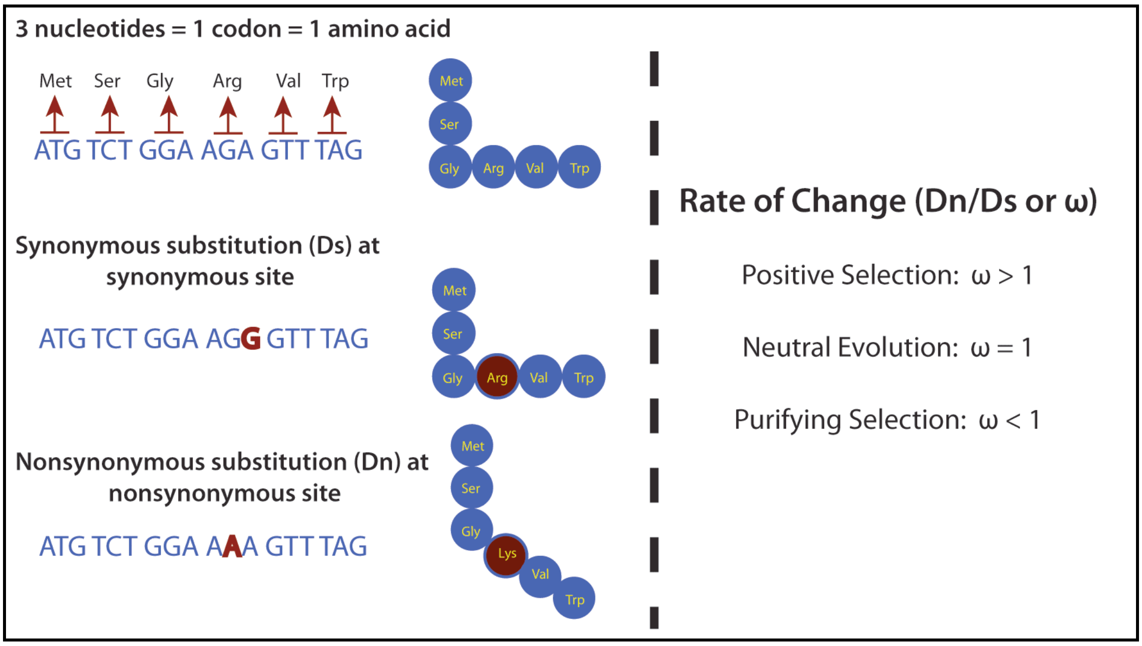

3 nucleotides = 1 codon = 1 amino acid
Met Ser Gly Arg Val Trp
ATG TCT GGA AGA GTT TAG
Met
Ser
Gly Arg Val Trp
Synonymous substitution (Ds) at synonymous site
ATG TCT GGA AGG GTT TAG
Met
Ser
Gly Arg Val Trp
Nonsynonymous substitution (Dn) at nonsynonymous site
ATG TCT GGA AAA GTT TAG
Met
Ser
Gly
Lys
Val
Trp
Rate of Change (Dn/Ds or ω)
Positive Selection: ω > 1
Neutral Evolution: ω = 1
Purifying Selection: ω < 1

Figure 1.7: Principles of selective pressure analysis.

Above is a cartoon depiction of the effects of synonymous and non-synonymous substitutions and how this is used in selective pressure analyis. To the left of the dashed line are three simplified protein structures, representing (from top to bottom): the wild type protein, a protein produced by a mutant form of the gene with a synonymous substitution, and a protein produced by a mutant with a non-synonymous substitution. Because the synonymous substitution does not affect the resultant protein, it is much less susceptible to the effects of natural selection and can provide a baseline estimate of the substitution rate. This forms the basis for selective pressure analysis by comparison of the rate of non-synonymous substitution per non-synonymous site (Dn) with the rate of synonymous substitution per synonymous site (Ds). The ratio of these two values (Dn/Ds or ω) can be used to identify the action of positive selection: a value of ω less than 1 is indicative of negative or purifying selection; a value of ω equal to 1 points to neutral evolution; and a value of ω greater than 1 indicates that the gene has undergone positive selection for functional shift. Figure taken with permission from Morgan (2012).

Confounding factors include transition-transversion rate ratios, which can cause overestimation of Ds and consequently underestimation of ω, (Li *et al.* 1985), and codon usage bias, which can cause underestimation of Ds and consequently overestimation of ω, (Bielawski *et al.* 2000). Later methods attempted to account for transition-transversion rate ratios, (Li *et al.* 1985, Nei and Gojobori 1986, Li 1993, Comeron 1995), and codon usage bias (Yang *et al.* 2000).

However, Yang and Nielsen (2000) found maximum likelihood methods of selective pressure analysis using explicit codon evolutionary models to have superior performance over distance-based methods: using a simulated dataset of codon sequences, the authors estimated the mean square error (MSE) of ω as estimated by three distance methods and maximum likelihood, finding that maximum likelihood had overall lowest MSE; especially so in the case of a strong transition-transversion bias.

Another popular method of selective pressure analysis makes use of sliding windows to identify sites under positive selection, (Endo *et al.* 1996, Hurst and Pál 2001, Fares *et al.* 2002). Software incorporating sliding window analysis include K-Estimator (Comeron 1999) and SWAPSC (Fares 2004). However, sliding window analysis is prone to artefactual positive selection: Schmid and Yang (2008) used sliding window analysis with simulated data and found that variation in Ds was greater than that of Dn, even in cases where the simulated Dn and Ds were constant along the full length of the sequence. This was attributed to multiple testing of codon sites as the sliding window moves along the sequence (Schmid and Yang 2008). Lin *et al.* (2011) have performed a sliding window analysis that does correct for multiple testing, using at gene level a Bonferroni correction for the number of windows tested in the given gene, and at genome level the Benjamini and Hochberg false discovery rate controlling method (Benjamini and Hochberg 1995). Nevertheless, the question of the most appropriate window size and offset to use remains an issue, since this can have a significant impact on the outcome of the sliding window analysis (Schmid and Yang 2008).

Due to their significant issues, neither sliding window nor averaging-based methods (i.e. methods that calculate a single Dn/Ds value for the entire gene/protein) are in common use in contemporary analyses of selective pressure variation. An example of the averaging method is the simplest model in the Codeml package, known as Model 0 (Goldman and Yang 1994, Yang 1998). This model calculates a single ω ratio for the entire protein-coding sequence (i.e. an averaging-based method). This is the model with the fewest estimated parameters that is implemented in Codeml, and is very unlikely to identify positive selection: this is because positive selection in a few sites in a sequence can be swamped by neutral evolution and/or negative selection in the majority of sites (Anisimova *et al.* 2001). Only very rarely will the signal of positive selection be strong enough for the value of ω to exceed 1 over the whole length of the gene sequence.

1.8.3 Selective Pressure Analysis with Codeml

More sophisticated models than Model 0 are therefore required to detect most cases of positive selection. Codeml from the PAML package (Yang 2007, Yang 1997) is arguably the most widely-used maximum likelihood method for selective pressure analysis. It supports a variety of models, each representing a particular class of evolutionary scenarios. The site-specific and branch-site specific models of Codeml allow for heterogeneity in ω by having multiple rate ratio (ω) categories across sites and across both branches and sites, respectively.

Codon Models of Evolution and Their Comparison

The site-specific models implemented in Codeml (Nielsen and Yang 1998, Yang *et al.* 2000, Wong *et al.* 2004) detect selective pressure variation on a site-specific basis: codon sites are grouped into site classes, each of which is associated with a particular ω value. Model 1a has two site classes with values $\omega < 1$ and $\omega = 1$. Model 2a has three site classes with values $\omega < 1$, $\omega = 1$ and $\omega > 1$. The discrete models Model 3 (k=2) and Model 3 (k=3), where k refers to the number of site classes in each model, allow for varying ω values to be estimated from the data. Model 7 (β), Model 8 (β&$\omega > 1$) and Model 8a (β&$\omega = 1$) each have ten site classes with $\omega \leq 1$. Models 8 and 8a both have an extra site class, with $\omega > 1$ and $\omega = 1$, respectively. For each of these models the likelihood of the model, given the data, is estimated, as is the proportion of sites (p) in the aligned homologous sequences that fit into each of these site classes.

The branch-site models implemented in Codeml (Yang and Nielsen 2002, Zhang *et al.* 2005) detect rate heterogeneity occurring in specific sites on particular lineages, or branches, of the phylogeny. Lineages of interest are chosen by the user *a priori* and each is labelled as a **foreground lineage**, with unlabelled lineages each serving as a **background lineage**. For example, if one wanted to test for positive selection in the Eutherian lineage, that is the lineage that should be labelled as foreground and all non-Eutherian species (e.g. chicken, platypus, opossum) would be considered as background taxa. Model A and Model A Null both have 3 ω value estimates across 4 site classes — see Zhang *et al.* (2005). Model A has constrained ω values as follows: $\omega_0 < 1$, $\omega_1 = 1$ and $\omega_2 \geq 1$, while Model A Null is the null model of Model A, with ω values: $\omega_0 < 1$, $\omega_1 = 1$ and $\omega_2 = 1$. The four site classes are 0, 1, 2a and 2b. The ω estimates for site classes 0 and 1 are respectively ω_0 and ω_1, in both foreground and background lineages. For site class 2a, the background lineage is constrained to ω_0, while the foreground lineage is constrained to ω_2. For site class 2b, the background lineage is constrained to ω_1, while the foreground lineage is again constrained to ω_2. Model B is similar to Model A, but differs in that the ω values are free to vary and are estimated entirely based on the data (Yang and Nielsen 2002); this is therefore the most parameter rich of all models. See Figure 1.8 for a graphical depiction of each of the currently available Codeml models and their parameters.

The models are then analysed using a **likelihood ratio test** (LRT) to compare the null hypothesis of neutral evolution and/or negative selection against the alternative hypothesis allowing for positive selection. The test statistic (*D*) of the LRT is calculated as shown in Equation 1.8.

$$D = c \cdot (\ln \ell_1 - \ln \ell_0)$$

Equation 1.8: Likelihood ratio test statistic.

The likelihood ratio test statistic *D* is shown in terms of the log-likelihood of the null and alternative models ($\ln \ell_0$ and $\ln \ell_1$, respectively), as well as a constant factor *c* specific to each pair of Codeml models.

For a pair of nested models, D follows a χ^2 distribution with degrees of freedom (ν) equal to the difference in the number of free parameters in the two models. The goodness-of-fit of the null and alternative models can be compared using a χ^2 test. See Table 1.1 for a summary of the LRTs that can be performed on currently available Codeml models.

Where Codeml infers that positive selection has occurred and the LRT confirms this, the positive selection model estimates may be used to estimate the posterior probability (PP) that a given site belongs to the category of positively selected sites. This can be done using a **Naïve empirical Bayes** (**NEB**) approach (Nielsen and Yang 1998) or a **Bayes empirical Bayes** (**BEB**) method (Yang *et al.* 2005). However, the Naïve empirical Bayes fails to account for uncertainties in parameter estimates and may infer sites to be positively selected with a high rate of false-positives, while on the other hand Bayes empirical Bayes can identify amino acid sites under positive selection more accurately. Indeed, in a simulation study by Yang *et al.* (2005), using 100 replicate datasets with a 30 taxon tree, in which half of the sites had an ω value of 1 and half were given an ω value of 1.5, BEB positively selected sites under Model 8 captured 9% of true-positives and 5% of false-positives, while NEB inference under the same model identified 16% of positively selected sites, but had a false-positive rate of 14%. The authors of this study concluded that where possible, positively selected sites inferred under Bayes empirical Bayes should be used in preference to those inferred by Naïve empirical Bayes.

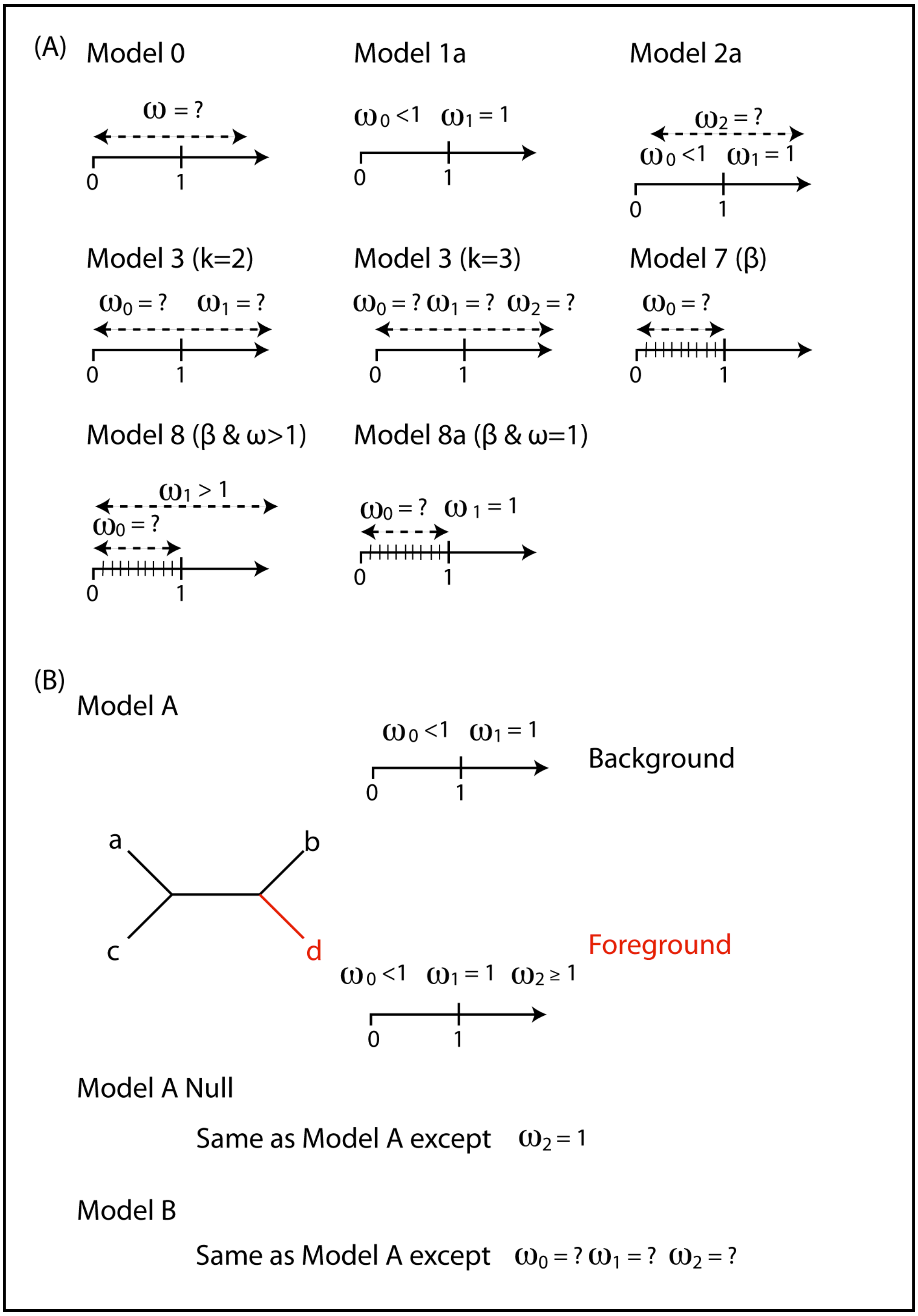

(A)
Model 0
$\omega = ?$
0
1
Model 1a
$\omega_0 < 1$ $\omega_1 = 1$
0
1
Model 2a
$\omega_2 = ?$
$\omega_0 < 1$ $\omega_1 = 1$
0
1
Model 3 (k=2)
$\omega_0 = ?$ $\omega_1 = ?$
0
1
Model 3 (k=3)
$\omega_0 = ?$ $\omega_1 = ?$ $\omega_2 = ?$
0
1
Model 7 (β)
$\omega_0 = ?$
0
1
Model 8 (β & ω>1)
$\omega_1 > 1$
$\omega_0 = ?$
0
1
Model 8a (β & ω=1)
$\omega_0 = ?$ $\omega_1 = 1$
0
1
(B)
Model A
$\omega_0 < 1$ $\omega_1 = 1$
Background
0
1
a
b
c
d
Foreground
$\omega_0 < 1$ $\omega_1 = 1$ $\omega_2 \geq 1$
0
1
Model A Null
Same as Model A except $\omega_2 = 1$
Model B
Same as Model A except $\omega_0 = ?$ $\omega_1 = ?$ $\omega_2 = ?$

Figure 1.8: Models of codon substitution implemented in Codeml.

Above is a graphical depiction of the codon substitution models used for selective pressure analysis by Codeml from the PAML package (Yang 2007, Yang 1997). Each model is shown with the set of possible values for *ω* under that model. These models are described in detail in Section 1.8.3. (A) Model 0 and the site-specific codon models in order of increasing complexity, with Model 0 being the simplest model implemented by Codeml and Model 8 being the most complex site-specific model used in this thesis. (B) Branch-site specific models are shown, with a phylogenetic tree illustrating the foreground lineage in red. Figure adapted with permission from Morgan (2012).

Table 1.1: Likelihood ratio tests used in Codeml analysis.

Null Model	Alternative Model	v	χ^2 Critical Value
Model 0	Model 3 (k=2)	2	5.99
Model 3 (k=2)	Model 3 (k=3)	1	1.00
Model 1a	Model 2a	2	5.99
Model 7 (β)	Model 8 (β&ω>1)	2	5.99
Model 8a (β&ω=1)	Model 8 (β&ω>1)	2	2.71
Model 1a	Model A	2	5.99
Model A Null	Model A	2	3.84
Model 3 (k=2)	Model B	2	5.99

Table 1.1 Legend:

This table shows the key statistics for each of the likelihood ratio tests used in the selective pressure analysis with Codeml in this thesis. The **Null Model** and **Alternative Model** columns show, respectively, the null and alternative models being tested in each LRT. The column marked **v** shows the degrees of freedom in the χ^2 distribution being used for each LRT, while the χ^2 **Critical Value** column indicates the critical value that must be exceeded by the LRT test statistic D (see Equation 1.8) in order for the null model to be rejected at a 5% significance level.

Accuracy and Power of Codeml

Although the models implemented in Codeml have been found to be both accurate and powerful, (Anisimova *et al.* 2001, Zhai *et al.* 2012), a number of factors can confound the identification of positively selected genes and amino acid sites. In some cases, biased gene conversion may give the appearance of positive selection, (Galtier and Duret 2007). For example, Galtier *et al.* (2009) found the proportion of AT→GC substitutions to be significantly greater in lineages undergoing accelerated amino acid evolution (43.6%), than in non-accelerating lineages (39.6%). Recombination itself can cause spurious signals of positive selection, although Codeml is robust to lower levels of recombination, (Anisimova *et al.* 2003). Population bottlenecks can also lead to false-positive identification of positive selection, because slightly deleterious substitutions are more likely to become fixed in a smaller population (Ohta 1973). Relaxation of selective constraint could similarly lead to spurious identification of positive selection, since such a relaxation can affect the non-synonymous substitution rate. Indeed, a selective pressure analysis was performed by Crandall and Hillis (1997) to determine whether selective pressure had been relaxed on rhodopsin genes in cave-dwelling crayfish. (No significant relaxation was inferred in that case, leading the authors to suggest that the rhodopsins in question may have some additional function other than phototransduction.) In addition to these factors, a reduced rate of adaptation in species with lower effective populations, combined with a higher proportion of neutral (or nearly neutral) substitutions, conspire to make adaptive change more difficult to accurately detect in such species (Gossmann *et al.* 2012, Lartillot 2012).

There are some aspects of selective pressure analysis that can be controlled more easily. The power of selective pressure analysis with Codeml is known to be low with less than 7 sequences (Anisimova *et al.* 2001) or with less than 100 codon sites (Bielawski and Yang 2003). In order to avoid the local optimum problem, Codeml should be run multiple times with different initial ω values, with at least one initial ω value below one, and one inital ω value above one (Bielawski and Yang 2003).

Schneider *et al.* (2009) identified three areas of selective pressure analysis in which avoidable errors can occur, as follows:

- **sequence quality**: genes with less than 3X sequencing coverage were found to have inferred rates of positive selection more than 3 times greater than those genes with more than 3X coverage.

- **annotation quality**: positive selection was inferred for 7.6% of Ensembl "known" genes, as opposed to 14.6% of those genes inferred by the Ensembl gene prediction pipeline.

- **sequence alignment**: genes in multiple sequence alignments with a 100% Heads or Tails (HoT) score (Landan and Graur 2007) were inferred to have positive selection in 9.5% of cases, as opposed to 15% of those gene families with a HoT score less than 100%.

Such errors can be minimised by using high-quality, high-coverage sequence data, by favouring more confidently annotated sequences and by careful choice of alignment method, respectively (Schneider *et al.* 2009). Increases in the quantity and quality of available genomic sequences in recent years (Flicek *et al.* 2012) can ameliorate the first two issues, while the active development of alignment methods (Edgar 2004b, Löytynoja and Goldman 2008), and alignment performance comparisons (Fletcher and Yang 2010, Markova-Raina and Petrov 2011), can help with the third issue.

The codon-based methods of Codeml have come under criticism, however, with a number of studies demonstrating false-positive inferences of positive selection using Codeml (Friedman and Hughes 2007, Suzuki 2008, Nozawa *et al.* 2009). For example, Friedman and Hughes (2007) found that inference of positive selection was correlated significantly ($\alpha = 0.05$) with the GC content in the third codon position (GC3), mean synonymous substitution rate and the length of input gene sequences, but found no significant correlation (at a 5% significance level) of positive selection with either mean non-synonymous substitution rate or mean estimated ω value. As a rationale for these results, the authors suggested that the correlation with sequence length may simply be due to the greater power of the LRT with more data; that the synonymous substitution rate was affected in their dataset by the presence of paralogous groups, in addition to the neutral substitution rate; and that the correlation with GC3 content reflected a more general GC bias in mammalian genomes. Zhai *et al.* (2012) sought to address the concerns of Friedman and Hughes (2007) in particular by repeating their analysis. However, Zhai *et al.* (2012) could not replicate the results of Friedman and Hughes (2007), finding instead that inference of positive selection was correlated, as expected, with mean estimated ω value and gene sequence length. Zhai *et al.* (2012) found no significant correlation between inference of positive selection and mean non-synonymous substitution rate, mean synonymous substitution rate or GC content.

Hughes and Friedman (2005) criticised the use of codon-based methods such as Codeml to infer positive selection, arguing that a certain proportion of sites are expected to have an ω value estimate greater than one, due to chance variation in the rate of synonymous and non-synonymous substitution. Indeed, the authors found that proportions of non-synonymous (p_N) and synonymous (p_S) substitution substitutions in nearly 2 million codons across more than 4 thousand genes from the aligned genomes of *S. cerevisiae* and *S. paradoxus*, found that 6.47% of codons had an excess of p_N over p_S, significantly greater than 6.31% — the rate expected by chance in that dataset ($P < 0.001$). Zhai *et al.* (2012) agreed that in many cases variation in synonymous and non-synonymous substitution rate affects the estimation of ω, but argued that in most such cases the LRT will not achieve significance in testing alternative models as a result of this variation.

Hughes (2007) argued that many of the models used lack a prior biological hypothesis. Zhai *et al.* (2012) countered that LRTs with site models do indeed have a set of prior statistical hyphotheses based on biological understanding of the effect of natural selection on nucleotide and amino acid sequences.

As an example, the authors cite codon models Model 1a (having two site classes with $\omega_0 < 1$ and $\omega_1 = 1$) and Model 2a (including an additional site class such that $\omega_2 > 1$). If an LRT rejects the null model, Model 1a, in this case, such a result constitutes statistical evidence that non-synonymous substitutions are preferentially fixed at some codon sites in the gene. The prior statistical hypothesis — that the ratio of non-synonymous to synonymous substitutions is greater than unity in some codons of a gene — can therefore be tested without reference to the specific process of adaptive substitution at play, or indeed prior knowledge of which sites are under positive selective pressure (Zhai *et al.* 2012).

While Zhai *et al.* (2012) appear to have addressed the narrow methodological concerns raised by Friedman and Hughes in their 2005 and 2007 studies, a broader epistemic question was raised by Hughes and Friedman (2010): the comparison of non-synonymous and synonymous substitutions is only suited to a scenario in which repeated amino acid substitution in the same gene has been favoured by natural selection. Further to this, they proposed several alternative scenarios of adaptive novelty that can not be modelled by codon-based comparisons of non-synonymous and synonymous substitutions. For a given gene these can include a single amino acid substitution, an insertion or deletion, a loss of splice signals that in turn leads to exon loss, gene fusion and a change in the regulation of the gene that affect its expression. It is to this last source of evolutionary novelty — gene regulation — that we turn next.

1.9 Gene Regulation

Sequence evolution plays a key role in creating evolutionary novelty, but the picture is inevitably incomplete without considering the role played by the regulation of gene expression (Hughes and Friedman 2010). Mechanisms of gene regulation are many and varied, including chromatin remodelling (Saha *et al.* 2006), cell signalling (Shin and Manley 2004), transcriptional regulation by transcription factors (Spitz and Furlong 2012), alternative splicing (Barash *et al.* 2010, Roy and Gilbert 2006), alternative polyadenylation (Di Giammartino *et al.* 2011), post-transcriptional regulation by miRNA (Pasquinelli 2012), DNA methylation (Suzuki and Bird 2008) and genomic imprinting (Ferguson-Smith 2011).

Transcription factors and miRNAs are arguably better studied than other forms of gene regulation (Chen and Rajewsky 2007). Both modes of gene regulation are believed to be critical for the development of multicellular eukaryotes. On one hand, transcription factors are known to play a role in co-ordination of body plans, (Levine and Tjian 2003, Davidson and Erwin 2006, Peter and Davidson 2011), while on the other, miRNAs have been shown to maintain phenotypic precision through tight control of the expression of target genes, (Cohen *et al.* 2006, Hornstein and Shomron 2006, Peterson *et al.* 2009a). The sequence-specific nature of regulatory binding sites for transcription factors and miRNAs might be seen as supporting a *prima facie* argument for the identification of such binding sites using bioinformatics. However, the reality is more complex — neither transcriptional nor miRNA regulation are completely understood — and the *in silico* identification of binding sites remains challenging both for transcription factors and for miRNAs. For example, high turnover of individual transcription factor binding sites can hinder their identification with conservation footprinting (Wasserman and Sandelin 2004), while miRNA-target interactions, being incompletely understood, cannot be modelled accurately without introducing some false-positives (Bartel 2009).

1.9.1 Transcription Factors

A transcription factor is a protein that binds to a specific DNA sequence to regulate the expression of its target gene(s). This can include general transcription factors (GTFs) that bind to core promoter elements, as well as activators and repressors that respectively enhance and repress target gene expression (Lee and Young 2000). By controlling the timing and extent of gene expression, transcription factors facilitate the development of complex structures in multicellular eukaryotes, such as the choriallantoic placenta (Cross *et al.* 2002, Rawn and Cross 2008).

Significant challenges remain with *in silico* identification of transcription factor binding sites (Wasserman and Sandelin 2004, Tompa *et al.* 2005). Such methods typically make use of sequence motifs and phylogenetic footprinting (i.e. positive weighting of conserved orthologous sequence) to identify putative transcription factor binding sites, but false-positive prediction rates remain intolerably high, such that the majority of predicted transcription factor binding sites are not biologically meaningful — due perhaps to the difficulties in modelling the relevant interactions, the high turnover of individual binding sites, or difficulties in validating interactions *in vivo* (Wasserman and Sandelin 2004, Tompa *et al.* 2005).

1.9.2 MicroRNAs

MicroRNAs are a class of short non-coding molecules of RNA of ~22 nucleotides in length that regulate genes post-transcriptionally by binding to their expressed transcripts (Lewin 2008). MiRNAs are believed to have widespread influence on gene expression levels (Bartel and Chen 2004), and estimates of the proportion of human genes regulated by these short RNA molecules range from 25% (Lewis *et al.* 2005) to 50% of the transcriptome (Shomron *et al.* 2009).

Among regulatory mechanisms, miRNA regulation lends itself to bioinformatics studies because these short non-coding RNA molecules are more highly conserved and more amenable to computational binding site prediction than transcription factors (Chen and Rajewsky 2007). Furthermore, miRNAs have a characteristic that is highly desirable in evolutionary biology — they are unique and rarely secondarily lost (Sperling and Peterson 2009). Once a miRNA emerges in a lineage it is rarely subsequently lost, therefore the patterns of presence and absence themselves reveal an appreciable amount of robust information on evolutionary history (Sperling and Peterson 2009).

MiRNAs were initially discovered by Lee *et al.* (1993), who found that the *C. elegans* gene lin-4 produced a short RNA that was complementary to a number of sites in the 3′ untranslated region (3′ UTR) sequence of lin-14. These sites had previously been identified as mediating the regulation of lin-14 by lin-4 (Wightman *et al.* 1991). It was proposed that the binding of lin-4 to the 3′ UTR of lin-14 repressed the translation of that gene, reducing levels of LIN-14 protein without significantly affecting lin-14 mRNA levels (Lee *et al.* 1993, Wightman *et al.* 1993).

It appeared as if this regulatory mechanism was specific to this one region in *C. elegans* until Reinhart *et al.* (2000) demonstrated that the *C. elegans* gene let-7 regulated several genes in a similar manner, and Pasquinelli *et al.* (2000) identified homologous sequences in multiple species including human, raising the prospect that this mode of regulation was more widespread than previously believed. This new class of regulators was termed miRNAs (Lagos-Quintana *et al.* 2001) and within a few years miRNAs were recognised as playing a significant role in gene regulation and development (Bartel 2004). MiRNAs are present in both animals (Brennecke *et al.* 2003, Johnston and Hobert 2003, Chen 2004a) and plants (Aukerman and Sakai 2003, Emery *et al.* 2003, Chen 2004b), although there are differences in miRNA biogenesis and modes of action between the two kingdoms (Millar and Waterhouse 2005, Axtell *et al.* 2011). The miRNA database miRBase (Griffiths-Jones 2004, Griffiths-Jones 2006, Griffiths-Jones *et al.* 2008, Kozomara and Griffiths-Jones 2011) now has over a thousand annotated miRNAs in human, and more than 20,000 entries in total.

1.9.3 Biogenesis of MicroRNAs

Vertebrate miRNA biogenesis can occur through a number of pathways, as shown in Figure 1.9. The primary miRNA transcript (pri-miRNA) may be derived from an intron (Calin *et al.* 2002) or exon (Tam 2001) of a non-coding transcript. Pri-miRNAs may also be derived from the intron (Bortolin-Cavaille *et al.* 2009) or exon (Lu *et al.* 2008) of a protein-coding transcript, although the latter is controversial (Berezikov *et al.* 2010) and is not believed to be a significant source of miRNAs (Berezikov 2011). MiRNAs are typically transcribed by RNA Polymerase II (Lee *et al.* 2004). Borchert *et al.* (2006) found evidence that the chromosome 19 microRNA cluster (C19MC) is transcribed by RNA polymerase III, but this was challenged by Bortolin-Cavaillé (2009), who demonstrated that this miRNA cluster lies within the introns of a non-coding gene that is transcribed by RNA Polymerase II.

In canonical miRNA biogenesis, the pri-miRNA is processed by the Microprocessor complex, to produce one or more hairpin loop precursor miRNAs (pre-miRNAs) of around 70 bases in length (Denli *et al.* 2004, Gregory *et al.* 2004). The Microprocessor complex contains Drosha (Lee *et al.* 2003) and its cofactor DiGeorge syndrome critical region gene 8 (DGCR8) in human, or Pasha in *C. elegans* and *D. melanogaster* (Han 2004, Landthaler *et al.* 2004).

Alternatively, pre-miRNAs may arise in a Drosha-independent manner as mirtrons. These are pre-miRNAs that form from spliced introns that undergo Lariat debranching (Ldbr) before progressing along the canonical miRNA biogenesis pathway (Kim *et al.* 2009). Mirtrons were first observed in *D. melanogaster* and *C. elegans* (Okamura *et al.* 2007, Ruby *et al.* 2007a) and shortly afterwards in mammals (Berezikov *et al.* 2007, Sibley *et al.* 2011).

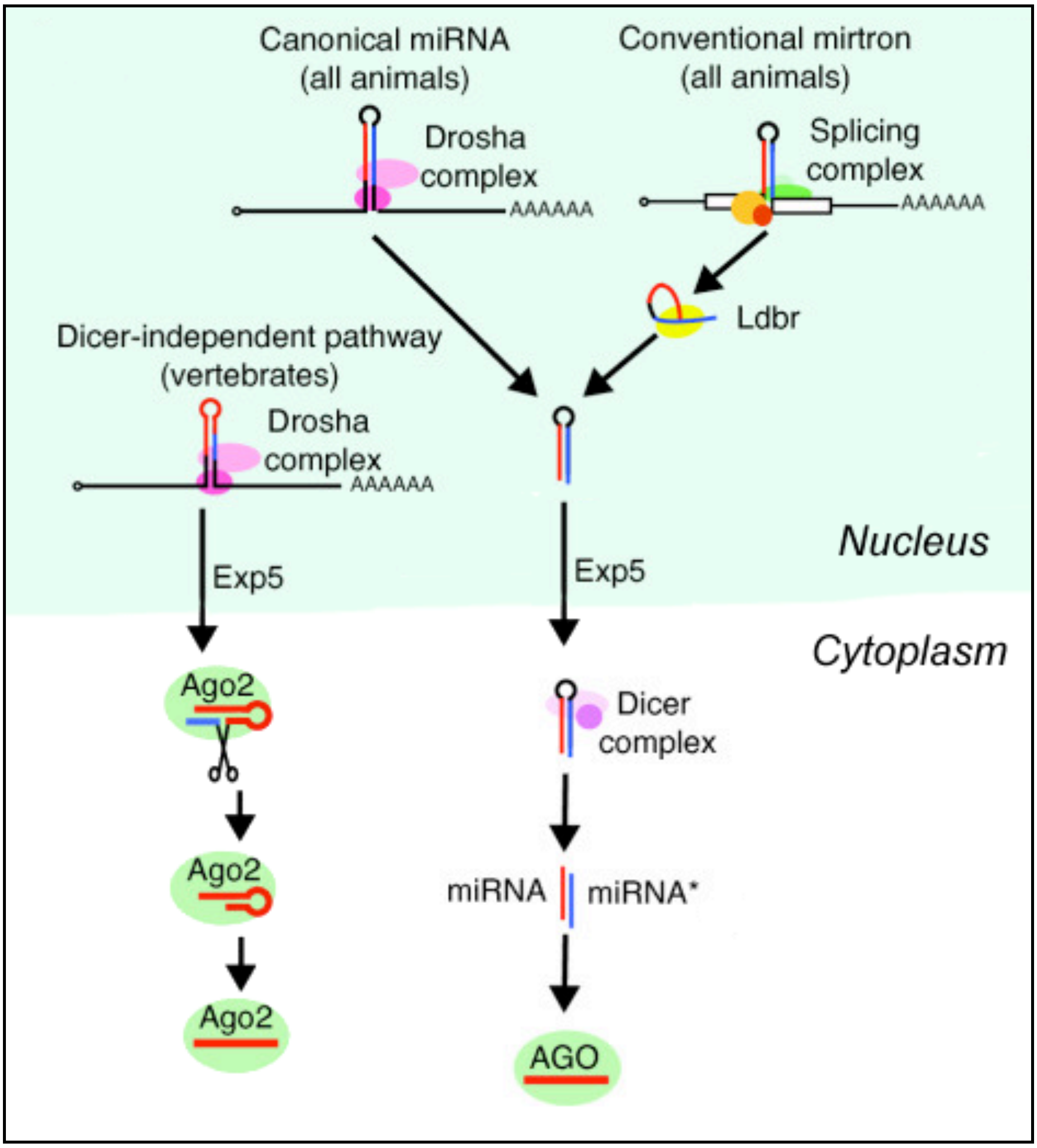


Figure 1.9: Pathways of microRNA biogenesis in vertebrates.

Three main miRNA biogenesis pathways are known in vertebrates: canonical miRNA biogenesis in which the primary miRNA is processed by the Drosha in the Microprocessor complex, the resulting precursor miRNA is exported from the nucleus, spliced by the Dicer complex and recruited by Argonaute; conventional mirtron biogenesis, in which a spliced intron undergoes Lariat debranching (Ldbr) before progressing along the canonical miRNA biogenesis pathway; and a Dicer-independent pathway, in which the pre-miRNA is both spliced and recruited by Argonaute-2. Figure adapted with permission from Axtell (2011).

Whatever the mechanism by which it arises, the pre-miRNA is exported from the nucleus to the cytoplasm in a process mediated by Exportin-5 (Kim 2004). Typically the pre-miRNA is then spliced by the Dicer complex, removing the hairpin loop and leaving only a miRNA duplex of about 22 bases in length (Bernstein *et al.* 2001, Ketting 2001, Hutvagner *et al.* 2001). Following cleavage by Dicer, the miRNA duplex is recruited by Argonaute, (Liu *et al.* 2004, Meister *et al.* 2004) and incorporated into the RNA-induced silencing complex (RISC), (Gregory *et al.* 2005). One strand — known as the **guide strand** — is retained in the RISC as a single-stranded mature miRNA, while its complementary strand — the **passenger strand** — is degraded (Schwarz *et al.* 2003, Khvorova *et al.* 2003). It is important to note that this bias in strand retention is by no means universal, and functional activity has been reported for many so-called passenger strands (Yang *et al.* 2011). A Dicer-independent pathway has also been observed in vertebrates, in which Argonaute-2 (Ago2) in particular serves a dual purpose of recruiting the pre-miRNA and mediating its cleavage (Cheloufi *et al.* 2010, Cifuentes *et al.* 2010, Yang *et al.* 2010). Once the mature miRNA is incorporated into the RISC, it is then available to assist the RISC in regulating mRNA transcripts (Axtell *et al.* 2011).

1.9.4 MicroRNA-Target Interactions

Although it was known from their initial discovery that miRNAs modulate the expression of their targets in a sequence-specific manner (Lee *et al.* 1993, Wightman *et al.* 1993), our understanding of the nature of miRNA-target interactions has evolved since that time to form a picture that is more rich, more complex, and in some ways more incomplete, (Bartel 2004, Brennecke *et al.* 2005, Breving and Esquela-Kerscher 2010, Bartel 2009, Berezikov 2011). The precise mechanisms by which miRNAs regulate their targets remain incompletely understood and are the subject of debate, although the most commonly reported mode of action remains post-transcriptional repression of the target mRNA by the mature miRNA in conjunction with RISC (Morozova *et al.* 2012).

The canonical function of a miRNA is to downregulate expression of a specific mRNA target by inducing cleavage or translational repression, where base pairing of the miRNA to the target mRNA is extensive or partial, respectively (Bartel 2009).

Cleavage of the target is more commonly observed in plants (Rhoades *et al.* 2002), while in animals extensive complementarity of miRNA and target is believed to be comparatively rare, and consequently translational repression is considered the more dominant mode of action, (Yekta *et al.* 2004, Davis *et al.* 2005). The repressive action of miRNAs is supported by the observation by Farh *et al.* (2005) that mRNAs that are conserved targets of a given miRNA tend to be expressed in the same tissue as the cognate miRNA, but that such miRNA-target pairs tend to have inverse expression patterns. For example, an accumulation of miR-1 and miR-133, after cell cycle arrest in differentiating myoblasts, was found to be associated with reduced expression of genes predicted to be targets of these two miRNAs (Farh *et al.* 2005).

In addition to repression activity, miRNAs have also been observed to upregulate expression (Vasudevan *et al.* 2007, Ørom *et al.* 2008) and to be linked to direct tuning of gene expression rather than outright repression (Poy *et al.* 2004, Karres *et al.* 2007). This expands the role of miRNA to both that of an expression switch and an expression fine-tuner, in a threshold-dependent manner (Mukherji *et al.* 2011). Reflecting this potentially broader set of functions, it has been proposed that miRNAs allow for more precise control of gene expression (Bartel and Chen 2004, Peterson *et al.* 2009a). The latter hypothesised that miRNAs effectively offer greater phenotypic precision (and therefore greater heritability) by exercising tight control over the level of target protein produced, without which elaborate morphological structures might not be possible.

MicroRNA-Target Site Features

Perhaps the most important feature of miRNA-target interactions — and probably the most utilised in early miRNA-target prediction (Lewis *et al.* 2003, John *et al.* 2004, Grün *et al.* 2005) — is binding of the microRNA seed region to the target site. The **microRNA seed** was defined by Lewis *et al.* (2003) as comprising bases 2-8 of the miRNA, numbering from the 5′ end. A **seed match site** is a region of a miRNA-target sequence that is complementary to the miRNA seed sequence. By identifying seed match sites that were conserved in orthologous sequences across human, mouse and rat, Lewis *et al.* (2003) predicted miRNA-target sites with a signal-to-noise ratio of 7:2, demonstrating that conserved seed match sites were a good indication of a functional miRNA-target site. Subsequent studies confirmed the importance of seed binding to miRNA-target interactions (Doench and Sharp 2004, Kloosterman *et al.* 2004, Krützfeldt *et al.* 2005), and its utility in predicting such interactions (Lewis *et al.* 2005, Krek *et al.* 2005, Brennecke *et al.* 2005).

Figure 1.10 shows the known types of seed match miRNA-target site and their relative proportions among conserved vertebrate miRNA targets as observed by Friedman *et al.* (2009). Both Figure 1.10 and the nomenclature used for seed match miRNA-target sites are taken from Bartel (2009). The **canonical miRNA-target site** types include 8mer, 7mer-m8 and 7mer-A1 sites (Lewis *et al.* 2005). The 8mer and 7mer-m8 sites both have Watson-Crick (WC) complementarity to sites 2-8 of the miRNA, while the 7mer-A1 site only has WC complementarity to miRNA sites 2-7. Both the 8mer and the 7mer-A1 sites are flanked by an adenosine opposite miRNA site 1, while a 7mer-m8 site lacks this. The 8mer is considered to be more indicative of a functional miRNA-target site than the 7mer site types, while a 7mer-m8 site is considered to have more efficacy than a 7mer-A1 site. The 6mer target sites pairing to miRNA sites 2-7 and 3-8 are considered to be a **marginal miRNA-target site** type and by themselves are not considered to be particularly discriminative (Grimson *et al.* 2007, Nielsen *et al.* 2007). Preferential conservation of miRNA-target site types reflects this, with the canonical and marginal site types, ordered by decreasing conservation signal-to-background ratios, as follows: 8mer, 7mer-m8, 7mer-A1 and 6mer (Friedman *et al.* 2009).

In some seed match sites, supplementary WC base pairing occurs between sites 12-17 of the miRNA and the target, in what is known as a **3′-supplementary site** (Brennecke *et al.* 2005, Lim *et al.* 2005). Base pairing between sites 12-17 can also compensate for imperfect binding of the seed region in a **3′-compensatory site** (Yekta *et al.* 2004), although these are believed to be relatively rare (Friedman *et al.* 2009).

Although seed match miRNA targets remain the most easily discriminated using sequence data, due to their (often conserved) high sequence complementarity to the miRNA seed sequence, binding of the miRNA seed region to the target is not always required for miRNA regulation. Functional miRNA targets have been observed that lack a seed region (Vella 2004, Hammell *et al.* 2008, Lal *et al.* 2009, Chi *et al.* 2009), and in some cases G:U base pairing in the seed region is tolerated (Didiano and Hobert 2006). Indeed, up to 27% of Ago-mRNA clusters in murine brain are **orphan clusters**: regions bound by Ago-miRNA complexes but lacking a canonical seed match site for the set of top Ago-bound miRNAs (Chi *et al.* 2009). A reliance on seed match sites is also widely acknowledged to introduce many false-positive miRNA-target predictions (Bartel 2009, Peter 2010), with as many as two-thirds of predicted targets by some seed match methods being found to be non-responsive to knockout of the cognate miRNA (Baek *et al.* 2008). This has led some authors to conclude that despite its success in identifying a significant class of functional miRNA-target sites, seed region binding is neither necessary nor sufficient for miRNA-target interactions generally (Hammell *et al.* 2008, Chi *et al.* 2012).

To complicate the picture further, at least two other classes of miRNA-target site type have been observed: the centred site and the bulge site. The **centred miRNA-target site** was observed by Shin *et al.* (2010) using pooled microarray data from miRNA and short interfering RNA (siRNA) transfection experiments (Lim *et al.* 2005, Birmingham *et al.* 2006, Jackson 2006, Jackson *et al.* 2006, Schwarz *et al.* 2006, Grimson *et al.* 2007, Anderson *et al.* 2008). This site type is characterised by WC base-pairing of 11 contiguous sites towards the centre of the miRNA (i.e. miRNA sites 4-14 or 5-15) and a lack of substantial pairing at either end of the miRNA (see Figure 1.11, part A). Centred miRNA-target sites were shown not to be an artefact of extended base pairing of canonical seed match sites, and were estimated to be about as common as 3′-compensatory sites (Shin *et al.* 2010). Friedman *et al.* (2009) estimated the proportion of conserved seed match miRNA-target sites that are 3′-compensatory to be about 2%, so while centred sites are an important new class of miRNA-target site, they are likely to be quite rare.

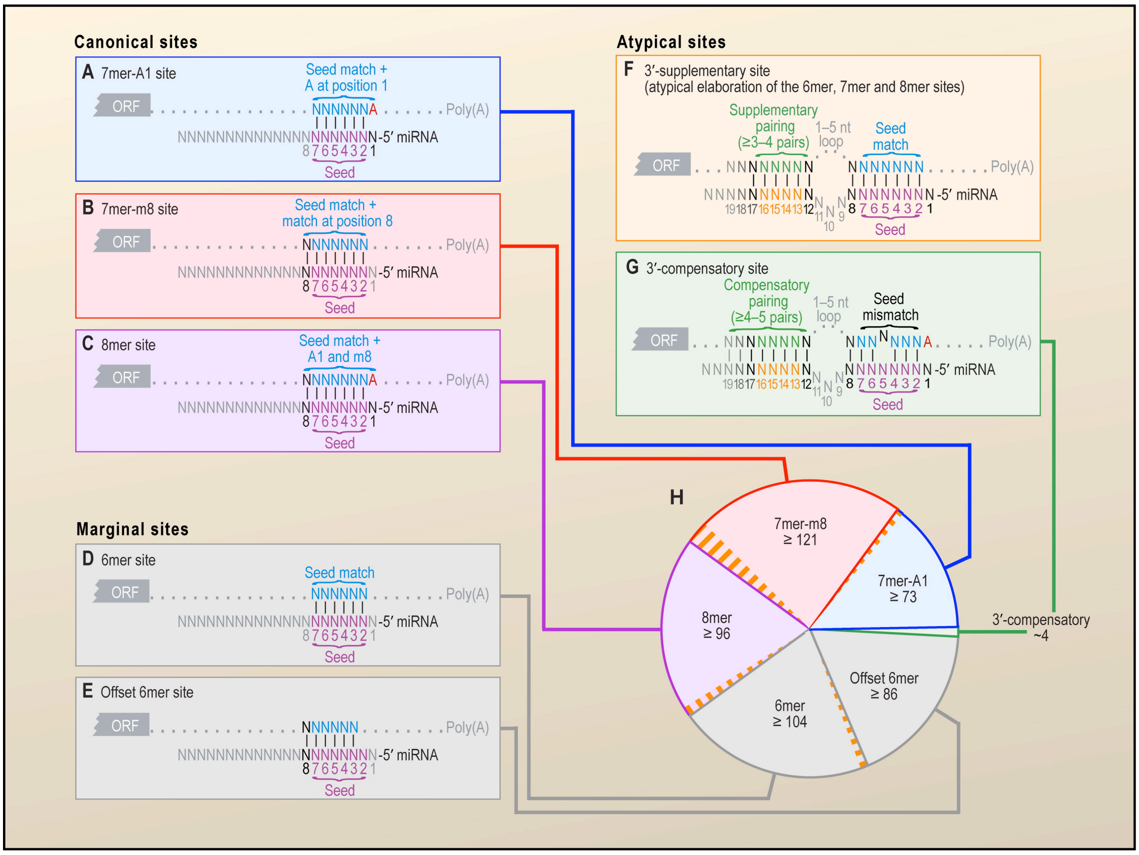


Figure 1.10: MicroRNA-target site types.

Above is a depiction of a number of known miRNA-target site types with seed binding. The canonical site types include (A) 7mer-A1, (B) 7mer-m8 and (C) 8mer, while the more marginal 6mer and offset 6mer are shown in (D) and (E). Watson-Crick binding of the miRNA to its target may occur towards the 3′ end of the miRNA, either to supplement seed binding (F) or to compensate for imperfect seed binding (G). The pie chart in (H) shows the proportions of each site type as estimated by Friedman *et al.* (2009). Figure taken from Bartel (2009). Permission obtained via the Copyright Clearance Center.

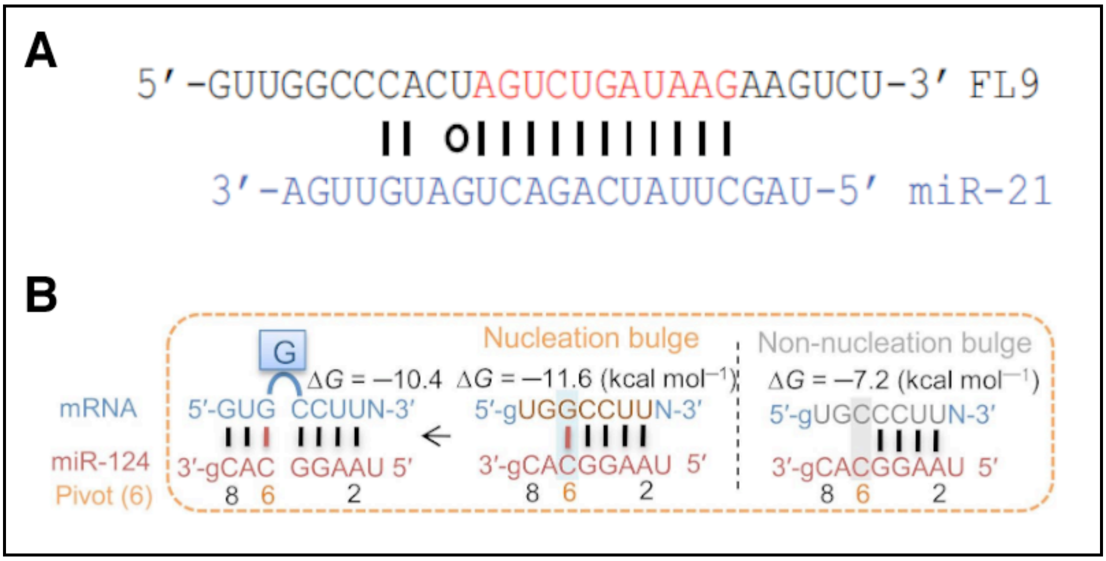


Figure 1.11: Alternative miRNA-target site types.

Shown are two observed miRNA-target site types that differ from those shown in Figure 1.10: (A) centred sites, in which base pairing of miRNA to target sequence is observed principally towards the centre of the miRNA (Shin *et al.* 2010), and (B) miRNA bulge sites, in which a functional miRNA-target interaction with a target sequence of imperfect complementarity to the miRNA seed is facilitated by a transitional nucleation bulge, such that miRNA sites 2-6 then bind to their corresponding nucleotides in the target sequence, the nucleotide opposite miRNA site 6 forms the bulge, and site 6 binds to the nucleotide on the target sequence immediately 5′ of the bulge nucleotide (Chi *et al.* 2012). Part (A) is taken from Shin *et al.* (2010), while part (B) is taken from Chi *et al.* (2012). Both figures are reproduced with permission obtained from the Copyright Clearance Center.

The **miRNA-target bulge site** was reported by Chi *et al.* (2012), who made use of a genome-wide Argonaute HITS-CLIP survey of Ago-miRNA binding sites in mouse brain cells (Chi *et al.* 2009). In a bulge site, miRNA sites 2-5 and 6-8 are bound to the target, but the target mRNA has a bulge between miRNA sites 5 and 6. Chi *et al.* (2012), proposed that binding of a miRNA to a bulge site is preceded by a transitional nucleation bulge, such that miRNA sites 2-6 bind to their corresponding nucleotides in the target sequence. The nucleotide opposite miRNA site 6 then forms the bulge, while site 6 binds to the target sequence immediately 5′ of the bulge nucleotide (Chi *et al.* 2012), see Figure 1.11, part B.

Using data from a meta-analysis by (Khan *et al.* 2009) of several miRNA transfection microarray experiments, Chi *et al.* (2012) showed that repression of genes mediated by bulge sites was significant (Kolmogorov-Smirnov test, P = 0.03), although less significant than that of genes containing canonical seed match sites (Kolmogorov-Smirnov test, $P = 1.05 \times 10^{-24}$). These authors pointed out that bulge sites have been observed before (Ha *et al.* 1996, Vella 2004), but were not recognised as a widespread miRNA-target site type. At least 15% of targets in mouse brain were found to contain miRNA-target bulge sites, which suggests that this site type is relatively common (Chi *et al.* 2012).

MicroRNA-Target Site Context

Perhaps the most significant aspect of miRNA-target site context is its location on the target mRNA. The 3′ untranslated region (UTR) of the target mRNA has historically been considered to host the majority of functional miRNA-target sites in animals, although whether this is due to mechanistic effects or historical contingency has been questioned (Bartel 2004). Most, if not all, of the earliest examples of animal miRNA-target sites were found in the 3′ UTR of their target gene (Bartel 2004). In addition, many of the early methods of miRNA-target prediction made use of conservation filters, and conserved miRNA-target sites would be more easy to detect against the relatively unconserved background of 3′ UTRs, compared to the more constrained coding sequence regions. Their relative ease of detection in 3′ UTRs would have exacerbated any initial acquisition bias, if one existed (Rigoutsos 2009, Peter 2010).

One suggested mechanistic explanation for enrichment of miRNA targets in 3′ UTR sequences is that ribosomal activity in coding regions prevents the miRISC from binding effectively (Bartel 2004). A mechanistic explanation is supported by Gu *et al.* (2009), who modified the stop codon of a reporter gene to effectively extend the expressed region of the transcript so as to include a miRNA-target site located in the 3′ UTR. This had the effect of eliminating repression of the reporter gene by the cognate miRNA. By addition of rare codons upstream of the miRNA-target site, ribosome stalling was induced and miRNA translational inhibition was restored (Gu *et al.* 2009).

Nevertheless, evidence has accumulated that at least some miRNA targets are to be found in coding regions (Duursma *et al.* 2008, Forman *et al.* 2008, Tay *et al.* 2008, Zhou *et al.* 2009, Elcheva *et al.* 2009, Huang *et al.* 2010, Fang and Rajewsky 2011) and in 5′ UTRs (Lytle *et al.* 2007, Ørom *et al.* 2008, Lee *et al.* 2009, Zhou *et al.* 2009). Schnall-Levin *et al.* (2011), in particular, noted a tendency for targeting by miR-181 to be unusually effective in the repeat-rich coding regions of its target genes retinoblastoma 1 (RB1) and RB-associated KRAB zinc finger (RBAK) (Schnall-Levin *et al.* 2011). The efficacy of miRNA-target sites in coding regions has been found to be marginal relative to those in 3′ UTRs; in a miRNA transfection experiment, the mean $\log_2$ fold change of target mRNA expression was -0.08 for genes with a miRNA-target site in their coding region, and -0.175 for genes with a miRNA-target site in their 3′ UTR (Baek *et al.* 2008). It has been proposed that miRNA-target sites in coding regions will be more effective for mRNAs that are not efficiently translated (Bartel 2009).

More local target features associated with functional miRNA-target sites include high local AU content of the target sequence and proximity of the miRNA-target site to either end of the 3′ UTR (Grimson *et al.* 2007, Majoros and Ohler 2007).[3]

[3] Grimson et al. 2007 did note one exception to this: miRNA-target sites in a 3′ UTR that are within ~15 base pairs of the stop codon were less preferentially conserved and were also found to be less effective than other target sites in the 3′ UTR at mediating downregulation.

The 3′ UTR length of a gene indicates whether it is likely to be the target of a miRNA, such that highly targeted genes tend to have longer 3′ UTRs (Stark *et al.* 2005). Indeed, even isoforms of the same gene with a longer 3′ UTR are more likely to be subjected to miRNA regulation (Sandberg *et al.* 2008). Structural accessibility of the mRNA in the vicinity of the putative target site has been shown to have a positive effect on miRNA-target interactions (Long *et al.* 2007, Kertesz *et al.* 2007). The presence of multiple miRNA-target sites on the same target sequence was found to have an additive effect on target repression, with synergistic effects being observed when miRNA-target sites were in close proximity (i.e. 8-40 bases) (Grimson *et al.* 2007, Saetrom *et al.* 2007, Hon and Zhang 2007).

A broader aspect of a given miRNA-target interaction is the cellular context in which it takes place, and the relative abundance of competing targets in the cell. The competing **target abundance** (TA) — defined by Garcia *et al.* (2011) as the number of nonoverlapping canonical miRNA-target sites in the available 3′ UTR sequences of the given transcriptome — has been shown to affect individual miRNA-targeting efficacy and specificity (Anderson *et al.* 2008, Arvey *et al.* 2010). For example, in a bioinformatics analysis following a set of miRNA transfection experiments, Arvey *et al.* (2010) found a significant correlation between difference in TA and difference in levels of downregulation (Spearman's rank correlation coefficient, ρ=0.59; P < 10^{-15}). Target abundance has also been shown to be a discriminative feature for identifying miRNA-target sites (Garcia *et al.* 2011, Ragan *et al.* 2011). Garcia *et al.* (2011) estimated TA from genomic data and weighted these according to the relative abundance of each mRNA in the cell of interest, but noted that in practice the genomic TA and cell-specific TA were highly correlated (Spearman's R^2 = 0.9813), suggesting that for practical purposes genomic TA may be informative for identifying genuine miRNA targets (Garcia *et al.* 2011).

MicroRNA Anti-Targets and MicroRNA Sponges

Given the short length of many miRNA seeds, it is not unreasonable to expect that many mRNAs in any given transcriptome will contain miRNA seed match sites by chance.[4] Where repression of a given gene would be deleterious, natural selection will favour the avoidance of seed matches in the gene sequence — since such deleterious repression would affect the survival of the organism — and the gene will become an anti-target (Bartel and Chen 2004). Genes that are more broadly expressed tend to have shorter 3′ UTRs and lower density of predicted miRNA-target sites than genes with more specific expression patterns (Farh *et al.* 2005, Stark *et al.* 2005).

In a given cellular context, the abundance of alternative targets has a negative effect on the efficacy of individual miRNA-target interactions. This has been exploited by the use of artificial miRNAs, known as microRNA sponges, that compete with a specific miRNA and effectively reverse repression of its targets (Franco-Zorrilla *et al.* 2007, Ebert *et al.* 2007). It has also been suggested that many putative miRNA-target genes are natural competitive inhibitors of specific miRNA-target interactions, i.e. natural miRNA sponges (Seitz 2009, Ebert and Sharp 2010).

1.9.5 Evolution of MicroRNAs

MiRNAs are believed to have emerged more or less continuously throughout the metazoa, with substitutions and secondary loss occurring only rarely (Wheeler *et al.* 2009). Significant expansions of the genomic complement of miRNAs are believed to have coincided with the ancestral lineage in Bilateria (Prochnik *et al.* 2006, Hertel *et al.* 2006), vertebrates (Heimberg *et al.* 2008, Hertel *et al.* 2006, Wang *et al.* 2010), Eutheria (Hertel *et al.* 2006) and primates (Wang *et al.* 2010).

[4] Assuming a miRNA seed 7mer and a random target DNA sequence in which each nucleotide is one of the 4 possible bases (i.e. A, C, G or T), one should expect to find a complementary match of the 7mer on average once every 4^7 = 16,384 bases, simply by chance.

In animals, *de novo* emergence of functional miRNAs is considered to be more likely than for protein-coding genes, since a functional miRNA need only be a transcribed locus that, when expressed, forms a hairpin that is recognised by the Microprocessor complex (Chen and Rajewsky 2007). A preliminary analysis by Chen and Rajewsky (2007) found that the set of miRNAs were distributed quite randomly in sequence space, which tentatively indicates that *de novo* emergence is a non-trivial source of new miRNAs. The presence in the human genome of large numbers of sequences capable of forming a hairpin structure (Bentwich *et al.* 2005) lends further support to a model of random emergence of miRNAs (Svoboda and Di Cara 2006, Tanzer and Stadler 2004).

A new miRNA can also emerge from the antisense strand of an existing miRNA (Stark *et al.* 2008, Tyler *et al.* 2008). This could be facilitated by the fact that the different arms of a miRNA hairpin are not always completely complementary to each other, so the emerging miRNA sequence could differ from that of the original miRNA (Berezikov 2011).

Duplication of existing miRNAs is also considered to be a significant source of novelty (Hertel *et al.* 2006); in plants, an inverted duplication model has been proposed — in which an inverted duplication event creates the upstream and downstream arms of the novel precursor sequence in each strand of DNA, forming a hairpin structure with the complementary sequences formed by the inverted duplication — that would account for the emergence of a hairpin structure with high WC complementarity between the hairpin arms (Allen *et al.* 2004). Yuan *et al.* (2011) reported that transposable elements and segmental duplications both play a role in the origin of new miRNAs.

Animal miRNAs have been noted for their tendency to form clusters (Lagos-Quintana 2003, Altuvia 2005, Yu *et al.* 2006), and the presence of a significant number of paralogs among these clusters points to the effect of segmental duplication in their expansion — of 326 human miRNAs in miRBase 7.1 (then known as the miRNA registry), 148 were found in 52 clusters, which contained 9 paralogous groups (Yu *et al.* 2006).

Shomron *et al.* (2009) has suggested that such clustering allows novel miRNAs to hitchhike on existing transcriptional mechanisms; this is supported by the finding that clustered miRNAs tend to be co-expressed with each other (Liang *et al.* 2007). Another way in which newly evolved miRNAs might hitchhike with existing transcriptional machinery is as an intronic miRNA: a miRNA residing within the intron of a host gene (Lin *et al.* 2006). About 25-40% of known miRNA genes reside within introns (Rodriguez *et al.* 2004, Shomron *et al.* 2009). However, not all intronic miRNAs are dependent on their host gene for expression; some have their own promoter regions (Martinez *et al.* 2008, Ozsolak *et al.* 2008). Introns may be a significant source of new miRNAs, mediated by a process called **intronic exaptation**, in which the miRNA emerges within a pre-existing intron and acquires independent functionality, with or without the retention of the original host gene (Campo-Paysaa *et al.* 2011).

Once a novel miRNA has emerged, there are several documented mechanisms by which it can undergo evolutionary change (Berezikov 2011).

- **Nucleotide substitutions** within miRNAs are rare, but when they do occur, they tend to happen away from the functionally critical seed region (Wheeler *et al.* 2009).

- **Seed shifting** occurs if the location of the seed has changed in either the 5′ or 3′ direction (Wheeler *et al.* 2009), and is relatively common between distantly related species (Grimson *et al.* 2008, Marco *et al.* 2010).

- **Arm switching** involves a swap in roles between the guide strand and passenger strand, such that the former passenger strand becomes the new guide strand (Okamura *et al.* 2008, de Wit *et al.* 2009). Recent studies have shown that the roles of guide and passenger strands are more plastic than previously thought: the passenger strand has been observed to accumulate to physiologically relevant levels (Ruby *et al.* 2007b), to associate with Argonaute and regulate targets (Okamura *et al.* 2008), and, for relative expression of the nominal guide and passenger strands to vary considerably between tissues (Kuchenbauer *et al.* 2011).

- **Hairpin shifting** is a more drastic mode of miRNA evolution, in which a new miRNA hairpin evolves from one arm of an existing hairpin and adjacent upstream or downstream sequence (de Wit *et al.* 2009).

1.9.6 Evolution of MicroRNA Targets

A novel miRNA needs novel binding sites through which to exert its effects. Given the relatively short length of complementarity between miRNAs and their target sites, novel binding sites can arise relatively quickly on an evolutionary timescale. Chen and Rajewsky (2007) reviewed several studies on the emergence of new binding sites (Stone and Wray 2001, Berg *et al.* 2004, MacArthur and Brookfield 2004, Durrett and Schmidt 2007). Chen and Rajewsky (2007) concluded that neutral evolution of new binding sites is most likely too slow — of the order of millions years for a 7mer to evolve from a null binding sequence — to constitute an appreciable source of novel miRNA binding sites, but that positive selection on 'pre-sites' (i.e. sites with one seed mismatch) could generate a new miRNA-target site in ~375,000 years (Chen and Rajewsky 2007). Shomron *et al.* (2009) estimated that a kilobase of 3′ UTR sequence would typically contain several such pre-sites, and gave an estimate for the emergence of a functional miRNA-target site of ~200,000 years.

Chen and Rajewsky (2007) further proposed a model of miRNA emergence followed by the evolution of miRNA-target interactions. In this model, the miRNA is initially expressed at low levels in a tissue-specific manner. The miRNA then exerts selective pressure on its targets (Chen and Rajewsky 2006, Saunders *et al.* 2007) and anti-targets[5] (Bartel and Chen 2004, Farh *et al.* 2005, Stark *et al.* 2005), while these in turn subject the miRNA itself to selective pressure (Sempere *et al.* 2006). As the miRNA, targets and anti-targets co-evolve, the miRNA increases in expression and may become more broadly expressed (Chen and Rajewsky 2007). This is supported by evidence that many miRNAs have tissue-specific expression patterns, (e.g. hsa-miR-516-5p, which was found, in a microarray experiment, to have placenta-specific expression) (Liang *et al.* 2007); while more broadly conserved miRNAs tend to be expressed at higher levels, (e.g. among miRNA genes in 6 *Drosophila* species, those

[5] A microRNA anti-target was defined by Bartel and Chen (2004) as a mRNA that is a non-target of a microRNA and is under selective pressure to remain a non-target.

miRNAs conserved in at least 5 species were more highly expressed than miRNAs conserved in fewer species) (Lu *et al.* 2008), an attribute associated with less tissue-specific expression (e.g. hsa-miR-98, which was found to have high expression but low tissue-specificity) (Landgraf *et al.* 2007).

1.9.7 MicroRNA-Target Prediction

As the mechanisms of interaction of miRNAs with their binding sites are not completely understood, miRNA-target prediction (miTP) remains incomplete. Table 1.2 lists some currently available miRNA-target prediction methods. Although it is by no means an exhaustive listing, it is representative of current methods, and includes the most popular general-purpose miRNA-target prediction algorithms. Many of the miRNA-target interaction features discussed in Section 1.9.4 are used by one or more of these methods. The problem of miRNA-target prediction is mainly approached from three general directions: evolutionary conservation, sequence features and thermodynamic considerations (Hammell 2010).

MicroRNA-Target Prediction using Evolutionary Conservation

With many early miRNA-target prediction studies, conservation of putative miRNA-target sites was extensively used as a filter for miRNA-target predictions, and many conserved candidate miRNA targets were identified as a result (Lewis *et al.* 2003, Brennecke *et al.* 2005, Krek *et al.* 2005, Lewis *et al.* 2005). It was reasoned that preferential conservation of the sequence of a putative binding site — and particularly that of the region complementary to the miRNA seed — indicates that the site is undergoing negative selection to preserve its function (Lewis *et al.* 2003). Conservation of the region flanking the putative miRNA-target site has also been shown to have some predictive value (Wen *et al.* 2011).

Requiring all species of interest to have an aligned orthologous miRNA-target site reduces the power of this method, since in many cases such a miRNA-target site may be absent due to lineage-specific duplication or loss, or simply due to lack of sequencing coverage or errors in alignment (Friedman *et al.* 2009). For example, in the 28-way genomic alignment by Miller *et al.* (2007) that was used for conservation analysis by Friedman *et al.* (2009), 10 genomes had a sequence coverage of 2X or less, which theoretically entails the lack of at least 10% of genomic sequence (Lander and Waterman 1988). Along with the challenging process of genomic alignment undertaken by Miller *et al.* (2007), this would have inevitably reduced the number of orthologous sequences included in the final 28-way alignment, even in the ideal case of perfectly conserved one-to-one orthologs. The effects of duplication and loss would compound this issue, to the extent that requiring perfect conservation of miRNA-target sites across all 28 species would markedly reduce the ability of this method to identify conserved miRNA-target sites (Friedman *et al.* 2009).

In response to this, methods developed more recently estimate instead the branch length over which a predicted miRNA-target site has been conserved using a phylogenetic tree of the species of interest, such that a longer branch length indicates greater probability of preferential conservation (Kheradpour *et al.* 2007, Friedman *et al.* 2009, Vejnar and Zdobnov 2012). Furthermore, Rajewsky (2006) distinguishes between conservation of miRNA-target sites and conservation of miRNA-target genes. While a miRNA-target site may be lost in one region of the target gene and arise in another region, the regulatory interaction itself is conserved. On the basis of this distinction, Rajewsky (2006) further suggests a strategy for conservation analysis of miRNA-targets: requiring conservation of miRNA-target sites for more closely related species, while requiring only the conservation of a miRNA-target interaction for more distantly related species.

While conservation filters have been very good at identifying miRNA targets, a large proportion of non-conserved miRNA-target sites are functional (Farh *et al.* 2005), so continuing to rely on conservation filters would severely limit the potential to identify novel miRNA targets (Bartel 2009).

MicroRNA-Target Prediction using Thermodynamics

In addition to identifying miRNA-target sites using primary sequence features, a parallel but complementary approach has explored the effect of RNA secondary structure thermodynamics on miRNA binding site efficacy (Long *et al.* 2007). Many early miRNA-target prediction (miTP) methods incorporated estimation of the differential of Gibbs free energy (ΔG) of miRNA-target hybridisation (Enright *et al.* 2003, Lewis *et al.* 2003, Stark *et al.* 2003), and some miTP methods rely predominantly on considerations of miRNA-target binding energy (Rehmsmeier *et al.* 2004, Krüger and Rehmsmeier 2006, Thadani and Tammi 2006).

The ViennaRNA suite of programs (Gruber *et al.* 2008, Hofacker *et al.* 1994, Hofacker 2003) has been widely used in miRNA-target prediction methods for estimation of folding energy of miRNA-target duplexes (Enright *et al.* 2003, Lewis *et al.* 2003, Rehmsmeier *et al.* 2004, Krüger and Rehmsmeier 2006, Thadani and Tammi 2006) and for modelling target site accessibility in addition to hybridisation energy (Kertesz *et al.* 2007, Sturm *et al.* 2010, Vejnar and Zdobnov 2012). The ViennaRNA package in turn makes use of the **minimum free energy** (MFE) algorithm devised by Zuker and Stiegler (1981), the ensemble partition function created by McCaskill (1990), the suboptimal folding algorithm described by Wuchty *et al.* (1999), and the RNA energy parameters in Mathews *et al.* (2004).

Unfortunately, large-scale studies of the effect of miRNAs on their targets have found the overall hybridisation energy of a miRNA and its putative target to be a poor indicator of a functional miRNA-target interaction (Grimson *et al.* 2007, Baek *et al.* 2008). In contrast, thermodynamic stability of the seed duplex has been shown to be a significant factor in siRNA interactions (Ui-Tei *et al.* 2008); siRNA modes of action include seed binding similar to that observed in miRNAs (Lim *et al.* 2005). Subsequently, estimated **seed-pairing stability** (SPS) of the miRNA-target duplex was shown to be a determinant of strength and specificity of repression in miRNA-target interactions (Garcia *et al.* 2011).

MicroRNA-Target Prediction using Sequence Features

As discussed in Section 1.9.4, the importance of miRNA seed binding to the prediction of miRNA-target sites was clear from its use in many early methods (Lewis *et al.* 2003, John *et al.* 2004, Grün *et al.* 2005). This importance is underscored by its continued use in more recently developed miRNA-target prediction methods (Liu *et al.* 2010, Sturm *et al.* 2010, Marín and Vaníček 2011 and 2012, Vejnar and Zdobnov 2012).

Binding of the miRNA in miRNA sites 12-17 to the target has been used to identify 3′-supplementary sites (Brennecke *et al.* 2005, Lim *et al.* 2005) and 3′-compensatory sites (Yekta *et al.* 2004). However, 3′-supplementary pairing is of relatively limited use, since these types of sites can be identified by a simple seed match analysis (Grimson *et al.* 2007), while 3′-compensatory sites are believed to be a relatively rare occurrence (Friedman *et al.* 2009). Some methods omit these site types (Lewis *et al.* 2005, Gaidatzis *et al.* 2007), while others incorporate them with such strict criteria that they have only marginal impact on miRNA-target predictions (Krek *et al.* 2005, Grimson *et al.* 2007).

To the knowledge of this author, no miRNA-target prediction method has yet incorporated bulge sites (Chi *et al.* 2012) or centred sites (Shin *et al.* 2010). Whether they are used in future methods will probably depend more on their predictive value than on their importance to miRNA-target binding. For example, the 6mer seed match site is considered to be very common among functional miRNA-target sites, but is not regarded as useful for miRNA-target prediction because of the preponderance of false-positive 6mer seed matches (Bartel 2009, Ellwanger *et al.* 2011). This emphasises the limitations of our current understanding of miRNA-target interactions, and shows that much remains to be done to improve that understanding (Bartel 2009). As miRNAs continue to be actively studied, progressively greater knowledge of these features is being used to improve the quality of miRNA binding site predictions.

Table 1.2: Current microRNA-target prediction methods.

miTP Method	Availability	Web Interface	Data Download	Standalone Software	References
DIANA-microT	diana.cslab.ece.ntua.gr/DianaTools/	Yes	Yes	No	Reczko *et al.* (2012), Reczko *et al.* (2011), Maragkakis *et al.* (2009a), Maragkakis *et al.* (2009b)
EIMMo	www.mirz.unibas.ch/ElMMo3/	Yes	Yes	No	Gaidatzis *et al.* (2007)
Hitsensor	Available on request	No	No	Yes	Zheng and Zhang (2010)
MicroTar	tiger.dbs.nus.edu.sg/microtar/	No	No	Yes	Thadani and Tammi (2006)
miRanda	www.microrna.org/	Yes	Yes	Yes	Betel *et al.* (2010), Betel *et al.* (2008), John *et al.* (2004), (Enright *et al.* 2003)
mirEE	didattica-online.polito.it/eda/miREE/	Yes	No	No	Reyes-Herrera *et al.* (2011)
miRmap	cegg.unige.ch/mirmap/	Yes	Yes	Yes	Vejnar and Zdobnov (2012)
MirTarget2	mirdb.org/miRDB/	Yes	Yes	No	Wang (2008), Wang and El Naqa (2008), Wang and Wang (2006)
MultiMiTar	www.isical.ac.in/~bioinfo_miu/	Yes	Yes	Yes	Mitra and Bandyopadhyay (2011), Bandyopadhyay and Mitra (2009)

Table 1.2: Current microRNA-target prediction methods. (continued)

miTP Method	Availability	Web Interface	Data Download	Standalone Software	References
PACCMIT	lcpt.epfl.ch/	No	Yes	No	Marín and Vaníček (2012), Marín and Vaníček (2011)
PicTar	pictar.mdc-berlin.de	Yes	Yes	No	Chen *et al.* (2006), Lall *et al.* (2006), Grün *et al.* (2005), Krek *et al.* (2005)
PITA	genie.weizmann.ac.il/pubs/mir07/	Yes	Yes	Yes	Kertesz *et al.* (2007)
RNA22	cm.jefferson.edu/rna22v1.0/	Yes	Yes	No	Miranda *et al.* (2006)
RNAhybrid	bibiserv.techfak.uni-bielefeld.de/	Yes	No	Yes	Krüger and Rehmsmeier (2006), Rehmsmeier *et al.*(2004)
SVMicrO	compgenomics.utsa.edu/svmicro.html	No	Yes	Yes	Liu *et al.* (2010)
TargetScan	www.targetscan.org/	Yes	Yes	Yes	Garcia *et al.* (2011), Friedman *et al.* (2009), Grimson *et al.*(2007), Lewis *et al.* (2005)
TargetSpy	www.targetspy.org/	Yes	Yes	Yes	Sturm *et al.* (2010)

1.10 Placenta

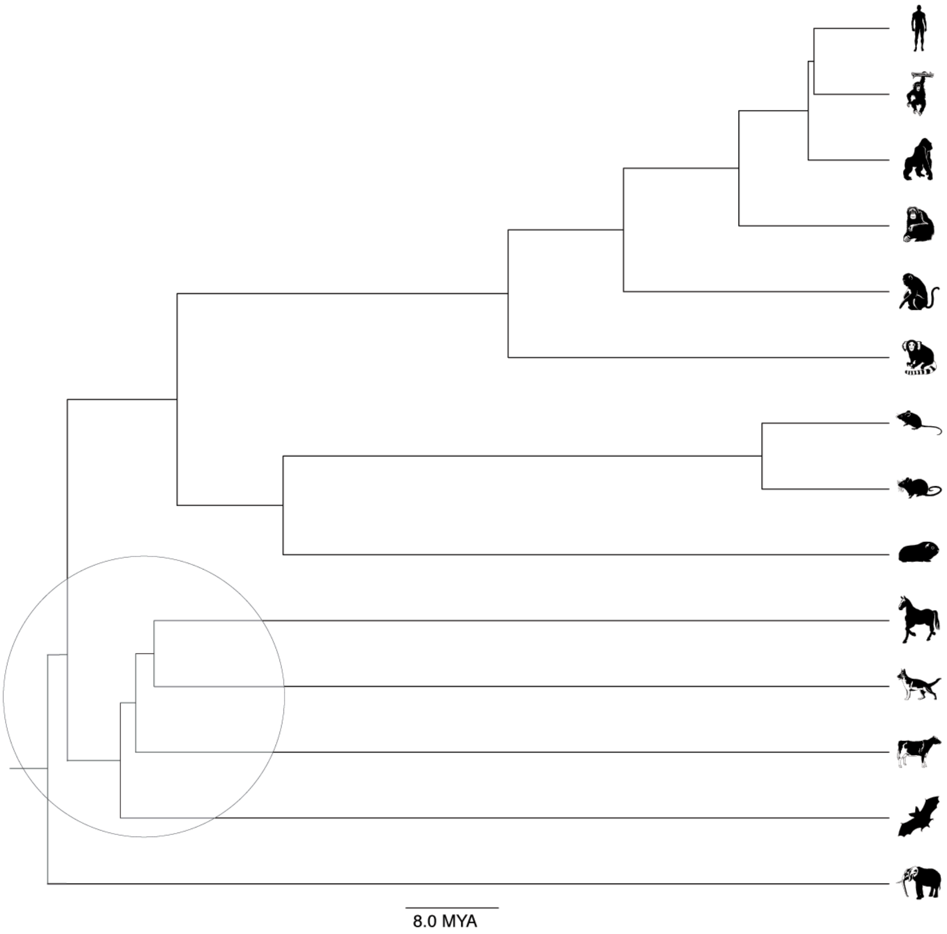
It is believed that the earliest Eutherian mammals evolved about 125 million years ago (MYA), and that the chorioallantoic placenta typical of Eutheria evolved (in evolutionary terms) shortly thereafter (Ji *et al.* 2002). See Figure 1.12 for a summary of placental characteristics across the species studied here. Formed from both foetal and maternal tissues, the chorioallantoic placenta forms an intimate connection between mother and offspring, mediating the transfer of nutrients, gases and waste products during gestation.

1.10.1 Evolution of the Placenta

The evolution of the placental structures appears to be intimately associated with the evolution of viviparity. As outlined by Crespi and Semeniuk (2004), viviparity carries several advantages: increases in offspring survival rates, birth weight and offspring vigour, as well as increased efficiency and flexibility in resource allocation by the mother.

Figure 1.12 Legend:

A phylogeny of the 14 placental mammal species in this study is shown overleaf, along with a breakdown of placental types by maternofaetal barrier, placental interdigitation and placental shape for each species. Species lacking a chorioallantoic placenta have been excluded. As with Figure 1.6, the phylogeny used is that of Benton and Donoghue (2007). The branches of the phylogeny are to scale except within the encircled area, where branch lengths are dilated to clarify the branching order used. The species in this phylogeny are as follows (from top to bottom): Human (*Homo sapiens*), Chimp (*Pan troglodytes*), Gorilla (*Gorilla gorilla*), Orangutan (*Pongo abelii*), Macaque (*Macaca mulatta*), Marmoset (*Callithrix jacchus*), Mouse (*Mus musculus*), Rat (*Rattus norvegicus*), Guinea Pig (*Cavia porcellus*), Horse (*Equus caballus*), Dog (*Canis familiaris*), Cow (*Bos taurus*), Bat (*Myotis lucifugus*) and Elephant (*Loxondonta africana*). Credit for all animal silhouettes goes to Dr Mary J. O'Connell, except for that of the bat, which is a public domain image sourced from Wikimedia Commons (http://commons.wikimedia.org). The information about placental types was taken from Kurt Benirschke's 'Comparative Placentation' website (Benirschke 2007) as cited in Elliot and Crespi (2009), except for that of the bat species *Myotis lucifugus*, which was taken from Gopalakrishna and Karim (1979).



Maternofoetal Barrier	Placental Interdigitation	Placental Shape
Hemochorial	Villous	Discoid
Hemochorial	Villous	Discoid
Hemochorial	Villous	Discoid
Hemochorial	Villous	Discoid
Hemochorial	Trabecular	Discoid
Hemochorial	Trabecular	Discoid
Hemochorial	Labyrinthine	Discoid
Hemochorial	Labyrinthine	Discoid
Hemochorial	Labyrinthine	Discoid
Epitheliochorial	Trabecular	Diffuse
Endotheliochorial	Labyrinthine	Zonary
Epitheliochorial	Villous	Cotyledonary
Hemochorial	Labyrinthine	Discoid
Endotheliochorial	Labyrinthine	Zonary

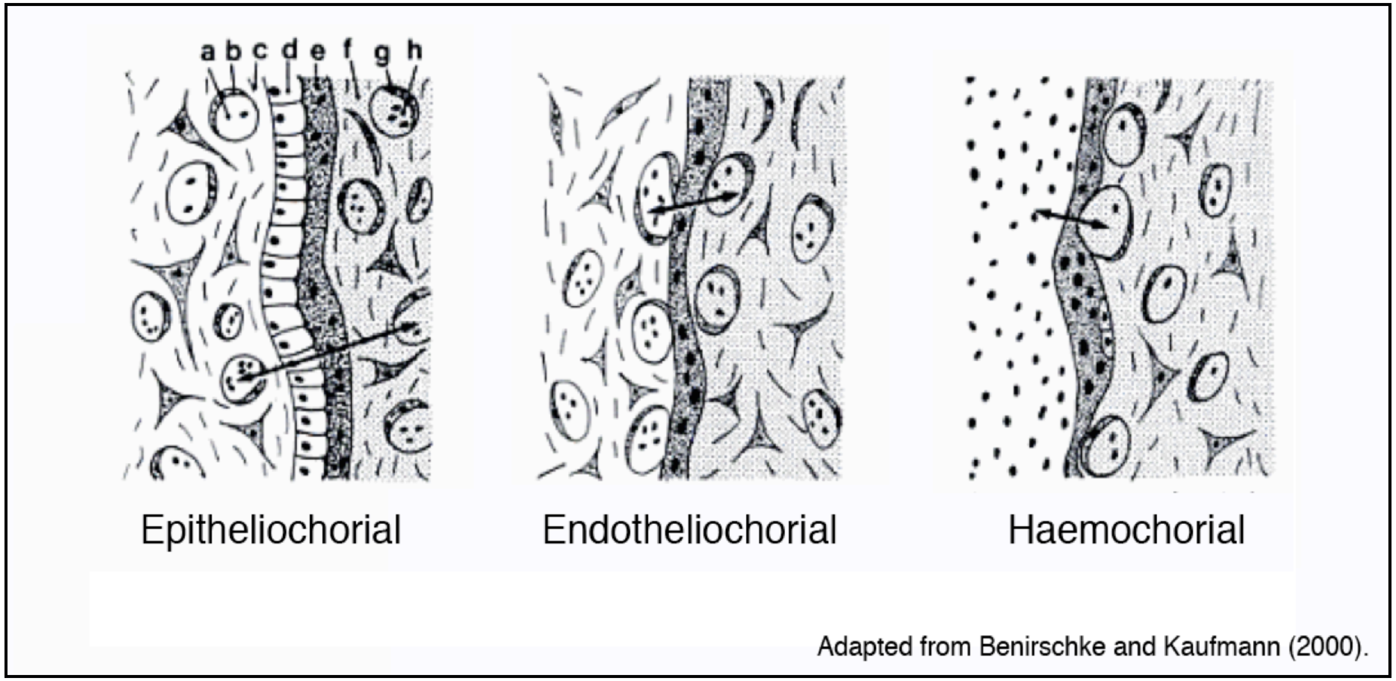
Figure 1.12: Placental characteristics of Eutherian species studied in this thesis.

That the chorioallantoic placenta has some compelling advantages is further supported by examples of viviparity and placental structures that have evolved independently in clades other than mammals, including some sharks (Hamlett 2005) and lizards (Flemming and Blackburn 2003).

The choriovitelline (or yolk-sac) placenta provides for limited exchange of nutrients in marsupials and monotremes, before the young are ensconced in a marsupium or laid in an egg, respectively. Eutherian mammals retain this yolk-sac placenta as a complement to the chorioallantoic placenta. In humans, it is responsible for nutrient exchange from the exocoelomic cavity in early pregnancy (Freyer and Renfree 2009), as well as playing a role in early haematopoiesis (King and Enders 1993). Rodents are a notable exception among Eutheria because, as in marsupials and monotremes, they retain a functional yolk-sac placenta in direct contact with the maternal endometrium until term (Freyer and Renfree 2009).

Notable exceptions to the yolk sac placenta among marsupials include the bandicoot, which has evolved a chorioallantoic placenta analogous to that found in Eutherian mammals (Tyndale-Biscoe and Renfree 1987). The bandicoot is an instructive example of the evolutionary rationale for the chorioallantoic placenta: accounting for its relatively short gestation period, neonates are more developed than comparable marsupials lacking the tissue, (Tyndale-Biscoe and CSIRO Publishing. 2005).

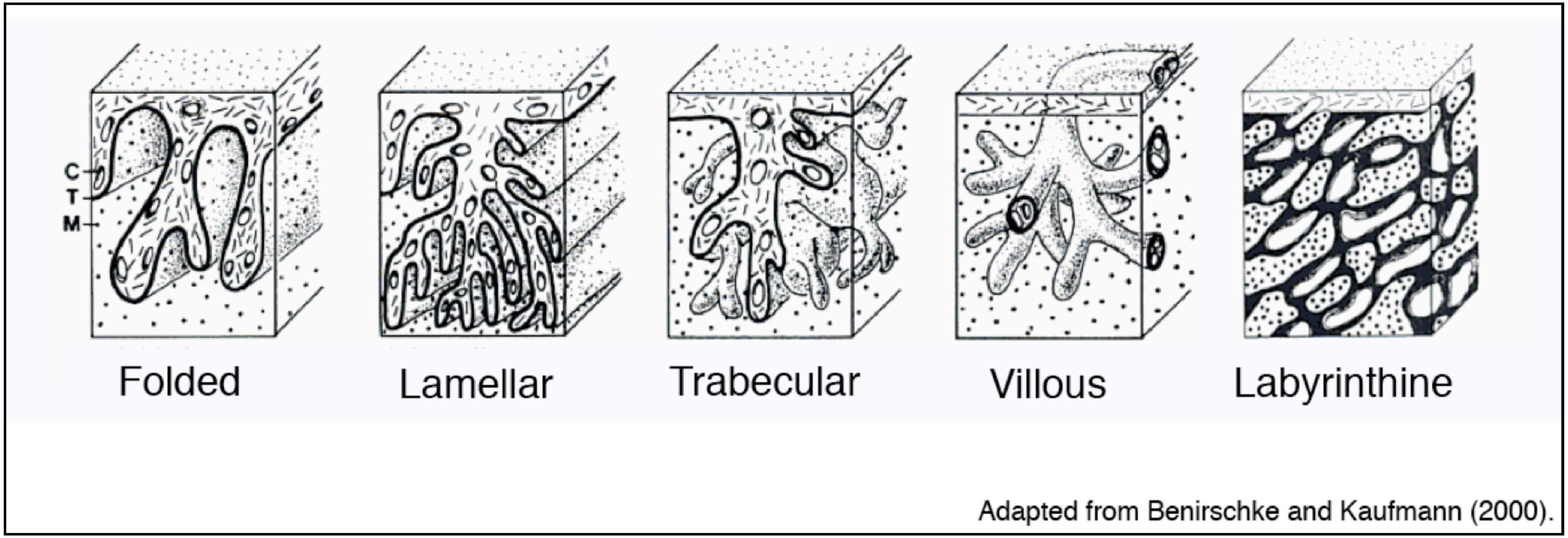
Despite this compelling rationale for its existence, aspects of the evolution of the chorioallantoic placenta remain incompletely understood, not least the form that the ancestral tissue would have taken. Chorioallantoic placentas are often grouped according to three general attributes: (i) the extent of the interhaemal barrier, (ii) the interdigitation between maternal and foetal tissues, and (iii) the shape of the placenta, as scored in Figure 1.12. In this analysis we have pursued a comparative genomics approach to elucidating the processes involved in placenta evolution. The 14 Eutherian species we have included in our analyses are shown in the phylogeny in Figure 1.12, each species being labelled with its placental attributes. The species were strategically selected, based on genome quality and phylogenetic placement.



Adapted from Benirschke and Kaufmann (2000).

Figure 1.13: Variety of placental interhaemal barrier.

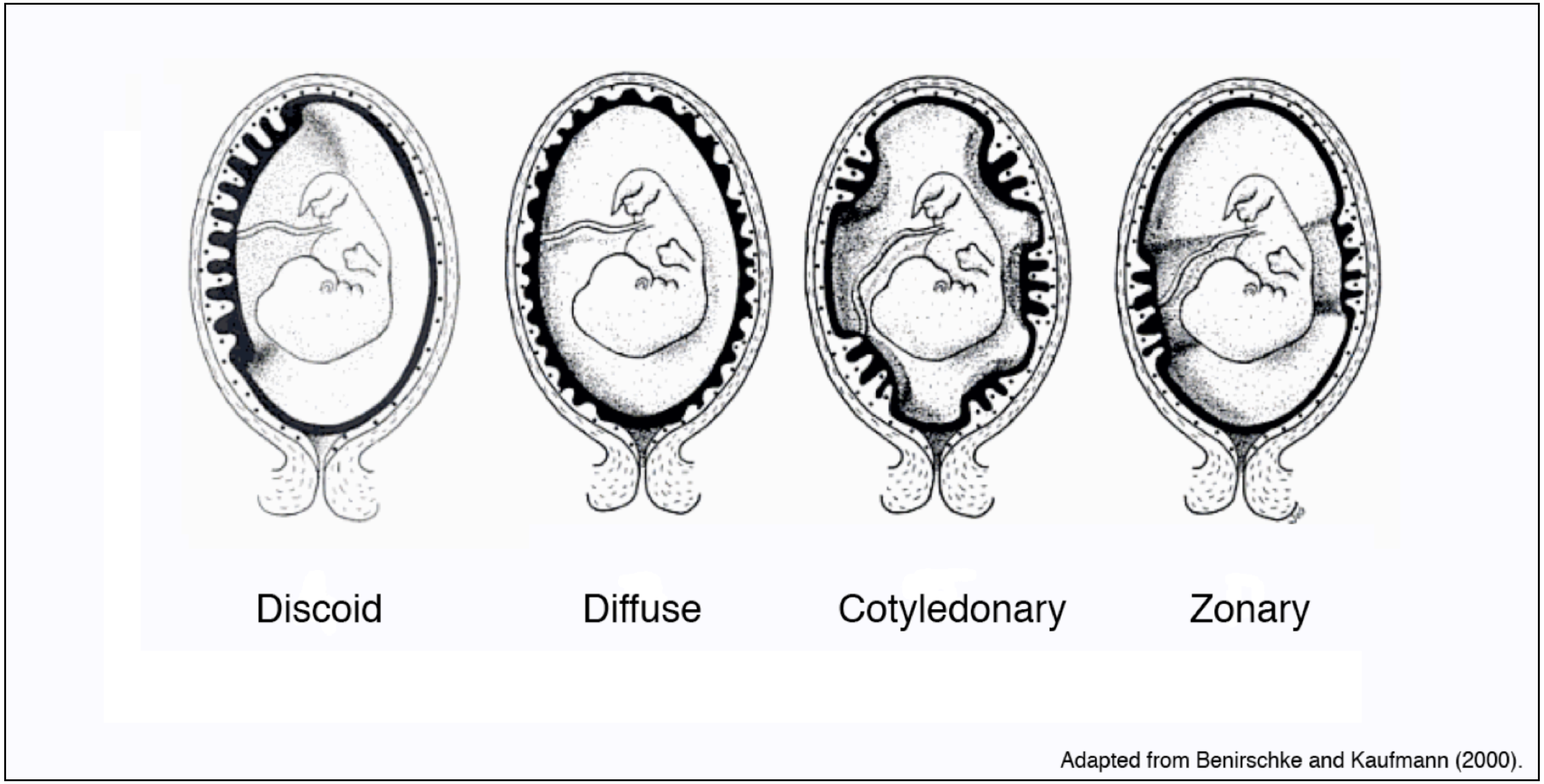
There are three general types of maternofoetal barrier found in the Eutherian chorioallantoic placenta: epitheliochorial, endotheliochorial and haemochorial (left, middle and right respectively in the diagram). Labelled in the epitheliochorial example above are a) maternal blood, b) maternal endothelium, c) endometrial connective tissue, d) endometrial epithelium, e) trophoblast, f) foetal connective tissue, g) foetal endothelium and h) foetal blood. Note that the endometrial epithelium is absent from the endotheliochorial example (middle), bringing trophoblast into contact with the maternal endothelium. Note further that the maternal endothelium is absent from the haemochorial example, bringing the trophoblast into direct contact with maternal blood. For examples from the species under study in this thesis, an epitheliochorial placenta is found in horse, an endotheliochorial placenta is found in dog, and a haemochorial placenta is found in human. Figure adapted with permission from Benirschke and Kaufmann (2000).



Adapted from Benirschke and Kaufmann (2000).

Figure 1.14: Variety of placental interdigitation.

The five main types of interdigitation in the chorioallantoic placenta are shown. From left to right, they are: folded, lamellar, trabecular, villous and labyrinthine. Labelled in the folded example above are: (M) maternal tissue or blood, (T) foetal trophoblast and (C) foetal capillaries. For all examples shown, dotted volumes represent maternal tissue/blood, dashed volumes represent foetal tissue and foetal trophoblast is shown in black. For examples from the species under study in this thesis, the macaque has trabecular interdigitation, the human has villous interdigitation, and that of mouse is labyrinthine. Figure adapted with permission from Benirschke and Kaufmann (2000).



Adapted from Benirschke and Kaufmann (2000).

Figure 1.15: Variety of placental shape.

The chorioallantoic placenta has a variety of shapes in different mammalian clades. From left to right, these include: discoid, diffuse, cotyledonary and zonary. For examples from the species under study in this thesis, the discoid placenta is found in human, the diffuse placenta is found in horse, cotyledonary placenta is found in cow, and zonary is found in elephant. Figure adapted with permission from Benirschke and Kaufmann (2000).

Under the former criterion, a species was included only if it had genome sequencing coverage above 6X. According to the latter requirement, and within the constraints of genome sequence availability, species were sampled as thoroughly as possible from across Eutheria in particular, and from vertebrates more generally. Thorough sampling of Eutheria was essential to ensure that as much as possible of the diversity of placental attributes could be represented. The attributes themselves are described in the following section.

1.10.2 Placental Variety

The interhaemal barrier is the barrier separating the foetal placental tissue from that of the mother, and different types involve different degrees of placental invasiveness. In order of increasing invasiveness, the main types are epitheliochorial, endotheliochorial and haemochorial (see Figure 1.13). Epitheliochorial placentation occurs when the trophoblast is in contact with the uterine epithelium, and offers relatively little oxygen diffusing capacity. Endotheliochorial placentation is where the trophoblast is in contact with the endothelial cells of the maternal blood vessels.

Haemochorial placentation is brought about by the trophoblast from the developing embryo that invades sufficiently into maternal tissue to provide direct access to the maternal blood — other things being equal, this offers the greatest oxygen diffusing capacity.

There is currently a lack of consensus as to what type of interhaemal barrier the ancestral Eutherian would have possessed, with some studies favouring endotheliochorial placentation (Vogel 2005, Mess and Carter 2007, Martin 2008), and others supporting a haemochorial placenta (Wildman *et al.* 2006, Elliot and Crespi 2009). However, Elliot and Crespi (2009) do make a somewhat compelling case that the ancestral placental mammal had a haemochorial placenta, based on evidence from maximum likelihood analysis of data on placenta and body mass for 340 species: 334 extant and 6 extinct. They note that the haemochorial form correlates with smaller body size, which would be consistent with the tendency to have a relatively small body size among early Eutherian mammals (Elliot and Crespi 2009). Although recent evidence suggests that these early Eutherian mammals were somewhat larger than previously believed (Romiguier *et al.* 2013), this would not be expected to affect the overall pattern of early Eutheria having had smaller body sizes.

The other two general attributes that are known to vary among placental mammals are maternofoetal interdigitation (the way in which maternal and foetal tissues interconnect, see Figure 1.14) and placental shape (see Figure 1.15).

Types of interdigitation in Eutheria include folded, lamellar, trabecular, villous or labyrinthine; while the shape of Eutherian placentae are similarly varied in different lineages, with diffuse, cotyledonary, zonary and discoid placentas being observed (Elliot and Crespi 2009). There is more consensus about the ancestral state of these attributes, with the earliest chorioallantoic placenta believed to be discoid and labyrinthine (Elliot and Crespi 2009).

This great variability among chorioallantoic placental forms adopted by different species of Eutherian mammals underlines the flexibility it affords the mother in delivering nutrients to, and removing waste from, the developing young. Although the exact mechanisms by which different placental structures are adapted to different constraints have yet to be completely understood, it is a good illustration of the benefit of the tissue itself, and perhaps a clue to why placental mammals have met with such success.

We have sought to study indirectly the process by which the placenta evolved by studying genes known to be critical for the function of placenta; the assumption being that genes that are critical for placental development are more likely to have played a critical role in early placental evolution.

1.11 Aim of Thesis

This project addresses a fundamental question about the origin and evolution of novel tissues in the animalia. We have attempted to address two complementary questions about the placental miRNAs and genes in our study. The first stage of the project has attempted to address the question: what role has been played by positive selection and functional shifts in the evolution of placental genes? The second stage of the project has involved a comparative benchmark study of current miRNA-target prediction methods, in order to identify the miRNA-target method best able to recover known miRNA-target interactions and to distinguish between known miRNA targets and non-targets. This was in preparation for the third stage of the project — and the second question to be addressed: what role has regulatory innnovation, with respect to miRNAs and their targets, played in the evolution of placenta? From a greater understanding of the evolution of these placental miRNAs and genes, it is possible to make reliable inferences about the evolution of the placenta itself and the molecular building blocks that have contributed to its origin.

Chapter 2: Selective Pressure Analysis of Placental Genes

2.1 Introduction

In this chapter the focus is specifically on placental genes, encompassing genes that are known to be critical to placental development and function, as well as genes that are known to have placenta-specific expression. However, it is worthwhile to consider the genomic backdrop against which these genes have been evolving. In a genome-wide study of 16,529 orthologous gene alignments across six mammalian genomes, a total of 544 genes were identified as containing the classical signatures of positive selection, 144 of these genes showed evidence of positive selection in the ancestral Eutherian lineage alone (Kosiol *et al.* 2008).

Although the study by Kosiol *et al.* (2008) was limited to six mammal species, this gives a conservative estimate of the level of positive selection in mammalian genomes: positive selection was detected in about 2% of genes overall, but this rose to about 6% in testes. On examination of the functional categories of genes most likely to have undergone positive selection, there is enrichment for functions such as sensory perception, reproduction and immune defense. This study also found, in a microarray experiment, that positively selected genes are expressed at significantly lower levels than their non-positively selected counterparts (one-sided Mann-Whitney *U* test, $P < 7 \times 10^{-25}$), and in a more tissue-specific manner. This is consistent with the finding, in the same study, of a negative correlation of Dn/Ds or ω with overall expression level in all 11 tissues sampled (Spearman's rank correlation coefficient, $-0.25 \leq \rho \leq -0.43$), and a positive correlation of ω with tissue-specific expression (Spearman's rank correlation coefficient, $\rho = 0.24$) (Kosiol *et al.* 2008).

A positive correlation of ω with tissue-specific expression of course implies that a study of placental genes is likely to find a large number of positively selected genes. Indeed, a prior study of genes with placenta-specific expression confirms this trend (Hou *et al.* 2009). Although the Hou 2009 study was restricted to genes with one-to-one orthologs, the genes show significant evidence of ancient adaptations, with a larger number of positively selected genes on the Eutherian stem lineage than on descendant branches, indicating that functional shift mediated by changes in protein sequence was taking place in placental genes in the ancestral Eutherian. Of 222 genes studied, 94 (~42%) showed evidence of positive selection — 62 of which were under positive selection in the Eutherian stem lineage (Hou *et al.* 2009). As described in Section 1.3.3, gene duplication is a known mechanism by which new sequences emerge in genomes (Ohno 1970). The analysis carried out in this chapter differs from that of Hou *et al.* (2009) in that it does not exclude those gene families with one-to-many ortholog relationships. Such gene families add to the complexity of the analysis but comprise a large proportion of gene families; their inclusion here should consequently offer, for the first time, a more complete analysis of the evolution of placental genes.

In another detailed study of genes preferentially expressed in placenta, Knox and Baker (2008) performed genome-wide expression profiling of the murine placenta. Samples of murine placenta were taken at nine different stages of gestation and a microarray analysis was performed on the resulting tissue samples. The microarray analysis identified a set of genes preferentially expressed in placenta (i.e. magnitude of fold change ≥ 1.5, placental PEM ≥ 4), many of which undergo a transition: being expressed, at higher levels, later in gestation than earlier, or *vice versa*. Estimating the time of origin of these preferentially expressed genes using homology relationships, it was shown that older, pre-existing genes were over-represented in the cohort of genes preferentially expressed early in gestation, while newer, duplicate genes were more likely to be expressed later in pregnancy (Knox and Baker 2008). From these results, Knox and Baker (2008) formed the hypothesis that early placental evolution was facilitated by the co-option of existing genes, while later placental evolution involved gene duplication and divergence. Three of the preferentially expressed placental genes identified by Knox and Baker (2008) were also identified as placental genes in this chapter: adrenomedullin (ADM), cyclin E1 (CCNE1) and placenta-specific 1 (PLAC1); all three were found to be preferentially expressed in developing placenta, as opposed to at term. While the study by Knox and Baker (2008) highlighted the importance of gene duplications in placental evolution, it stopped short of performing an evolutionary analyis of the selective pressures at work on these genes. An important part of understanding the process of new tissue formation, maintenance and regulation is to examine how each of the genes in the process evolved.

In this chapter we have performed evolutionary analyses of selective pressure on 110 placental genes, 73 of which are known to be critical to placental function and 40 of which are preferentially expressed in placenta. Our aim in doing so was to: (a) determine if protein functional shift (positive selection) is evident in these genes, (b) to what extent and in what lineages (with a particular focus on the ancestral Eutherian lineage), and (c) to determine if the sites inferred to be positively selected are related to specific protein functions.

2.2 Materials

Two complementary sets of placental genes were assembled: those known to be critical to the function of the placenta and those known to be specifically or exclusively expressed in placenta. The genomes of 22 species — 14 Eutherian and 6 outgroup species — were selected, from which gene family members were to be obtained (see Table 2.1). The gene family sequence data of each placental gene was then downloaded from the Ensembl genomic database server (Flicek *et al.* 2012) through the BioMart interface (Smedley *et al.* 2009). All scripts used during this process can be accessed from the following location in the electronic appendix: Appendix / home / code / scripts.html.

(Please note that throughout this thesis, unless otherwise specified, Ensembl data has been obtained from Ensembl release 65. Note also that Ensembl release 65 provides access to the human genome assembly version GRCh37.p5 (Church *et al.* 2011). This includes patch release 5 — genome patches are minor changes made to a genome assembly between major assembly releases. Patch release 5 includes 40 fix patches (a **fix patch** contains an alternative genomic assembly sequence that corrects genome assembly errors in the standard genome release), 65 novel patches (a **novel patch** contains a genomic assembly sequence that adds new alternate loci to the standard genome release), and 9 haplotype sequences that are mainly located in the major histocompatibility complex region of chromosome 6 (Flicek *et al.* 2012).

2.2.1 Selection of Species of Interest

An interspecies analysis of selective pressure requires homologs from multiple species by definition. Selective pressure analysis by maximum likelihood should ideally involve 7 or more taxa (Anisimova *et al.* 2001). Indeed, good taxon sampling can affect many aspects of phylogenetic analysis, with increased taxon sampling improving phylogenetic reconstruction (Philippe *et al.* 2011), and conferring resistance to artefacts such as long branch attraction (Yang and Rannala 2012), so it is important to sample as many taxa (i.e. species) as possible.

On the other hand, higher sequencing coverage and its positive impact on alignment quality is critical to avoid detection of spurious signals of positive selection (Schneider *et al.* 2009) — see Section 1.8.3. This constraint limits the set of species genomes that can be selected, since genome data is inevitably heterogeneous across species in the Ensembl genomic database, ranging from less than 2X coverage in, for example, hedgehog (specifically *Erinaceus europaeus*), to the high coverage assemblies of human and mouse (Flicek *et al.* 2012). Of the 54 full genome assemblies available in the standard Ensembl release 65, 22 species — each with a genome sequencing coverage of 6X or greater — were used in this thesis (see Table 2.1). These 22 genomes were chosen so as to strike a balance between sequence coverage and annotation quality on one hand, and taxon sampling on the other.

2.2.2 Placental Gene Set Assembly

To identify a set of placental genes, two complementary approaches were taken: (I) genes critical to the development and/or function of placenta were identified using a keyword search of Online Mendelian Inheritance in Man (OMIM) (Hamosh *et al.* 2005), and (II) genes specifically expressed in placenta were identified using tissue-specificity filters on microarray expression data, as consolidated by Russ and Futschik (2010). These two sets of genes were then merged into one combined set of (III) placental genes. The processes involved are shown in the bioinformatics pipeline in Figure 2.1.

I. Placenta-Critical Genes

To obtain a set of **placenta-critical** genes — genes critical to placental function and/or development — using Online Mendelian Inheritance in Man (OMIM) (Hamosh *et al.* 2005), the full set of OMIM entries was searched for the following keywords that are associated with placental tissue or cell types: 'placenta', 'trophoblast', 'syncytium', 'blastocyst', 'chorioallantoic', 'chorion' and 'allantois'. This was done using the Perl script FilterOMIMbyKeywords.pl, which filters OMIM with respect to user-specified keywords. This script then outputs a table in a **tab-separated variable** (TSV) file — a text format for representing tables in which rows are on different lines and columns are separated by tab characters — showing information for the set of OMIM entries with a keyword match.

Table 2.1: Ensembl genome assemblies used in this thesis.

Species	Scientific Name	Genome Assembly	Genome Coverage	Assembly Type
Bat	*Myotis lucifugus*	myoLuc2	7X	NC
Chicken	*Gallus gallus*	WASHUC2	~6X	C
Chimp	*Pan troglodytes*	CHIMP2.1.4	6X	C
Cow	*Bos taurus*	UMD3.1	7.1X	C
Dog	*Canis familiaris*	CanFam2.0	7.6X	C
Elephant	*Loxodonta africana*	loxAfr3	7X	NC
Frog	*Xenopus tropicalis*	JGI_4.2	7.65X	NC
Fugu	*Takifugu rubripes*	FUGU4.0	8.5X	NC
Gorilla	*Gorilla gorilla*	gorGor3.1	35X	C
Guinea Pig	*Cavia porcellus*	cavPor3	6.79X	NC
Horse	*Equus caballus*	EquCab2	6.79X	C
Human	*Homo sapiens*	GRCh37.p5	High	C
Lizard	*Anolis carolinensis*	AnoCar2.0	7.1X	C
Macaque	*Macaca mulatta*	MMUL_1.0	~5.1X	C
Marmoset	*Callithrix jacchus*	C_jacchus3.2.1	6X	C
Mouse	*Mus musculus*	NCBIM37	High	C
Opossum	*Monodelphis domestica*	monDom5	7.33X	C
Orangutan	*Pongo abelii*	PPYG2	6X	C
Platypus	*Ornithorhynchus anatinus*	OANA5	6X	C
Rat	*Rattus norvegicus*	RGSC3.4	7X	C
Zebra Finch	*Taeniopygia guttata*	taeGut3.2.4	~6X	C
Zebrafish	*Danio rerio*	Zv9	~7.5X	C

Table 2.1 Legend:

This table lists the common name, species name, genome assembly, genome coverage and assembly type. Assemblies are grouped into chromosomal (C) and non-chromosomal (NC).

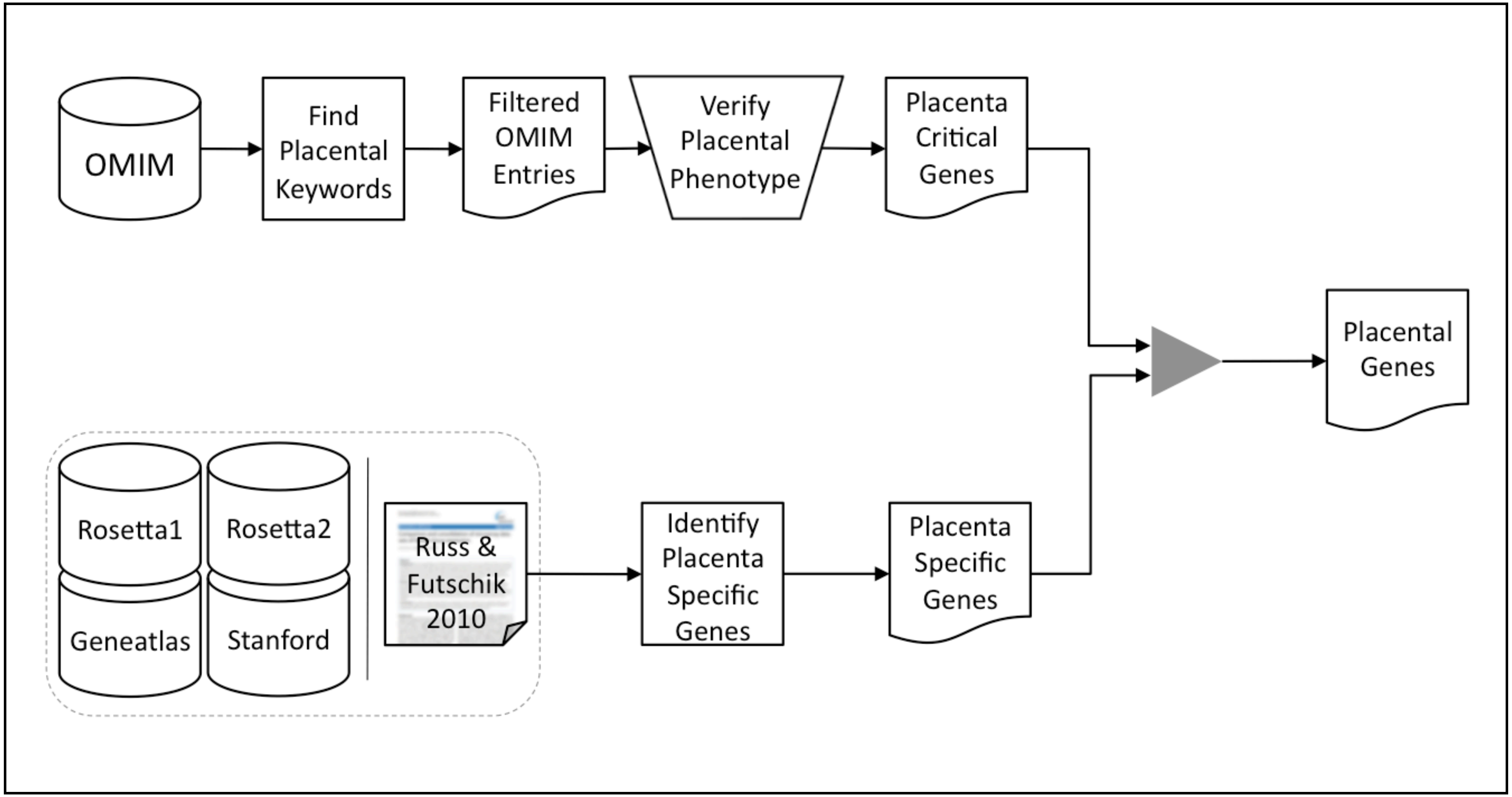


Figure 2.1: Placental gene set assembly pipeline.

Shown is a bioinformatics pipeline depicting the steps involved in placental gene set assembly. Arrows indicate the direction of process flow. For information on symbols used, see the pipeline key in Figure 2.2.

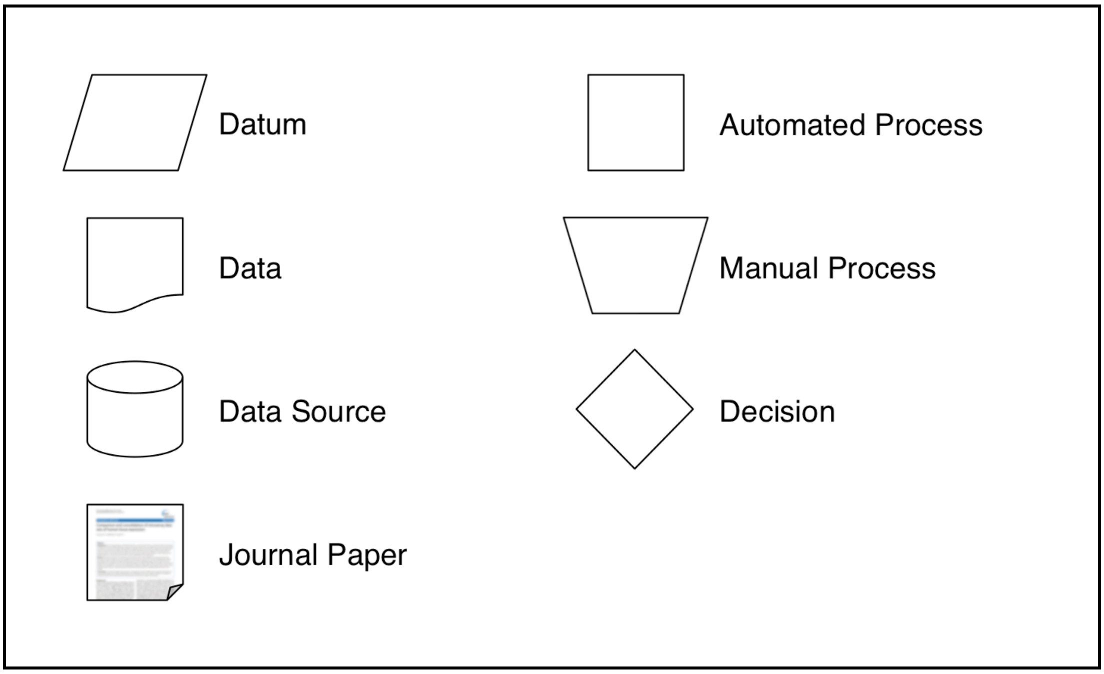


Figure 2.2: Bioinformatics pipeline key.

Shown is a key for the symbols used in bioinformatics pipelines throughout this study. Note that, since it is rare for a process to be purely automated or manual, an automated process is labelled as such if the process is mainly automated, while manual processes are labelled so if they are predominantly manual.

For each of the 2,368 filtered OMIM entries, the output table shows the OMIM ID, gene ID, number of keyword matches and the context of each keyword match. Only OMIM entries for which the relevant gene locus is known were checked for keyword matches.

The set of OMIM entries with keyword matches was then manually reviewed by this author, and retained only if a placental phenotype was verified. Three broad categories of phenotype were accepted: human allelic variants that are associated with a placental disease state (e.g. pre-eclampsia), genes that are upregulated or downregulated in a placental disease state and genes that have been observed to have a null phenotype in animal experiments.

For each retained entry, a brief description of the phenotype was added to the dataset. Also, gene symbols were checked against Ensembl and, where necessary, replaced by the correct HGNC symbol. A total of 73 placenta-critical genes remained after manual review. A table of these can be found in the appendix at the following location: Appendix / home / projects / placenta / data / placental-genes / pcgs.html.

II. Placenta-Specific Genes

Placenta-specific genes — those genes specifically expressed in placenta — were identified using microarray expression data as consolidated from four separate studies by Russ and Futschik (2010). Following the nomenclature of Russ and Futschik (2010), the four microarray datasets were Rosetta1 (Johnson *et al.* 2003), Rosetta2 (Schadt *et al.* 2004), Stanford (Shyamsundar *et al.* 2005) and Geneatlas (Su *et al.* 2002). Four criteria of tissue-specific gene expression were reviewed for possible use in this process, all of which take tissue expression data as input.

Minimum Akaike Information Criterion Estimate (MAICE) identifies a set of tissues with outlying expression values. Given a set of expression values of a given gene across multiple tissues, the set of tissue expression values chosen as outliers (i.e. outlier candidates) that minimises the statistic *U* (see) is taken as the set of tissues in which the gene is specifically expressed (Kadota *et al.* 2003, Ueda 1996, Akaike 1973).

$$U = \frac{1}{2}AIC = n\log(\sigma) + \sqrt{2} \cdot s \cdot \frac{\log(n!)}{n}$$

Equation 2.1: MAICE statistic.

The Minimum Akaike Information Criterion Estimate (MAICE) is obtained by minimising U, shown here in terms of the AIC and also in terms of the number of outlier candidates *s*, the number of tissues that are designated non-outliers *n*, and the standard deviation (σ) of expression values in those tissues that are designated non-outliers.

The outcome of the MAICE process is simply a list of outlier tissues: those for which the gene of interest is relatively over-expressed. With respect to a specific tissue, this can be codified in binary form, such that 1 and 0 indicate, respectively, that the gene *is* and *is not* specifically expressed in the given tissue.

For example, Table 2.2 shows the estimates of MAICE derived from 4 microarray datasets with respect to the expression of the gene pregnancy-associated plasma protein A, pappalysin 1 (PAPPA) in the placenta, corresponding to the expression profile shown in Figure 2.3. PAPPA has highly placenta-specific expression, and this is reflected in the MAICE outcomes for each dataset, which show that expression of PAPPA in placenta is an outlier in the distribution of tissue expression values for that gene.

The MAX statistic developed by Russ and Futschik (2010) was devised to identify genes that are almost exclusively expressed in a unique tissue. Given a vector of *n* tissue expression levels for the gene of interest in order of decreasing expression $\mathbf{T}=(t_1, t_2,..., t_{n-1}, t_n)$, such that t_i is the expression level in the tissue of interest, the MAX statistic is obtained as in Equation 2.2.

$$MAX = \begin{cases} t_1 / t_2 \ \ if \ t_i = t_1 \\ 0 \ \ if \ t_i \neq t_1 \end{cases}$$

Equation 2.2: MAX tissue-specificity measure.

The MAX tissue-specificity measure is shown for the tissue of interest with expression level t_i in two cases. If the tissue of interest is that in which the given gene is most highly expressed ($t_i = t_1$), then MAX is the ratio of expression in the tissue in which a gene has highest expression (t_1) to that of the tissue with next highest expression of the gene (t_2). Otherwise, the value of MAX is zero.

MAX effectively indicates the extent to which the expression of a given gene is specific to a tissue of interest; its value can conceivably range from 1 to ∞, where higher values indicate more preferential expression in the tissue of interest. For example, the placenta-specific expression of PAPPA (see Figure 2.3) is reflected in the MAX values shown in Table 2.2. In each dataset examined, expression levels of PAPPA in placenta were at least twice those in any other tissue.

Huminiecki *et al.* (2003) devised the Preferential Expression Measure (PEM), which scores expression of a gene in a given tissue relative to its average expression in all tissues (see Equation 2.3).

$$PEM = \log_2(o / e)$$

Equation 2.3: Preferential Expression Measure.

The Preferential Expression Measure (*PEM*) is shown in terms of the observed expression level of a gene of interest in a given tissue (*o*) and the expected expression level of the gene of interest (*e*).

PEM is perhaps most useful for assessing both under- and over-expressed genes in the tissue of interest: values below 0 indicate that a gene is under-expressed in the given tissue, while values above 0 indicate that the given is over-expressed in the tissue of interest. For example, the placenta-specific expression of PAPPA (see Figure 2.3) is reflected in the PEM values shown in Table 2.2. In each dataset examined, expression of PAPPA in placenta was many times greater than the levels that would be expected for a gene that is not expressed in a tissue-specific manner.

Xiao *et al.* (2010) defined a specificity measure (SPM) to estimate the tissue-specificity of gene expression, which is evaluated as in Equation 2.4.

$$SPM = \frac{\|P\|}{\|T\|}$$

Equation 2.4: SPM specificity measure.

Given a vector of expression levels for a gene in different tissues T=(t_1, t_2,..., t_{n-1}, t_n), such that t_i is the expression level in the tissue of interest, a second vector T′=(0, 0,..., t_i,...,0), in which all tissue expression values except t_i have been set to zero, and ||P||, the scalar projection of T′ on T, the SPM is evaluated as the ratio of ||P|| to the magnitude of the overall expression vector T (i.e. ||T||).

SPM effectively indicates the proportion of total expression of a gene that is observed in the tissue of interest. SPM can range from 0 to 1, such that a value of 0 indicates that no expression was observed for that gene in the tissue of interest, and a value of 1 implies that the gene is expressed exclusively in that tissue. For example, the SPM values in Table 2.2 effectively capture the impression given by Figure 2.3 that the vast majority of detected expression of PAPPA occurs in the placenta.

An ideal criterion in this instance is one that would select only genes that are exclusively (or almost exclusively) expressed in placenta. Because of this strict requirement, MAICE and PEM were not used. The stringency of MAICE can not be easily adjusted, while PEM does not differentiate sufficiently between those genes with over-expression in a subset of tissues and those with highly specific expression in one tissue. Genes with placenta-specific expression were identified using the Perl script IdentifyTSGs.pl. This script identifies genes with tissue-specific expression based on user-selected tissue-specificity criteria (i.e. one or more of MAICE, MAX, PEM or SPM) and using the dataset provided by Russ and Futschik (2010). MAX and SPM were set to strict cutoff values of 2 and 0.9, respectively. To satisfy both these criteria, 90% of the expression of a particular gene would have to take place in the placenta, and its expression level in placenta would have to be 4-fold higher than in any other tissue (MAX is on a $\log_2$ scale).

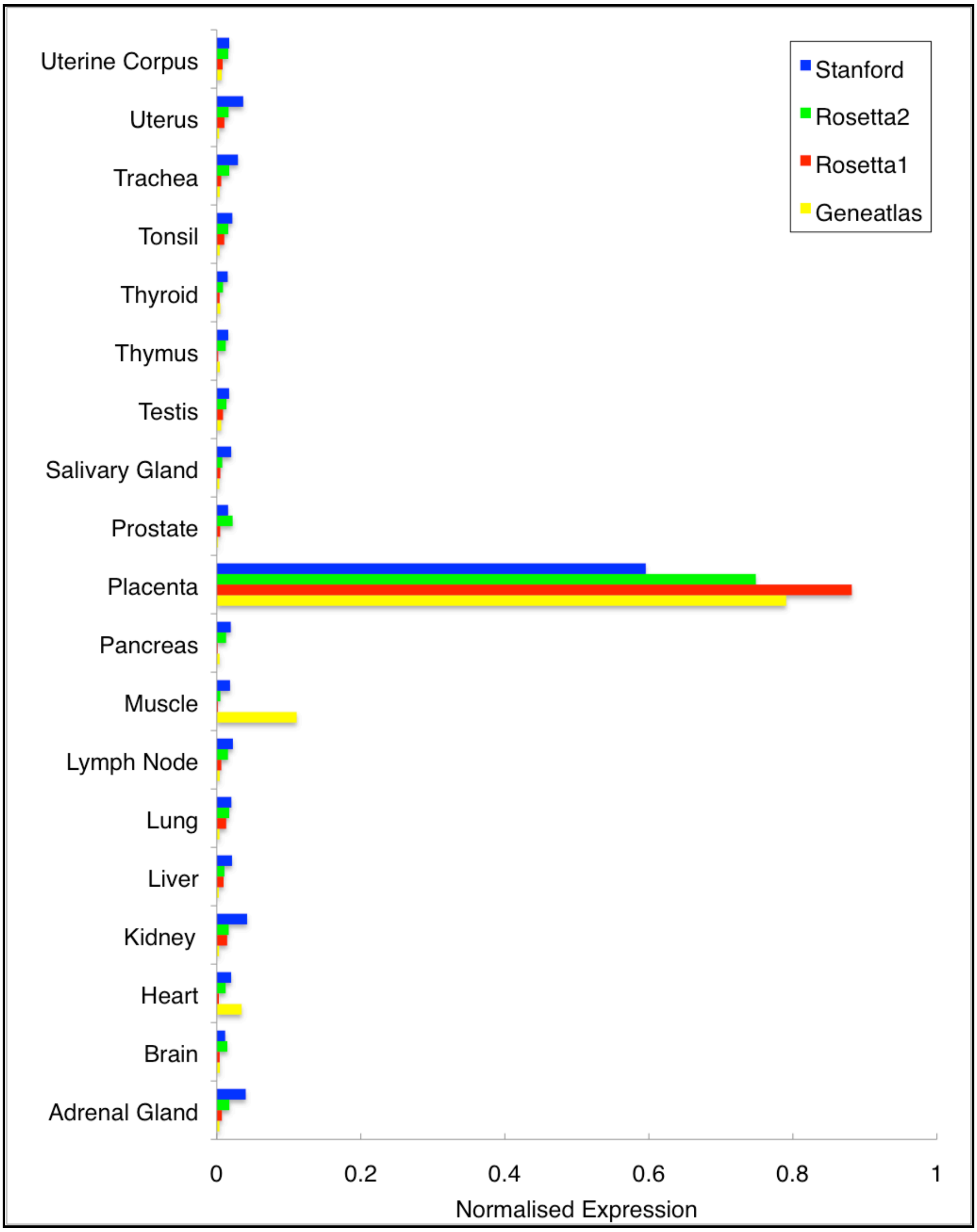


Figure 2.3: PAPPA tissue expression profile.

The above figure is a tissue expression profile for pregnancy-associated plasma protein A, pappalysin 1 (PAPPA), showing normalised expression of PAPPA in 19 tissues across 4 microarray datasets consolidated by Russ and Futschik (2010) — the dataset nomenclature used in that study is used again here. For each tissue, the normalised expression level indicates the proportion of all PAPPA expression that was detected in that tissue. It can be seen from this graph — and from Table 2.2 — that PAPPA has a highly placenta-specific expression pattern.

Table 2.2: PAPPA expression tissue-specificity.

	MAICE	MAX	PEM	SPM
Geneatlas	1	2.83	3.91	0.9784
Rosetta1	1	5.94	4.07	0.9987
Rosetta2	1	5.09	3.83	0.9933
Stanford	1	3.82	3.50	0.9715

Table 2.2 Legend:

Shown are values of 4 tissue-specificity measures with respect to placenta, across 4 microarray expression datasets, for pregnancy-associated plasma protein A, pappalysin 1 (PAPPA). The 4 microarray datasets used were those consolidated by Russ and Futschik (2010) — the dataset nomenclature used in that study is used again here. The 4 tissues specificity measures shown are the Minimum Akaike Information Criterion Estimate (**MAICE**), **MAX**, Preferential Expression Measure (**PEM**) and specificity measure (**SPM**). It can be seen from the specificity measures shown — and from Figure 2.3 — that PAPPA has a highly placenta-specific expression pattern. Note that the value of MAICE takes a value of either 1 or 0, depending respectively on whether or not the expression of PAPPA in placenta constitutes an outlier in the distribution of tissue expression values.

A gene was deemed **placenta-specific** if the tissue-specificity criteria in at least 2 of the 4 component datasets were satisfied. It is important to note that a gene might be accepted as placenta-specific even if it did not satisfy the tissue-specificity criteria in all datasets, as long as the median value of the tissue-specificity measure across datasets exceeded the threshold. Effectively, if the gene was present in 2 or 3 datasets, it would have to meet the threshold in at least 2 datasets. If the gene was present in all 4 datasets, it was required to meet the threshold in at least 3 datasets. The output file was manually reviewed to check all HGNC symbols and insert the correct symbol where necessary.

A total of 44 genes were identified by this process as having placenta-specific expression. Figure 2.3 shows the expression profile of one of these genes — pregnancy-associated plasma protein A, pappalysin 1 (PAPPA) — while Table 2.2 shows the tissue-specificity estimates for PAPPA as derived from each dataset. A table of these can be found in the appendix at the following location: Appendix / home / projects / placenta / data / placental-genes / psgs.html.

III. Placental Genes

The set of placenta-critical genes and the set of placenta-specific genes were combined to form a dataset of "**placental genes**": those that are known to be critical to the development and/or function of placenta, or have placenta-specific expression. The combined placental gene set comprised 114 genes. Three genes were both placenta-critical and placenta-specific: glial cells missing homolog 1 (GCM1), leptin (LEP) and pregnancy specific beta-1-glycoprotein 1 (PSG1). Seven genes were also present in the study by Hou *et al.* (2009) that examined rates of positive selection in one-to-one orthologous gene families on the Eutherian lineage. These were: ADAM metallopeptidase domain 12 (ADAM12), adrenomedullin (ADM), kisspeptin (KISS1), leptin (LEP), met proto-oncogene (MET), nuclear receptor coactivator 6 (NCOA6) and protein C receptor, endothelial (PROCR). These will serve as points of comparison for results generated in this analysis. Gene names and identifiers for all 114 placental genes are shown in Table 2.3.

Table 2.3: Placental genes.

Gene Symbol	Gene Name	Ensembl Gene ID	Representative Transcript Ensembl ID	Swiss-Prot ID
ADAM12	ADAM metallopeptidase domain 12	ENSG00000148848	ENST00000368679	O43184
ADM	adrenomedullin	ENSG00000148926	ENST00000278175	P35318
ADRA2A	adrenoceptor alpha 2A	ENSG00000150594	ENST00000280155	
AGT	angiotensinogen	ENSG00000135744	ENST00000366667	P01019
ALB	albumin	ENSG00000163631	ENST00000295897	P02768
ALPP	alkaline phosphatase, placental	ENSG00000163283	ENST00000392027	P05187
ALPPL2	alkaline phosphatase, placental-like 2	ENSG00000163286	ENST00000295453	P10696
AQPEP	aminopeptidase Q	ENSG00000172901	ENST00000357872	Q6Q4G3
ASCL2	achaete-scute complex homolog 2	ENSG00000183734	ENST00000331289	Q99929
BIRC6	baculoviral IAP repeat containing 6	ENSG00000115760	ENST00000421745	Q9NR09
CAPN6	calpain 6	ENSG00000077274	ENST00000324068	Q9Y6Q1
CCBP2	chemokine binding protein 2	ENSG00000144648	ENST00000273145	O00590
CCNE1	cyclin E1	ENSG00000105173	ENST00000262643	P24864
CCNE2	cyclin E2	ENSG00000175305	ENST00000308108	O96020
CDX2	caudal type homeobox 2	ENSG00000165556	ENST00000381020	Q99626
CGA	glycoprotein hormones, alpha polypeptide	ENSG00000135346	ENST00000369582	P01215
COMT	catechol-O-methyltransferase	ENSG00000093010	ENST00000361682	P21964
CRH	corticotropin releasing hormone	ENSG00000147571	ENST00000276571	P06850
CRHR1	corticotropin releasing hormone receptor 1	ENSG00000120088	ENST00000398285	P34998
CSF3R	colony stimulating factor 3 receptor	ENSG00000119535	ENST00000361632	Q99062

Table 2.3: Placental genes. (continued)

Gene Symbol	Gene Name	Ensembl Gene ID	Representative Transcript Ensembl ID	Swiss-Prot ID
CSH1	chorionic somatomammotropin hormone 1	ENSG00000136488	ENST00000316193	P01243
CSH2	chorionic somatomammotropin hormone 2	ENSG00000213218	ENST00000392886	P01243
CSHL1	chorionic somatomammotropin hormone-like 1	ENSG00000204414	ENST00000309894	Q14406
CUL7	cullin 7	ENSG00000044090	ENST00000265348	Q14999
CYP19A1	cytochrome P450, family 19, subfamily A, polypeptide 1	ENSG00000137869	ENST00000260433	P11511
CYR61	cysteine-rich, angiogenic inducer, 61	ENSG00000142871	ENST00000451137	O00622
DLX3	distal-less homeobox 3	ENSG00000064195	ENST00000434704	O60479
EBI3	Epstein-Barr virus induced 3	ENSG00000105246	ENST00000221847	Q14213
EDN1	endothelin 1	ENSG00000078401	ENST00000379375	P05305
EED	embryonic ectoderm development	ENSG00000074266	ENST00000263360	O75530
EGFL6	EGF-like-domain, multiple 6	ENSG00000198759	ENST00000361306	Q8IUX8
ELF5	E74-like factor 5	ENSG00000135374	ENST00000312319	Q9UKW6
EOMES	eomesodermin	ENSG00000163508	ENST00000295743	O95936
ERF	Ets2 repressor factor	ENSG00000105722	ENST00000222329	P50548
ERN1	endoplasmic reticulum to nucleus signaling 1	ENSG00000178607	ENST00000433197	O75460
ESRRB	estrogen-related receptor beta	ENSG00000119715	ENST00000380887	O95718
ESX1	ESX homeobox 1	ENSG00000123576	ENST00000372588	Q8N693
ETS2	v-ets erythroblastosis virus E26 oncogene homolog 2	ENSG00000157557	ENST00000360214	P15036
FBN2	fibrillin 2	ENSG00000138829	ENST00000262464	P35556
FOSL1	FOS-like antigen 1	ENSG00000175592	ENST00000312562	P15407

Table 2.3: Placental genes. (continued)

Gene Symbol	Gene Name	Ensembl Gene ID	Representative Transcript Ensembl ID	Swiss-Prot ID
GCM1	glial cells missing homolog 1	ENSG00000137270	ENST00000259803	Q9NP62
GH1	growth hormone 1	ENSG00000259384	ENST00000323322	P01241
GH2	growth hormone 2	ENSG00000136487	ENST00000423893	P01242
GINS1	GINS complex subunit 1	ENSG00000101003	ENST00000262460	Q14691
GSTP1	glutathione S-transferase pi 1	ENSG00000084207	ENST00000398606	P09211
HAND1	heart and neural crest derivatives expressed 1	ENSG00000113196	ENST00000231121	O96004
HGF	hepatocyte growth factor	ENSG00000019991	ENST00000222390	P14210
HRAS	v-Ha-ras Harvey rat sarcoma viral oncogene homolog	ENSG00000174775	ENST00000311189	P01112
HSD11B2	hydroxysteroid (11-beta) dehydrogenase 2	ENSG00000176387	ENST00000326152	P80365
IGF2	insulin-like growth factor 2	ENSG00000167244	ENST00000300632	P01344
INSL4	insulin-like 4	ENSG00000120211	ENST00000239316	Q14641
ISM2	isthmin 2 homolog	ENSG00000100593	ENST00000342219	Q6H9L7
ITGAV	integrin, alpha V	ENSG00000138448	ENST00000261023	P06756
KISS1	kisspeptin	ENSG00000170498	ENST00000367194	Q15726
KRT8	keratin 8	ENSG00000170421	ENST00000293308	P05787
LEP	leptin	ENSG00000174697	ENST00000308868	P41159
LGALS13	lectin, galactoside-binding, soluble, 13	ENSG00000105198	ENST00000221797	Q9UHV8
LGALS14	lectin, galactoside-binding, soluble, 14	ENSG00000006659	ENST00000392052	Q8TCE9
MAP2K1	mitogen-activated protein kinase kinase 1	ENSG00000169032	ENST00000307102	Q02750
MAP3K3	mitogen-activated protein kinase kinase kinase 3	ENSG00000198909	ENST00000361733	Q99759

Table 2.3: Placental genes. (continued)

Gene Symbol	Gene Name	Ensembl Gene ID	Representative Transcript Ensembl ID	Swiss-Prot ID
MAPK14	mitogen-activated protein kinase 14	ENSG00000112062	ENST00000229794	Q16539
MET	met proto-oncogene	ENSG00000105976	ENST00000397752	P08581
MFN2	mitofusin 2	ENSG00000116688	ENST00000235329	O95140
MTHFD1	methylenetetrahydrofolate dehydrogenase (NADP+ dependent) 1	ENSG00000100714	ENST00000555709	P11586
NCOA2	nuclear receptor coactivator 2	ENSG00000140396	ENST00000452400	Q15596
NCOA6	nuclear receptor coactivator 6	ENSG00000198646	ENST00000359003	Q14686
NLRP7	NLR family, pyrin domain containing 7	ENSG00000167634	ENST00000340844	Q8WX94
NOS3	nitric oxide synthase 3	ENSG00000164867	ENST00000297494	P29474
NR2F2	nuclear receptor subfamily 2, group F, member 2	ENSG00000185551	ENST00000394166	P24468
PAGE4	P antigen family, member 4	ENSG00000101951	ENST00000218068	O60829
PAPPA	pregnancy-associated plasma protein A, pappalysin 1	ENSG00000182752	ENST00000328252	Q13219
PAPPA2	pappalysin 2	ENSG00000116183	ENST00000367662	Q9BXP8
PBRM1	polybromo 1	ENSG00000163939	ENST00000296302	Q86U86
PCSK4	proprotein convertase subtilisin/kexin type 4	ENSG00000115257	ENST00000300954	Q6UW60
PEG10	paternally expressed 10	ENSG00000242265	ENST00000482108	Q86TG7-2
PHLDA2	pleckstrin homology-like domain, family A, member 2	ENSG00000181649	ENST00000314222	Q53GA4
PISD	phosphatidylserine decarboxylase	ENSG00000241878	ENST00000439502	Q9UG56
PLAC1	placenta-specific 1	ENSG00000170965	ENST00000359237	Q9HBJ0
PNPLA6	patatin-like phospholipase domain containing 6	ENSG00000032444	ENST00000221249	Q8IY17-2
POU5F1	POU class 5 homeobox 1	ENSG00000204531	ENST00000259915	Q01860

Table 2.3: Placental genes. (continued)

Gene Symbol	Gene Name	Ensembl Gene ID	Representative Transcript Ensembl ID	Swiss-Prot ID
PPARG	peroxisome proliferator-activated receptor gamma	ENSG00000132170	ENST00000287820	P37231
PROCR	protein C receptor, endothelial	ENSG00000101000	ENST00000216968	Q9UNN8
PSG1	pregnancy specific beta-1-glycoprotein 1	ENSG00000231924	ENST00000436291	P11464
PSG11	pregnancy specific beta-1-glycoprotein 11	ENSG00000243130	ENST00000320078	Q9UQ72
PSG2	pregnancy specific beta-1-glycoprotein 2	ENSG00000242221	ENST00000329509	P11465
PSG3	pregnancy specific beta-1-glycoprotein 3	ENSG00000221826	ENST00000327495	Q16557
PSG4	pregnancy specific beta-1-glycoprotein 4	ENSG00000243137	ENST00000405312	Q00888
PSG5	pregnancy specific beta-1-glycoprotein 5	ENSG00000204941	ENST00000342951	Q15238
PSG6	pregnancy specific beta-1-glycoprotein 6	ENSG00000170848	ENST00000292125	Q00889
PSG7	pregnancy specific beta-1-glycoprotein 7	ENSG00000221878	ENST00000406070	
PSG9	pregnancy specific beta-1-glycoprotein 9	ENSG00000183668	ENST00000270077	Q00887
ROCK2	Rho-associated, coiled-coil containing protein kinase 2	ENSG00000134318	ENST00000315872	O75116
RTL1	retrotransposon-like 1	ENSG00000254656	ENST00000534062	
S100P	S100 calcium binding protein P	ENSG00000163993	ENST00000296370	P25815
SENP1	SUMO1/sentrin specific peptidase 1	ENSG00000079387	ENST00000004980	Q9P0U3
SERPINE1	serpin peptidase inhibitor, clade E, member 1	ENSG00000106366	ENST00000223095	P05121
SLC2A1	solute carrier family 2 (facilitated glucose transporter), member 1	ENSG00000117394	ENST00000426263	P11166
SLC40A1	solute carrier family 40 (iron-regulated transporter), member 1	ENSG00000138449	ENST00000261024	Q9NP59
SOCS3	suppressor of cytokine signaling 3	ENSG00000184557	ENST00000330871	O14543

Table 2.3: Placental genes. (continued)

Gene Symbol	Gene Name	Ensembl Gene ID	Representative Transcript Ensembl ID	Swiss-Prot ID
STK11	serine/threonine kinase 11	ENSG00000118046	ENST00000326873	Q15831
SULT1E1	sulfotransferase family 1E, estrogen-preferring, member 1	ENSG00000109193	ENST00000226444	P49888
TAC3	tachykinin 3	ENSG00000166863	ENST00000415231	Q9UHF0
TFAP2C	transcription factor AP-2 gamma	ENSG00000087510	ENST00000201031	Q92754
TFPI2	tissue factor pathway inhibitor 2	ENSG00000105825	ENST00000222543	P48307
TGFBR1	transforming growth factor, beta receptor 1	ENSG00000106799	ENST00000374994	P36897
THBD	thrombomodulin	ENSG00000178726	ENST00000377103	P07204
TJP1	tight junction protein 1	ENSG00000104067	ENST00000346128	Q07157
UBP1	upstream binding protein 1 (LBP-1a)	ENSG00000153560	ENST00000283628	Q9NZI7
VEGFA	vascular endothelial growth factor A	ENSG00000112715	ENST00000523873	P15692
VGLL1	vestigial like 1	ENSG00000102243	ENST00000370634	Q99990
VHL	von Hippel-Lindau tumor suppressor, E3 ubiquitin protein ligase	ENSG00000134086	ENST00000256474	P40337
XAGE2	X antigen family, member 2	ENSG00000185751	ENST00000330906	Q96GT9
XAGE3	X antigen family, member 3	ENSG00000171402	ENST00000346279	Q8WTP9
ZFP36L1	ZFP36 ring finger protein-like 1	ENSG00000185650	ENST00000336440	Q07352

Table 2.3 Legend:

Shown on pages 105 to 110 inclusive are tables showing gene names and identifiers for all 114 placental genes studied in this thesis. Aside from each gene symbol and gene name, the corresponding Ensembl Gene ID is given, as is the Ensembl Transcript ID of the representative transcript used, in addition to the corresponding entry in Swiss-Prot, if available.

2.2.3 Placental Gene Data Assembly

Gene family homologs and their representative transcript coding sequences were obtained for each placental gene — the steps taken in doing so are shown in the bioinformatics pipeline in Figure 2.4. For all 22 species of interest, the set of homologs of each placental gene — as inferred by the Ensembl Compara pipeline (Vilella *et al.* 2009) — were obtained from Ensembl (Flicek *et al.* 2012). The process undertaken by Vilella *et al.* (2009) to identify these homologous families is outlined here and shown in Figure 2.5.

1. **Gene sequence download**: a set of representative gene sequences are loaded from all Ensembl genomes. For each gene, the longest available coding sequence is taken as representative.

2. **Homology search**: Protein homology relationships are inferred using WUBLASTP 2.0, in an all-against-all homolog search for the set of representative gene sequences.

3. **Homology network construction**: a sparse graph is constructed of protein homology relationships identified during the homolog search, in which each node represents a protein and each edge represents a homology relationship between two proteins.

4. **Gene family alignment**: Gene families are formed from clusters in the homology graph, and the amino acid sequences of every protein in each gene family are aligned using MUSCLE (Edgar 2004b).

5. **Gene family phylogeny**: TreeBeST (Li 2007) makes use of the aligned homologous genes to infer the phylogeny of the gene family, which is reconciled with a species phylogeny.

6. **Orthology/paralogy inference**: the reconciled gene tree is used to infer orthology and paralogy relationships in each gene family (Vilella *et al.* 2009).

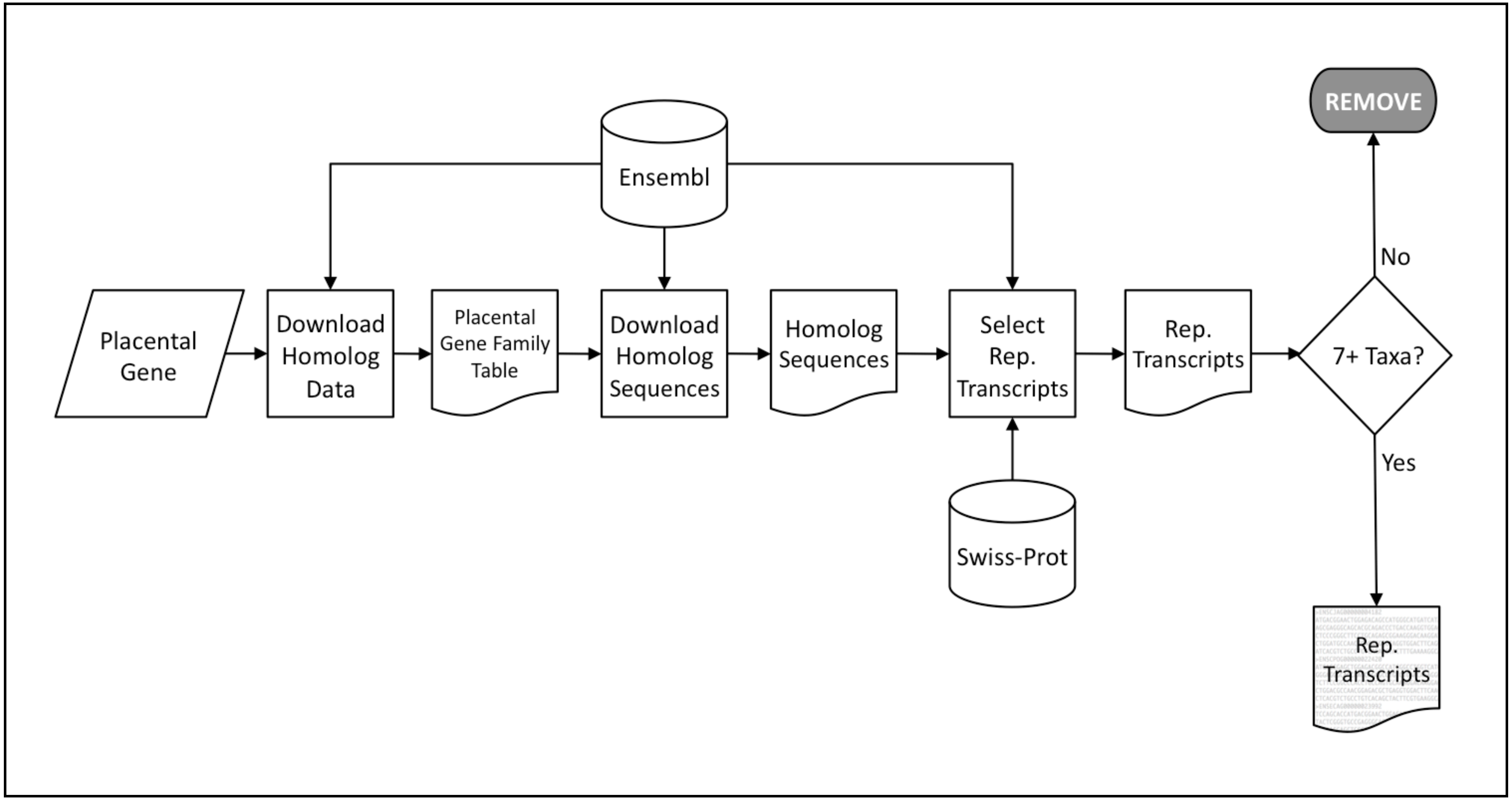


Figure 2.4: Placental gene data assembly pipeline.

Shown is a bioinformatics pipeline depicting the steps involved in placental gene data assembly. Arrows indicate the direction of process flow. For information on symbols used, see the pipeline key in Figure 2.2.

A key aspect of the Ensembl Compara pipeline is the use of a **duplication consistency score** (DCS) developed by Vilella *et al.* (2009) to resolve ambiguous proposed duplications. Given two daughter lineages descending from a proposed duplication event, the DCS is calculated from the number of species represented in both lineages divided by the number of species represented in either lineage. A duplication without subsequent gene loss results in all species being represented in both daughter lineages, this example would attain a perfect DCS of 1. A worst-case scenario with a duplication and reciprocal complementary gene loss in the daughter lineages attains a DCS of 0 — this is considered unlikely (Vilella *et al.* 2009).

For each of the 114 placental human genes described in Section 2.2.2, a set of pre-defined Ensembl Compara homologs was obtained from Ensembl as follows: each gene was used as a query in a BioMart 'Homologs' search on the Ensembl server. The following attributes were requested from the server: (1) the **Ensembl Gene ID** — a unique identifier associated with each gene in the Ensembl genome database; (2) the Human Paralog Ensembl Gene ID — this is the Ensembl Gene ID of a human gene paralogous to one of the placental genes; and (3) the Ancestor — the most recent common ancestor of the placental gene and its paralog (as inferred by the Ensembl Compara pipeline). The result of this query was a paralog table that showed the pre-defined Ensembl Compara human paralogs of each placental gene and the inferred ancestor of each paralogous pair. This table was filtered to retain only paralogs whose common ancestor with a placental gene was within Eutheria. This ensured that only paralogs resulting from duplications within the clade of interest (i.e. Eutheria) would be retained for further analyses.

With the set of placental genes and their paralogs in human, a series of Ensembl BioMart 'Homologs' queries were made to obtain the orthologs of each placental gene and paralog in every other species under study. The attributes sought in each such query were: (1) the Ensembl Gene ID, and (2) either Ortholog or Possible Ortholog. Both of these return the Ensembl Gene ID of an ortholog of the query gene, but possible orthologs are from a part of the gene family phylogeny with a duplication consistency score equal to or less than 0.25 (Vilella *et al.* 2009).

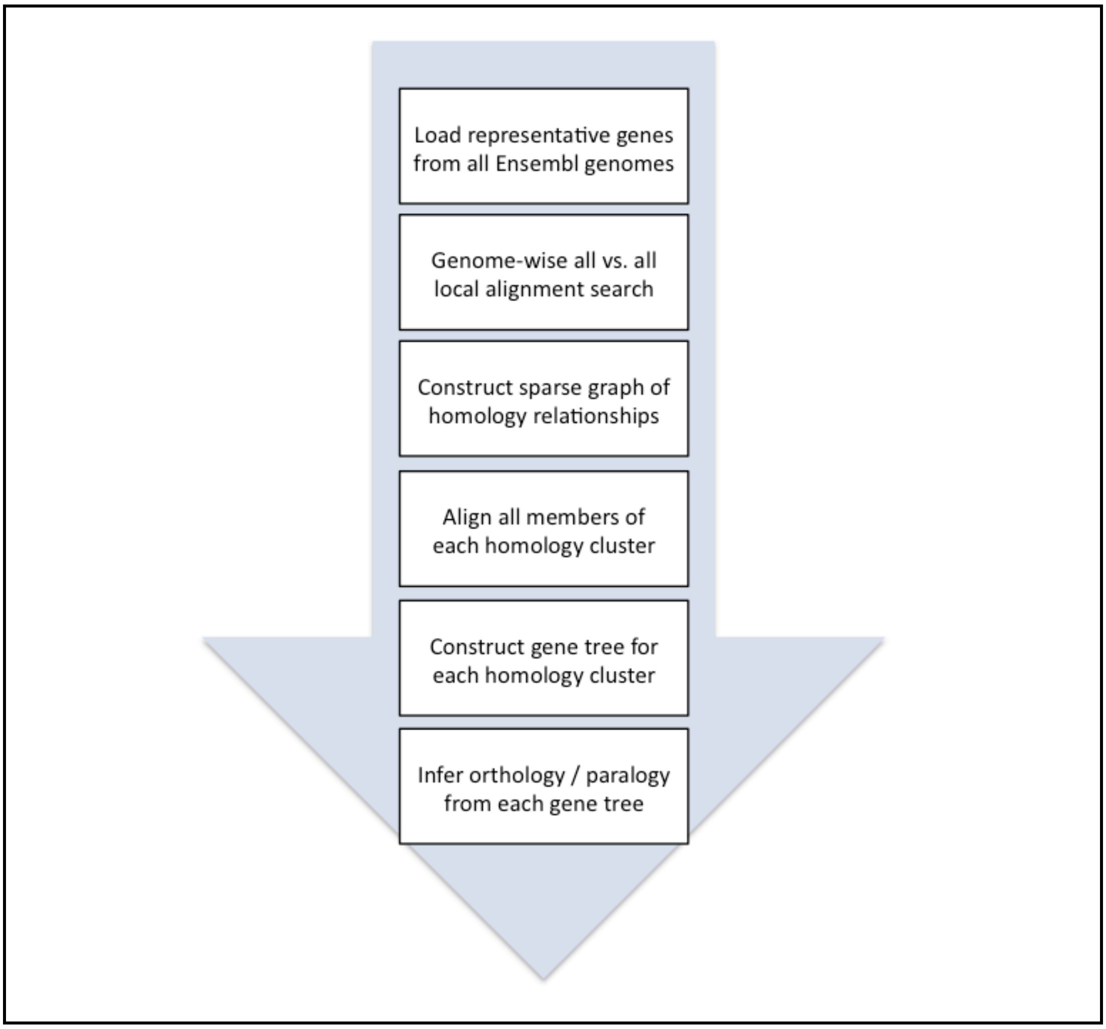

Load representative genes from all Ensembl genomes
Genome-wise all vs. all local alignment search
Construct sparse graph of homology relationships
Align all members of each homology cluster
Construct gene tree for each homology cluster
Infer orthology / paralogy from each gene tree

Figure 2.5: Ensembl Compara pipeline.

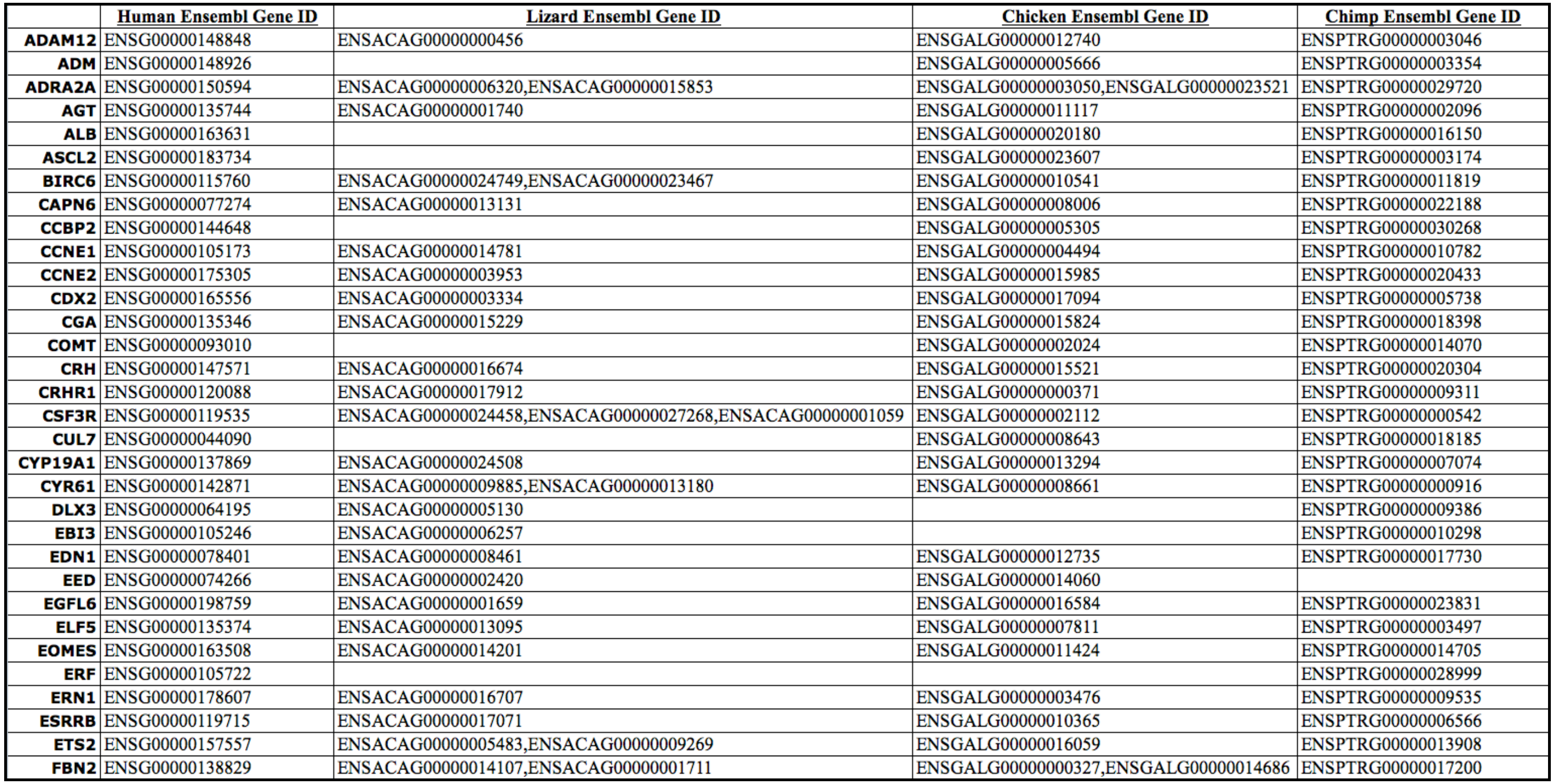
Shown is an outline of the Ensembl Compara pipeline (Vilella *et al.* 2009). This involves 6 main steps: (1) loading a set of representative gene sequences from every Ensembl genome; (2) performing a genome-wise all-against-all search, to identify homology relationships between the set of representative genes in all genomes; (3) constructing a sparse graph of homology relationships between representative genes, such that clusters of genes represent gene families; (4) aligning the sequences of all the genes in each homology cluster; (5) constructing a gene tree for each homology cluster, while reconciling that gene tree with a known species phylogeny; and (6) inferring orthology and paralogy relationships for each pair of homologs using the gene phylogeny.

In summary, the result of these Ensembl BioMart queries was a set of tables, each with two columns: one with the Ensembl Gene ID of the gene of interest, the other with the Ensembl Gene ID of an Ensembl Compara ortholog. Using the Perl script MergeTables.pl, these were combined into one ortholog table, such that every row showed the orthologs of a given placental gene or paralog, and every column showed all orthologs for a given species. With the paralog table and merged ortholog table as input, the script CreateEnsemblGeneFamilyTable.pl was used to construct a gene family table, in which the leftmost column contained the 114 placental genes, and the other columns contained gene family members — including the placental genes themselves — in all 22 species of interest, including human. Figure 2.6 shows an example of such a gene family table.

It is optimal when performing analysis of selective pressure using maximum likelihood that a minimum of 7 species are represented in the dataset. Anisimova *et al.* (2001) found that with a lower number of sequences, selective pressure analysis with maximum likelihood lacks power: in two comparable simulated datasets of length 100 codons, one of which had 6 taxa, the other having 17 sequences, the LRT had a power of 66% and 92%, respectively ($\alpha = 0.05$).

Only 110 gene families were found to have genes in 7 or more species. The other 4 gene families were: pregnancy specific beta-1-glycoprotein 7 (PSG7), X antigen family members 2 and 3 (XAGE2 and XAGE3) and insulin-like 4 (INSL4). These gene families were not analysed for selective pressure variation as the test would lack power. Of the remaining 110 placental gene families, 58 genes had one-to-one orthologs within Eutheria, and 25 genes had one-to-one orthologs throughout all 22 species.

Gene sequences were downloaded for the placental gene families as follows. All annotated coding DNA sequences of each placental gene, paralog and ortholog were obtained using an Ensembl BioMart 'Sequences' query, with the following FASTA header attributes requested: (1) Ensembl Gene ID; and (2) **Ensembl Transcript ID** — a unique identifier associated with each gene transcript in the Ensembl genome database.



	Human Ensembl Gene ID	Lizard Ensembl Gene ID	Chicken Ensembl Gene ID	Chimp Ensembl Gene ID
ADAM12	ENSG00000148848	ENSACAG00000000456	ENSGALG00000012740	ENSPTRG00000003046
ADM	ENSG00000148926		ENSGALG00000005666	ENSPTRG00000003354
ADRA2A	ENSG00000150594	ENSACAG00000006320,ENSACAG00000015853	ENSGALG00000003050,ENSGALG00000023521	ENSPTRG00000029720
AGT	ENSG00000135744	ENSACAG00000001740	ENSGALG00000011117	ENSPTRG00000002096
ALB	ENSG00000163631		ENSGALG00000020180	ENSPTRG00000016150
ASCL2	ENSG00000183734		ENSGALG00000023607	ENSPTRG00000003174
BIRC6	ENSG00000115760	ENSACAG00000024749,ENSACAG00000023467	ENSGALG00000010541	ENSPTRG00000011819
CAPN6	ENSG00000077274	ENSACAG00000013131	ENSGALG00000008006	ENSPTRG00000022188
CCBP2	ENSG00000144648		ENSGALG00000005305	ENSPTRG00000030268
CCNE1	ENSG00000105173	ENSACAG00000014781	ENSGALG00000004494	ENSPTRG00000010782
CCNE2	ENSG00000175305	ENSACAG00000003953	ENSGALG00000015985	ENSPTRG00000020433
CDX2	ENSG00000165556	ENSACAG00000003334	ENSGALG00000017094	ENSPTRG00000005738
CGA	ENSG00000135346	ENSACAG00000015229	ENSGALG00000015824	ENSPTRG00000018398
COMT	ENSG00000093010		ENSGALG00000002024	ENSPTRG00000014070
CRH	ENSG00000147571	ENSACAG00000016674	ENSGALG00000015521	ENSPTRG00000020304
CRHR1	ENSG00000120088	ENSACAG00000017912	ENSGALG00000000371	ENSPTRG00000009311
CSF3R	ENSG00000119535	ENSACAG00000024458,ENSACAG00000027268,ENSACAG0000001059	ENSGALG00000002112	ENSPTRG00000000542
CUL7	ENSG00000044090		ENSGALG00000008643	ENSPTRG00000018185
CYP19A1	ENSG00000137869	ENSACAG00000024508	ENSGALG00000013294	ENSPTRG00000007074
CYR61	ENSG00000142871	ENSACAG00000009885,ENSACAG00000013180	ENSGALG00000008661	ENSPTRG00000000916
DLX3	ENSG00000064195	ENSACAG00000005130		ENSPTRG00000009386
EBI3	ENSG00000105246	ENSACAG00000006257		ENSPTRG00000010298
EDN1	ENSG00000078401	ENSACAG00000008461	ENSGALG00000012735	ENSPTRG00000017730
EED	ENSG00000074266	ENSACAG00000002420	ENSGALG00000014060	
EGFL6	ENSG00000198759	ENSACAG00000001659	ENSGALG00000016584	ENSPTRG00000023831
ELF5	ENSG00000135374	ENSACAG00000013095	ENSGALG00000007811	ENSPTRG00000003497
EOMES	ENSG00000163508	ENSACAG00000014201	ENSGALG00000011424	ENSPTRG00000014705
ERF	ENSG00000105722			ENSPTRG00000028999
ERN1	ENSG00000178607	ENSACAG00000016707	ENSGALG00000003476	ENSPTRG00000009535
ESRRB	ENSG00000119715	ENSACAG00000017071	ENSGALG00000010365	ENSPTRG00000006566
ETS2	ENSG00000157557	ENSACAG00000005483,ENSACAG00000009269	ENSGALG00000016059	ENSPTRG00000013908
FBN2	ENSG00000138829	ENSACAG00000014107,ENSACAG00000001711	ENSGALG00000000327,ENSGALG00000014686	ENSPTRG00000017200

Figure 2.6: Gene family table example.

Shown is an example gene family table, in which each row shows the gene identifiers of members of a particular gene family across all species, while each column shows the identifiers of genes in a particular species across all gene families. In this example, the gene families are named for the gene of interest, while the individual gene identifiers are Ensembl Gene IDs. Where multiple homologs are present for a given gene in a given species, these are separated by commas. Where no homologs are known for a given gene in a given species, the table entry is left blank.

Where multiple alternative transcripts were present for a single gene, a **representative transcript** was chosen for each gene, such that the sequence of that transcript was used for the given gene throughout the analyses in this thesis. Data quality is critical for selective pressure analysis, so the best quality transcript should be selected where possible (Schneider *et al.* 2009).

A single representative transcript was chosen for each gene using the Perl script GetRepresentativeTranscripts.pl in conjunction with a local copy the Swiss-Prot sequence data file and a table of Ensembl transcript annotation data. This script takes as input a FASTA file containing a set of sequences such that for every gene represented in the file, the coding DNA sequence (CDS) of every transcript of that gene available in the Ensembl genomic database (Flicek *et al.* 2012) is included in the file. It then outputs a FASTA file containing a single representative transcript CDS for each gene included in the input FASTA file. The process of representative transcript selection is shown in Figure 2.7 and outlined below.

1. **Quality check**: the set of transcripts for each gene is first quality checked to ensure that the CDS has a complete reading frame and no internal stop codons. Any transcript that fails this quality check (e.g. by having multiple internal stop codons) is removed. If every available transcript of a gene fails this quality check, then it is not possible to select a valid protein coding transcript for that gene, so the gene is removed from consideration entirely. Where only one transcript is available for a gene, it is automatically taken as the representative transcript. Otherwise, this script filters the candidate transcripts by Ensembl annotation.

2. **Ensembl annotation**: transcripts are filtered by their Ensembl annotated 'Transcript Biotype' and 'Status (transcript)', respectively. First, if there are protein-coding transcripts for a gene, all transcripts that are not protein-coding are removed. Then, the transcripts are grouped by Ensembl-annotated evidence status (i.e. 'KNOWN', 'NOVEL' or 'PUTATIVE'), and only the set of transcripts with the best status level are retained. Known transcripts are given preference over novel transcripts, while both are given preference over putative transcripts. This is done on the basis that a transcript with more experimental support is less likely to produce an unreliable alignment and downstream spurious signals of positive selection (Schneider *et al.* 2009). If this filter selects only one candidate transcript, that transcript is taken as representative. Otherwise, this script filters the remaining candidate transcripts by Swiss-Prot annotation.

3. **Swiss-Prot annotation**: if a given gene has a corresponding entry in Swiss-Prot (UniProt Consortium 2011), the transcripts are filtered for sequence similarity to that of the corresponding Swiss-Prot entry. Only those transcripts with the best match to the corresponding Swiss-Prot sequence are kept. If this filter selects only one candidate transcript, that transcript is taken as representative. Otherwise, this script filters the remaining candidate transcripts by transcript length.

4. **Transcript length**: transcripts with the longest CDS are retained. If one candidate transcript has a longer transcript than all others, it is taken as representative. Otherwise, this script arbitrarily selects a transcript as representative from among those with the longest CDS.

5. **Arbitrary selection**: the remaining transcripts are sorted by Ensembl Transcript ID and the first transcript in the list is arbitrarily selected as the representative transcript.

For each species, a gene-transcript mapping table was output, showing the identifiers of each gene and its corresponding representative transcript, in addition to the details of the representative transcript selection process. Where a representative transcript was successfully mapped to a Swiss-Prot entry, the Swiss-Prot accession was included in the information for that representative transcript.

Swiss-Prot annotation informed representative transcript selection in only 17 of the 22 species under study. The following genomes had no corresponding Swiss-Prot entries for Ensembl Compara homologs: Bat (*Myotis lucifugus*), Fugu (*Takifugu rubripes*), Lizard (*Anolis carolinensis*) and Platypus (*Ornithorhynchus anatinus*). Furthermore, Ensembl release 65 lacked Swiss-Prot accession mapping information for Gorilla (*Gorilla gorilla*).

Three genes had no transcript with a valid coding sequence, so these were removed from the dataset. This included the human placental gene pregnancy specific beta-1-glycoprotein 7 (PSG7), which despite the highly placenta-specific expression pattern of its mRNA in 3 of the 4 microarray experiments consolidated by Russ and Futschik (2010), is a polymorphic pseudogene and is designated an orphan (no annotated orthologs) in Ensembl release 65. (In any case, PSG7 was excluded from further analysis because it lacked the requisite number of homologs.) The two other genes found to lack a valid coding sequence were *Danio rerio* genes ENSDARG00000092174 and ENSDARG00000093343. Both these genes were found to have annotated coding sequences whose length is not a multiple of 3, so these were removed from the gene family table created previously.

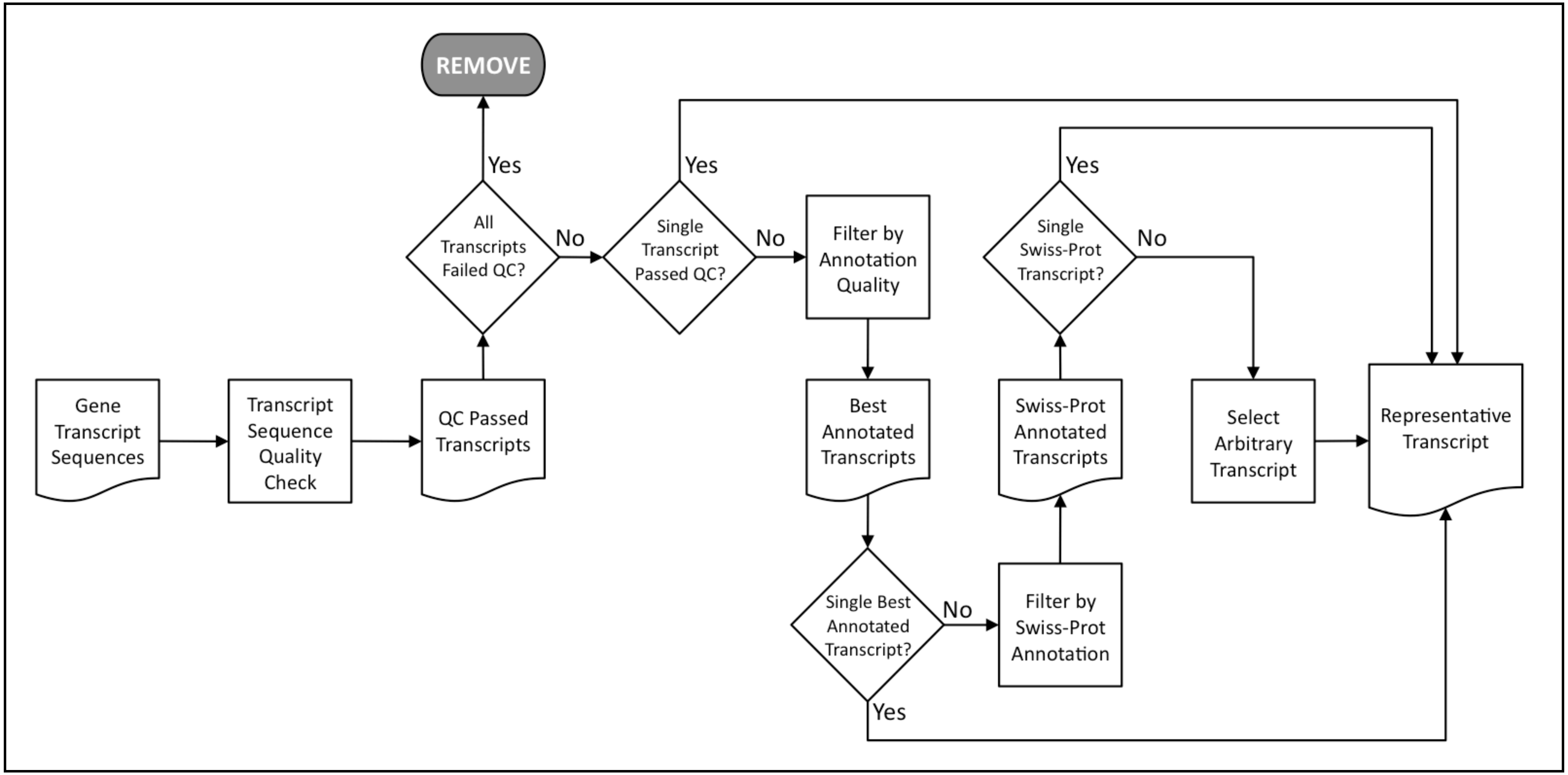

REMOVE
Yes
All Transcripts Failed QC?
No
Yes
Single Transcript Passed QC?
No
Filter by Annotation Quality
Yes
Single Swiss-Prot Transcript?
No
Gene Transcript Sequences
Transcript Sequence Quality Check
QC Passed Transcripts
Best Annotated Transcripts
Swiss-Prot Annotated Transcripts
Select Arbitrary Transcript
Representative Transcript
Single Best Annotated Transcript?
No
Filter by Swiss-Prot Annotation
Yes

Figure 2.7: Representative transcript selection pipeline.

Shown is a bioinformatics pipeline depicting the steps involved in representative transcript selection. Arrows indicate the direction of process flow. For information on symbols used, see the pipeline key in Figure 2.2.

2.3 Methods

This section describes the procedures followed in a selective pressure analysis pipeline conducted for 114 placental genes. The phylogeny of each gene family was reconstructed, as was a nucleotide alignment (see Figure 2.8, part A). These were to be used as input to Codeml for selective pressure analysis (Yang 2007, Yang 1997) (see Figure 2.8, part B). All scripts used during this process can be accessed from the following location in the electronic appendix: Appendix / home / code / scripts.html.

2.3.1 Phylogeny Reconstruction

The reconstruction of the phylogeny of each gene family involved three main steps: (I) the alignment of the sequences of all gene family members, so that homologous amino acid sites could be compared, (II) the selection of the model of best fit for each gene family, and (III) the reconstruction of a phylogenetic tree of the gene family, using the gene family multiple sequence alignment and the model of best fit.

I. Multiple Sequence Alignment

Coding DNA sequences for each gene family were translated to amino acid sequences using the Perl script TranslateFASTA.pl, which takes as input a FASTA file containing protein-coding nucleotide sequences and outputs a FASTA file containing the translated amino acid sequences corresponding to the input nucleotide sequences. All stop codons were removed, and where possible ambiguous codons were resolved and translated (e.g. 'ACN' would be translated as 'T'). The amino acid sequences of each gene family were then aligned using PRANK (Löytynoja and Goldman 2008).

Many different programs are available to perform multiple sequence alignment (MSA) (e.g. Clustal, MUSCLE, PRANK, see Section 1.6.2). PRANK was chosen in this instance because it was found to outperform other methods when used in selective pressure analysis on both simulated data (Fletcher and Yang 2010) and empirical datasets (Markova-Raina and Petrov 2011).

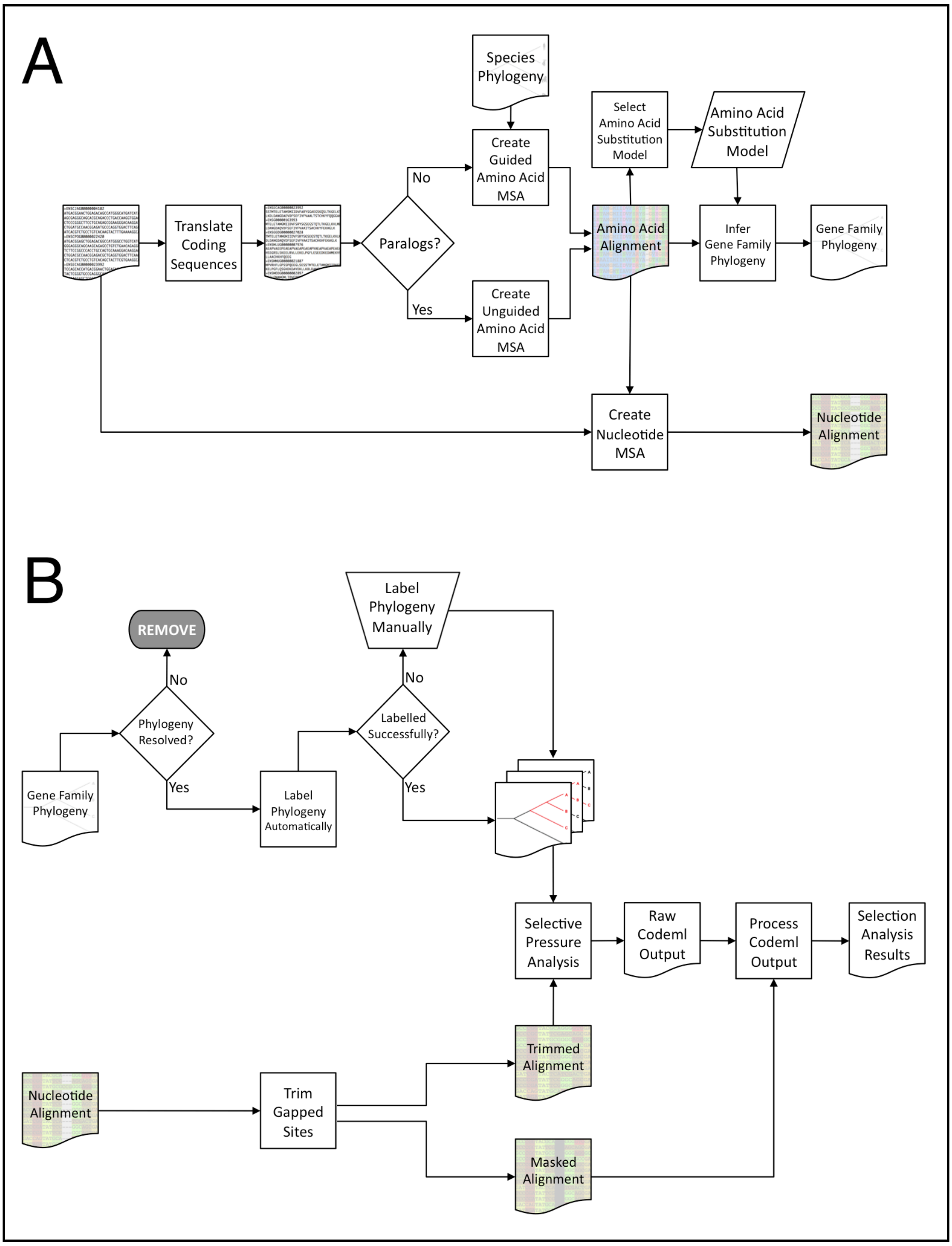


Figure 2.8: Selective pressure analysis pipeline.

Shown is a bioinformatics pipeline depicting the steps involved in selective pressure analysis. Part (A) depicts the process of creating a gene family phylogeny and nucleotide sequence alignment, while part (B) shows the process of preparing and using this data in selective pressure analysis. Arrows indicate the direction of process flow. For information on symbols used, see the pipeline key in Figure 2.2.

For example, in a ROC analysis of selective pressure analysis outcomes in a simulation study by Fletcher and Yang (2010), the true alignment — known from the simulated data — had an area under the ROC curve[6] (AUC) of 0.63, the AUC of PRANK ranged from 0.61-0.63, while AUC did not exceed 0.6 for other methods such as ClustalW (Larkin *et al.* 2007, Thompson *et al.* 1994) and MUSCLE (Edgar 2004b). These other, worse performing methods were found to align non-homologous sites together in the same column, resulting in an alignment that is more compact but less correct with respect to modelling sitewise homology across a set of sequences. However, both studies warn that high false-positive rates remain a potential issue even when using PRANK, with false-positive estimates of selective pressure analysis with a PRANK alignment ranging from 2% to 29%, depending on the foreground lineage, synonymous substitution rate and sequence type (Fletcher and Yang 2010). To reduce this risk, all results from the selective pressure analysis have been carefully monitored (see Section 2.3.2).

There are a number of different versions of the PRANK software, with each iteration incorporating more features and fixing bugs. As mentioned above, 25 of the 110 gene families comprised one-to-one orthologs and lacked paralogs throughout all species sampled. PRANK v.100802 was used to create amino acid MSAs for these 25 gene families, using a guide tree taken from a recent authoritative species phylogeny for these species (Benton and Donoghue 2007).

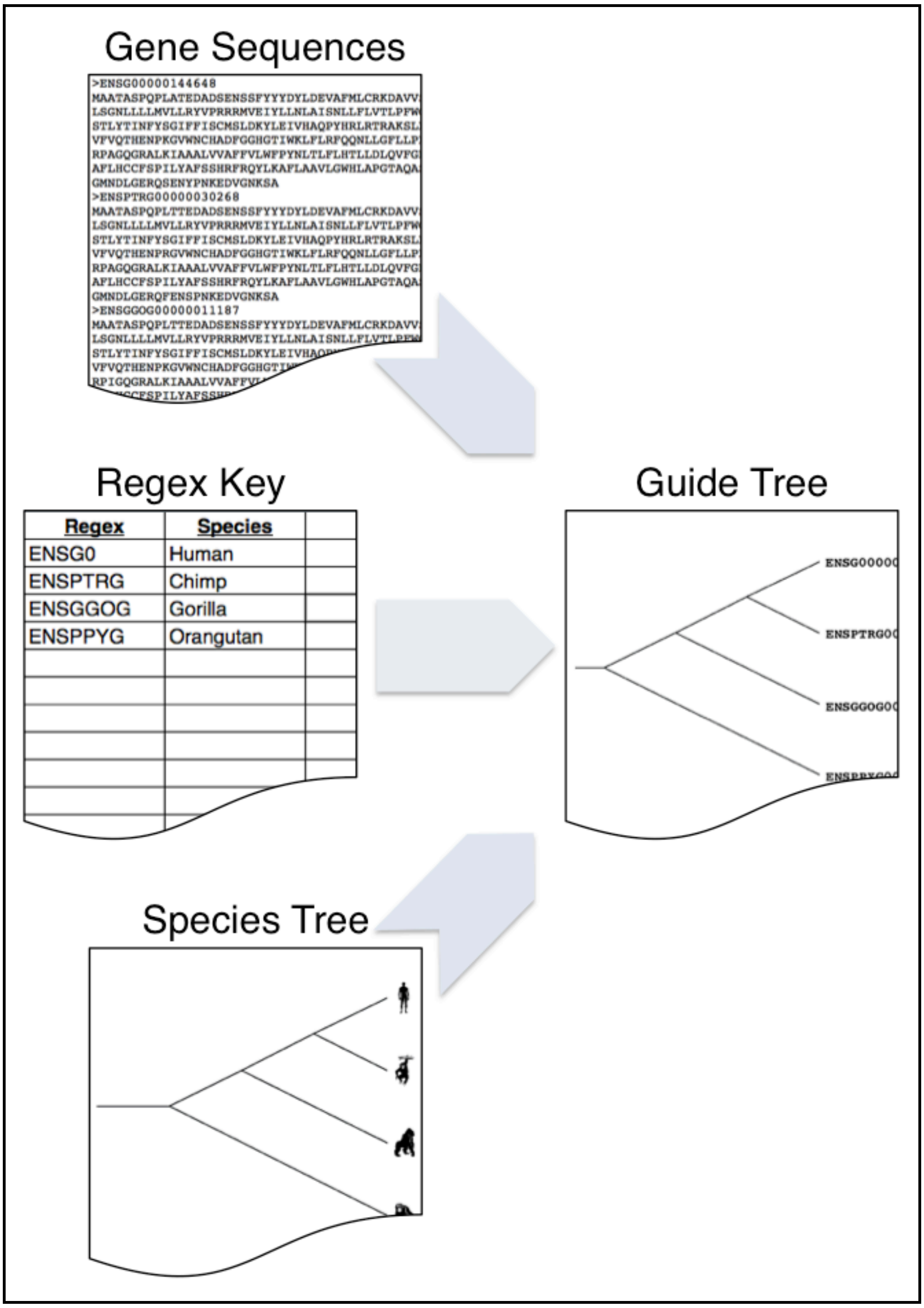
For these 25 gene families comprising one-to-one orthologs, each guide tree was created with the Perl script CreateGuideTree.pl, using the species phylogeny of Benton and Donoghue (2007) as a template (see Figure 2.9). This script takes as input a FASTA file containing the members of a gene family comprising one-to-one orthologs, a template species tree and a regex (i.e. regular expression) key file

[6] ROC curves can be used to compare method performance where the true-positive rate and false-positive rate can be estimated. Essentially, the ROC curve of a method is drawn by plotting a graph of the true-positive rate and false-positive rate of the algorithm at different levels of stringency. The area under the ROC curve (AUC) of each method can then provide a measure of the performance of that method, such that a higher AUC is associated with a higher true-positive rate and lower false-positive rate. A prediction method that identifies all true-positives and rejects all true negatives will have an AUC of 1, while a random method will have an AUC of 0.5.

mapping gene identifier regular expressions to species (e.g. an Ensembl Gene ID containing 'ENSG0' would correspond to 'Human'). For each species with a corresponding gene identifier in the FASTA file, this script replaces the species name with the gene identifier. The resulting guide tree containing gene identifiers is then saved to file and can be used as a pre-specified guide tree for multiple sequence alignment. Version 100802 of PRANK was used in this case because it includes a pruning option: the full species tree can be input as a guide tree to PRANK, which prunes the tree dynamically to retain only those taxa for which an input sequence is present.

For the remaining 85 gene families with paralogs, this Perl script could not be used to create a guide tree, because genes could not be unambiguously placed on the template species tree in those cases. In those cases, PRANK v.081202 was used to create amino acid MSAs with a PRANK-generated guide tree. All PRANK jobs were run on the Ampato high performance computing (HPC) cluster in Dublin City University, operated under the aegis of the Sci-Sym Research Center (http://sci-sym.dcu.ie/).

There is currently no parallel implementation of PRANK, so in order to make optimal use of the parallel architecture of Ampato, a Sun Grid Engine array job script was used to manage the parallel computation of serial PRANK jobs. Array jobs are useful for **embarrassingly parallel** jobs such as this: parallel computation jobs with multiple serial tasks that lack interdependency, such that parallelising the workload is trivial. The array job allocated each individual PRANK job to one of 32 individual computer processors, both at the start of the array job and on completion of each PRANK task. See Figure 2.10 for an illustration.

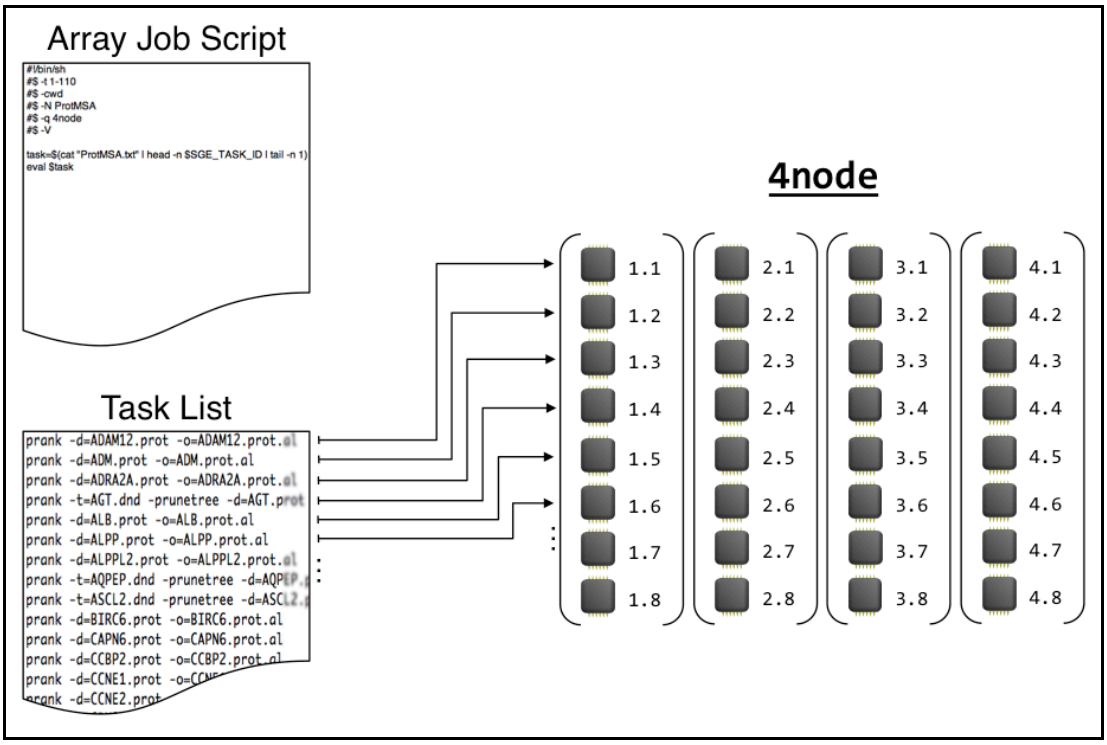


```
>ENSG00000144648
MAATASPQPLATEDADSENSSFYYYDYLDEVAFMLCRKDAVV
LSGNLLLLMVLLRYVPRRRMVEIYLLNLAISNLLFLVTLPFW
STLYTINFYSGIFFISCMSLDKYLEIVHAQPYHRLRTRAKSL
VFVQTHENPKGVWNCHADFGGHGTIWKLFLRFQQNLLGFLLP
RPAGQGRALKIAAALVVAFFVLWFPYNLTLFLHTLLDLQVFG
AFLHCCFSPILYAFSSHRFRQYLKAFLAAVLGWHLAPGTAQA
GMNDLGERQSENYPNKEDVGNKSA
>ENSPTRG00000030268
MAATASPQPLTTEDADSENSSFYYYDYLDEVAFMLCRKDAVV
LSGNLLLLMVLLRYVPRRRMVEIYLLNLAISNLLFLVTLPFW
STLYTINFYSGIFFISCMSLDKYLEIVHAQPYHRLRTRAKSL
VFVQTHENPRGVWNCHADFGGHGTIWKLFLRFQQNLLGFLLP
RPAGQGRALKIAAALVVAFFVLWFPYNLTLFLHTLLDLQVFG
AFLHCCFSPILYAFSSHRFRQYLKAFLAAVLGWHLAPGTAQA
GMNDLGERQFENSPNKEDVGNKSA
>ENSGGOG00000011187
MAATASPQPLTTEDADSENSSFYYYDYLDEVAFMLCRKDAVV
LSGNLLLLMVLLRYVPRRRMVEIYLLNLAISNLLFLVTLPFW
STLYTINFYSGIFFISCMSLDKYLEIVHAQ
VFVQTHENPKGVWNCHADFGGHGTI
RPIGQGRALKIAAALVVAFFV
CCFSPILYAFSS
```

Regex	Species	
ENSG0	Human	
ENSPTRG	Chimp	
ENSGGOG	Gorilla	
ENSPPYG	Orangutan	

Figure 2.9: Guide tree creation.

Shown is an illustration of the input and output of the process of creation of guide trees for use by PRANK alignment software. Guide trees were created only for those gene families with no more than one gene per species, so that each member of the gene family could be matched to its corresponding species. Taking the set of gene sequences and their corresponding identifiers, a species tree derived from that of Benton and Donoghue (2007), and a regex key mapping gene identifiers to species names, a guide tree was created in which each species name had been replaced by the relevant gene identifier. Credit for animal silhouette images is due to Dr Mary J. O'Connell.



```
#!/bin/sh
#$ -t 1-110
#$ -cwd
#$ -N ProtMSA
#$ -q 4node
#$ -V

task=$(cat "ProtMSA.txt" l head -n $SGE_TASK_ID l tail -n 1)
eval $task
```

```
prank -d=ADAM12.prot -o=ADAM12.prot.al
prank -d=ADM.prot -o=ADM.prot.al
prank -d=ADRA2A.prot -o=ADRA2A.prot.al
prank -t=AGT.dnd -prunetree -d=AGT.prot
prank -d=ALB.prot -o=ALB.prot.al
prank -d=ALPP.prot -o=ALPP.prot.al
prank -d=ALPPL2.prot -o=ALPPL2.prot.al
prank -t=AQPEP.dnd -prunetree -d=AQPEP.p
prank -t=ASCL2.dnd -prunetree -d=ASCL2.p
prank -d=BIRC6.prot -o=BIRC6.prot.al
prank -d=CAPN6.prot -o=CAPN6.prot.al
prank -d=CCBP2.prot -o=CCBP2.prot.al
prank -d=CCNE1.prot -o=CCN
prank -d=CCNE2.prot
```

Figure 2.10: Array job example.

Shown is an illustration of an array job performed under the Sun Grid Engine (SGE) batch system on the '4node' queue of the Ampato computing cluster. The array job script (upper left) instructs the batch system to run each serial task from a task list (lower left) on the Ampato 4node queue (at right), which comprises 4 nodes, each with 8 processing cores. A cartoon representation of each core in the 4node queue is shown, numbered by node and core, so that for example, the 3rd core on the 4th node is numbered 4.3. The arrows shown indicate that serial tasks are each allocated to a single core in the parallel architecture, and thus run in an 'embarrassingly parallel' manner.

While the phylogenetic reconstruction for each gene family required an amino acid MSA, the downstream selective pressure analysis would require an alignment at the nucleotide level as the nucleotides contain data about synonymous substitution rates. To generate the precise nucleotide alignment for each gene family, the Perl script MapGapsFASTA.pl was used to map alignment gaps from the amino acid MSA onto the corresponding position in the unaligned nucleotide sequences, such that each single amino acid gap maps to a triplet of gaps in the nucleotide coding sequence. This ensured that the sitewise homology represented in the resulting nucleotide multiple sequence alignment corresponded to that of the gene family amino acid alignment. These nucleotide alignments were retained for use in selective pressure analysis.

II. Phylogenetic Model Optimisation

For each gene family, the evolutionary model of best fit was chosen using the software ProtTest 3.2 (Darriba *et al.* 2011), which proceeds through two main stages:

1. **Model parameter optimisation** using PhyML (Guindon and Gascuel 2003). The following models were included as candidates in this process: JTT, Dayhoff, Blosum62, VT, WAG and LG (see Section 1.6.2). For each model, parameters were optimised using the gene family alignment and a neighbour-joining guide tree created by BIONJ (Gascuel 1997). Parameter optimisation was also performed for variants of each model that account for invariant sites (Reeves 1992), among-site rate variation using a discrete approximation to a gamma distribution (Yang 1996), or a combination of the two. These 3 model variants were denoted by the suffixes '+I', '+G' and '+I+G', respectively.

2. **Model selection** using a suitable criterion. The criterion used in this case was the Bayesian information criterion (BIC) (Schwarz 1978) — see Section 1.6.2. Since none of the models being compared are nested, comparison of a sequence of models by hierarchical LRTs was not a suitable approach. The set of models and model variants were ranked according to increasing BIC, and the posterior probability of each model (conditional on the model of best fit being among those tested) was estimated in a manner analagous to that for 'Akaike weights' (Burnham and Anderson 2002).

In each case, the model with the lowest BIC — and by implication the closest to the 'true' model — was taken to be the model of best fit. If a model variant, such as JTT+G — i.e. Jones, Taylor and Thornton model incorporating a discrete gamma distribution to account for among-site rate variation — was selected, then this model variant was of course chosen for use in the subsequent phylogenetic reconstruction. The results of model selection are shown in Table 2.4.

For 105 of 110 gene families, the model of best fit was found to be JTT incorporating a discrete gamma distribution to account for substitution rate variation among sites (i.e. JTT+G). This was in line with expectations, as a previous study had found JTT to be the model of best fit for a majority (i.e. 57%) of vertebrate genes studied (Keane *et al.* 2006).

Of the remaining 5 gene families, JTT without a gamma distribution was found to be the model of best fit for 3 genes of interest: kisspeptin (KISS1), P antigen family member 4 (PAGE4) and tachykinin 3 (TAC3). The other two genes were albumin (ALB) and caudal type homeobox 2 (CDX2), for which the models of best fit were WAG+G and Dayhoff+G, respectively.

Table 2.4 Legend:

Shown on pages 129 to 131 inclusive are tables listing each gene family for which a phylogenetic model was selected, showing the model, its log-likelihood (**lnL**), the value of the Bayesian Information Criterion (**BIC**) and the posterior probability of the given model **P(*M*)**, given the set of models tested.

Table 2.4: Results of phylogenetic model selection.

Gene	Model	lnL	BIC	P(M)
ADAM12	JTT+G	-13045.63	26372.85	0.97
ADM	JTT+G	-4103.18	8439.42	0.94
ADRA2A	JTT+G	-10707.31	21784.97	0.97
AGT	JTT+G	-12552.62	25371.82	0.96
ALB	WAG+G	-12429.58	25118.98	0.77
ALPP	JTT+G	-21819.53	44345.64	0.98
ALPPL2	JTT+G	-21819.53	44345.64	0.98
AQPEP	JTT+G	-17811.7	35889.66	0.97
ASCL2	JTT+G	-3296.02	6786.95	0.95
BIRC6	JTT+G	-35250.53	70897.43	0.99
CAPN6	JTT+G	-8910.84	18096.17	0.96
CCBP2	JTT+G	-4863.3	9921.74	0.95
CCNE1	JTT+G	-7103.86	14555.78	0.96
CCNE2	JTT+G	-5467.83	11226.95	0.95
CDX2	Dayhoff+G	-2583.62	5389.19	0.93
CGA	JTT+G	-1797.23	3801.09	0.92
COMT	JTT+G	-6155.38	12628.92	0.95
CRH	JTT+G	-3603.15	7441.9	0.94
CRHR1	JTT+G	-3985.18	8232.03	0.96
CSF3R	JTT+G	-22808.21	45939.01	0.97
CSH1	JTT+G	-7546.85	15592.77	0.96
CSH2	JTT+G	-7546.85	15592.77	0.96
CSHL1	JTT+G	-7546.85	15592.77	0.96
CUL7	JTT+G	-26510.77	53311.76	0.98
CYP19A1	JTT+G	-8943.27	18208.75	0.96
CYR61	JTT+G	-7344.22	15012.94	0.96
DLX3	JTT+G	-2443.22	5094.55	0.93
EBI3	JTT+G	-5165.21	10534.79	0.95
EDN1	JTT+G	-3819.05	7850.12	0.94
EED	JTT+G	-2982.04	6219.28	0.96
EGFL6	JTT+G	-10049.77	20399.76	0.96
ELF5	JTT+G	-2712.21	5634.15	0.95
EOMES	JTT+G	-8044.87	16403.91	0.97
ERF	JTT+G	-3402.48	6998.84	0.96
ERN1	JTT+G	-14434.52	29254.85	0.98

Table 2.4: Results of phylogenetic model selection. (continued)

Gene	Model	lnL	BIC	P(M)
ESRRB	JTT+G	-6991.44	14302.47	0.97
ESX1	JTT+G	-12749.19	26034.44	0.97
ETS2	JTT+G	-7925.78	16168.1	0.96
FBN2	JTT+G	-51376.83	103360.12	0.99
FOSL1	JTT+G	-4040.9	8327.4	0.96
GCM1	JTT+G	-7167.25	14582.03	0.96
GH1	JTT+G	-7456.62	15404.78	0.96
GH2	JTT+G	-7456.31	15404.16	0.96
GINS1	JTT+G	-1733.16	3702.02	0.94
GSTP1	JTT+G	-3878.51	8023.18	0.94
HAND1	JTT+G	-1701.08	3586.91	0.88
HGF	JTT+G	-11893.49	24139.74	0.97
HRAS	JTT+G	-1803.2	3898.15	0.95
HSD11B2	JTT+G	-6451.23	13162.02	0.96
IGF2	JTT+G	-3714.76	7649.19	0.95
ISM2	JTT+G	-9581.59	19424.91	0.97
ITGAV	JTT+G	-13194.57	26707.9	0.97
KISS1	JTT	-2057.37	4223.44	0.93
KRT8	JTT+G	-13603.36	27731.63	0.97
LEP	JTT+G	-2062.78	4279.46	0.93
LGALS13	JTT+G	-4448.02	9198.08	0.94
LGALS14	JTT+G	-4448.02	9198.08	0.94
MAP2K1	JTT+G	-2427.93	5113.46	0.96
MAP3K3	JTT+G	-9767.95	19929.32	0.97
MAPK14	JTT+G	-2699.09	5685.28	0.93
MET	JTT+G	-18987.2	38282	0.97
MFN2	JTT+G	-6590.03	13500.36	0.97
MTHFD1	JTT+G	-9378.37	19053.43	0.97
NCOA2	JTT+G	-17842.86	36016.09	0.98
NCOA6	JTT+G	-31619.31	63620.44	0.98
NLRP7	JTT+G	-26532.64	53435.58	0.98
NOS3	JTT+G	-10318.46	20881.99	0.97
NR2F2	JTT+G	-4129.63	8584.32	0.96
PAGE4	JTT	-2876.25	6104.41	0.93
PAPPA	JTT+G	-21160.01	42694.91	0.98

Table 2.4: Results of phylogenetic model selection. (continued)

Gene	Model	lnL	BIC	P(M)
PAPPA2	JTT+G	-30904.49	62148.46	0.98
PBRM1	JTT+G	-18000.31	36472.71	0.98
PCSK4	JTT+G	-12665.66	25577.57	0.97
PEG10	JTT+G	-15190.18	31563.42	0.97
PHLDA2	JTT+G	-2168.64	4511.16	0.94
PISD	JTT+G	-5736.34	11766.7	0.97
PLAC1	JTT+G	-6213.59	12823.11	0.95
PNPLA6	JTT+G	-15321.1	30941.65	0.98
POU5F1	JTT+G	-10564.45	21533.73	0.96
PPARG	JTT+G	-4097.35	8478.31	0.96
PROCR	JTT+G	-4820.75	9831.97	0.94
PSG1	JTT+G	-19096.19	39240.29	0.98
PSG2	JTT+G	-19129.79	39307.49	0.98
ROCK2	JTT+G	-16117.21	32715.58	0.98
RTL1	JTT+G	-13612.97	27362.27	0.98
S100P	LG+G	-1269.37	2661	0.88
SENP1	JTT+G	-9884.51	20066.37	0.97
SERPINE1	JTT+G	-6358.33	12947.62	0.95
SLC2A1	LG+G	-7861.2	16073.05	0.96
SLC40A1	JTT+G	-8992.59	18274.91	0.96
SOCS3	JTT+G	-4156.7	8616.01	0.95
STK11	JTT+G	-3847.11	7955.82	0.96
SULT1E1	JTT+G	-7026.28	14367.76	0.96
TAC3	JTT	-1966.82	4075.17	0.91
TFAP2C	JTT+G	-5788.75	11908.65	0.96
TFPI2	JTT+G	-11229.44	22940.64	0.96
TGFBR1	JTT+G	-4752.34	9837.39	0.97
THBD	JTT+G	-12025.87	24285.03	0.96
TJP1	JTT+G	-24467	49323.82	0.98
UBP1	JTT+G	-5989.95	12279.1	0.96
VEGFA	JTT+G	-5161.85	10621.49	0.96
VGLL1	JTT+G	-6745.56	13753.94	0.96
VHL	JTT+G	-3221.18	6675.74	0.94
ZFP36L1	JTT+G	-7896.97	16109.29	0.97

The 8 pregnancy-specific beta-1-glycoproteins (PSGs) that remained in the dataset at this point form part of a PSG superfamily with multiple gene duplications within the Eutherian clade — according to the Ensembl Compara data. The PSG superfamily members were as follows: PSG1, PSG2, PSG3, PSG4, PSG5, PSG6, PSG9 and PSG11. Model selection for the PSG superfamily was performed only with the MSAs for PSG1 and PSG2 as representatives of the PSG superfamily. Because both alignments contained all 8 members of the PSG superfamily, and JTT+G was the model of best fit in both cases, this was taken as the model of best fit for all PSGs.

For the subsequent Bayesian phylogeny reconstruction (see below), PSG1 and PSG2 were taken as representative of the PSG superfamily, and analysis for the PSGs was performed only on the data for PSG1 and PSG2. Exclusion of the other 6 PSG gene families reduced the total number of phylogenetic trees to be reconstructed to 104. However, each of the remaining 6 PSGs are included in the PSG superfamily phylogeny, so for example, PSG3 is represented in the phylogeny of both PSG1 and PSG2.

III. Bayesian Phylogeny Reconstruction

The parallel implementation of the Bayesian phylogeny software, MrBayes version 3.2, was used to reconstruct the phylogeny of each gene family (Altekar *et al.* 2004, Ronquist and Huelsenbeck 2003, Huelsenbeck and Ronquist 2001). For every gene family, MrBayes was run on the Stokes HPC cluster operated by the Irish Centre for High-End Computing (ICHEC), in 2 runs of 6 chains each for 5 million generations. Trees were sampled every 1000 generations, and the relative burn-in was set to 0.25. For example, at 1 million generations, the burn-in would be 250,000 generations. The convergence diagnostic in MrBayes is the standard deviation of split frequencies (see Section 0). This convergence diagnostic reflects the divergence between sampled trees, so that as the MCMC process converges on an optimal region of tree space, sampled trees become increasingly similar and the differences in splits of the phylogenies are reduced. When the convergence diagnostic reached 0.01, the analysis was stopped and a majority rule consensus tree was generated using the trees from the phase of the run after burn-in.

For 26 gene families for which the phylogenies failed to converge after 5 million generations, MrBayes was rerun for 7 million generations; for the 13 of these that failed to converge, MrBayes was rerun for 20 million generations with chain temperatures adjusted, where the temperature of an MCMC chain determines the probability with which a less optimal phylogeny will be accepted. The adjustments of temperature and MCMC chain length were made in order to improve mixing within tree space, so that the distribution of optimal trees were more effectively sampled. After these reruns of MrBayes, the phylogenies of all but 5 gene families attained a convergence diagnostic of 0.01. For the remaining 5 gene families (i.e. ESX1, FBN2, PEG10, PSG1, PSG2), the convergence diagnostic was not below 0.01. However, for ESX1, FBN2, PEG10 and PSG1, the convergence diagnostic was below 0.05 for at least one run. This was regarded as acceptable and these phylogenies were used in selective pressure analysis, with the caveat that their respective phylogenies were less well resolved and supported than those that achieved a convergence diagnostic below 0.01, and that the results of selective pressure analysis for these gene families should be interpreted with caution, since the tree topology is a key element of selective pressure analysis. The phylogeny of PSG1 was taken as representative for the entire PSG superfamily, reducing the number of gene families to be analysed further to 103.

2.3.2 Selective Pressure Analysis using Maximum Likelihood

Codeml takes as input a gene family phylogeny and a multiple alignment of protein coding DNA sequences, such that the set of genes in the phylogeny matches those in the alignment. This process involves three main steps: (I) removal of phylogenies with poor resolution and *a priori* selection of foreground lineages of interest, (II) removal from nucleotide alignments of those sites with a high proportion of gaps that contribute little to the analysis of selective pressure, and (III) selective pressure analysis and likelihood ratio tests. For an overview, see Figure 2.8, part B.

I. Phylogeny Preparation and Lineage Selection

The gene family phylogeny must satisfy the following requirements in order for a Codeml analysis to produce sensible results: (a) the phylogeny must be reasonably well resolved, at minimum splitting the foreground and background lineages; (b) the phylogeny should be unrooted, if a molecular clock is not assumed; (c) the foreground lineage should be labelled for branch-site models, so that Codeml can distinguish foreground and background taxa. Codeml reads the input phylogeny in Newick format and regards a lineage as foreground if the part of the Newick tree representing that lineage is followed by the text '**#1**'. (See Figure 1.5 for an illustration of the Newick tree format.) For Model 0 and site-specific models, branch labels are ignored, such that identical trees could be used for both site- and lineage-specific models for each specific alignment.

The resolution of each gene family phylogeny was checked using the nw_stat program from the Newick utilities (Junier and Zdobnov 2010). The maximum number of dichotomies (i.e. internal nodes with two children) in a phylogeny with **n** leaves is **n-1**. Phylogenies with less than half this number of dichotomies were considered to be poorly resolved and were not retained for selective pressure analysis; in this case out of 103 phylogenies in total, 17 were deemed to be poorly resolved and were not retained for Codeml analysis. In many cases, the gene family phylogeny was poorly resolved due to high conservation of the gene family member sequences, so in these cases at least, one would not expect to detect positive selection even if they were retained for Codeml analysis.

As is required by Codeml when a molecular clock is not assumed, all phylogenies were unrooted. However, due to the nature of the Newick format, Newick trees have an implied root (see Figure 1.5). To assist automated labelling of phylogenies, for the 28 gene family phylogenies in which ingroup taxa were placed at the implied root, the tree was rerooted around a known outgroup species and then unrooted again. This was done using the known species phylogeny as a guideline (Benton and Donoghue 2007).

The foreground lineages of interest in this analysis can be seen in Figure 1.6. These are the apes, primates, Murinae, rodents, dog, Fereuungulata and Eutheria. For each gene family, the Perl script CreateLabelledTrees.pl was used to automatically generate an appropriately labelled phylogeny for every foreground lineage of interest. This script labels the foreground of a phylogeny in the trivial case where the foreground taxa form one or more distinct paralagous groups, each forming a clear monophyletic clade separated by background taxa. The script takes as input an unlabelled tree template file and one or more files listing tree foreground regexes (i.e. regular expressions). For example, the ape lineage might be labelled with the regexes 'ENSG0', 'ENSPTRG', 'ENSGGOG' and 'ENSPPYG', where these are regular expressions found in the Ensembl Gene IDs of all genes in human, chimp, gorilla and orangutan, respectively. Using the input tree and foreground regexes, this script generates a set of labelled Codeml tree files, one for each foreground clade.

With 7 lineages of interest and 86 gene families, there was a maximum of 602 possible labelled phylogenies. Of these, 430 labelled trees were created by CreateLabelledTrees.pl automatically. Of the remaining 172 possible labelled phylogenies, 15 lacked sequences in the given foreground lineage; 4 lacked background taxa with which to compare the foreground gene sequences; and 153 did not have a clear set of monophyletic clades separated by background taxa. The 153 trees with ambiguous foreground clades were labelled manually. For the 19 trees without either foreground or background taxa, no sensible labelled tree could be produced. In 4 cases, due to nested clades having identical foreground genes, automatic tree labelling produced phylogenies with identically labelled foreground lineages. Where this occurred, the redundant labelled tree was identified and removed manually.

II. Nucleotide Coding Sequence Alignment Preparation

It is not uncommon for regions of a gene family nucleotide MSA to be found in which the majority of codon sites are composed of gaps (i.e. '---') or ambiguous nucleotides (e.g. 'NNN'), typically reflecting missing data or indel events. For example, with a species-specific insertion in a human gene, the human gene sequence will align with no other sequence, so all other alignment sequences will contain gaps at that position in the MSA. Although these may reflect interesting evolutionary events in their own right (e.g. positive selection for indels was observed by Podlaha *et al.* (2005), in the first exon of the rodent sperm protein Catsper1), they are not very informative in a Codeml analysis, since Codeml primarily models substitution rates within homologous sites across species.

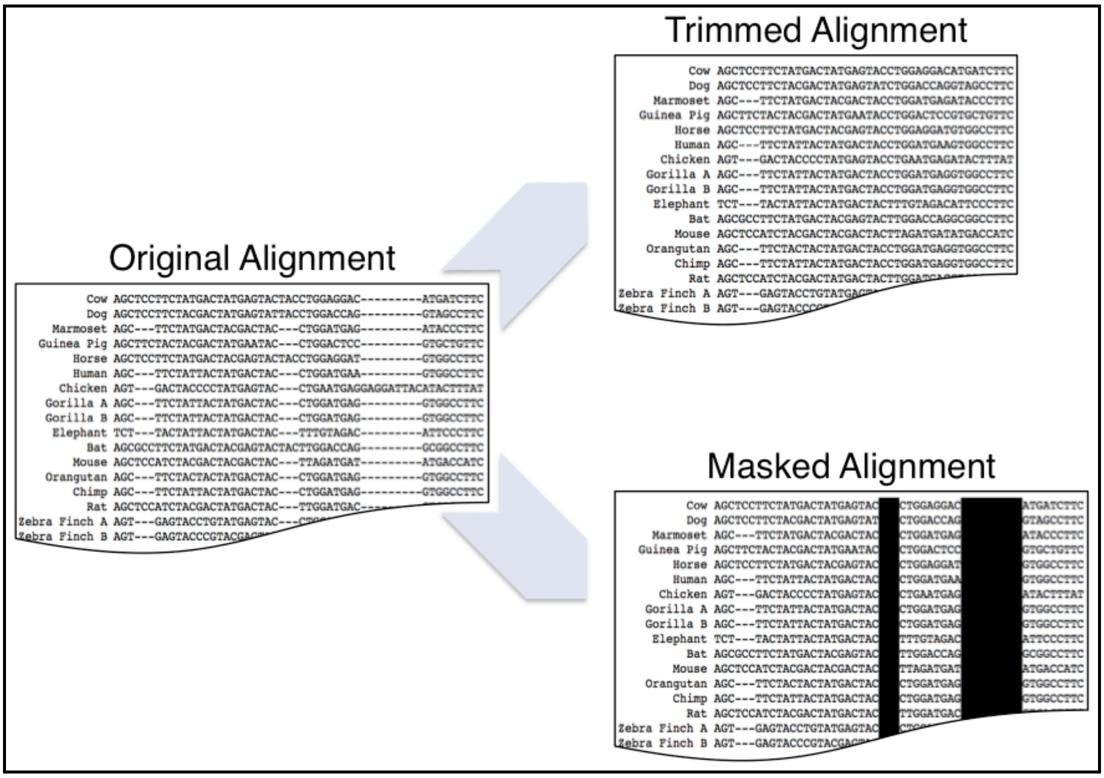
Therefore, for each of the 86 remaining well-resolved gene family phylogenies, the corresponding gene family nucleotide alignment was trimmed using the Perl script CreateTrimmedAlignment.pl. Alignments were trimmed as follows: sites were removed if the number of sequences with gap characters exceeded **n-7**, where **n** was the total number of sequences in the gene family alignment. Sites with such a large number of gaps were trimmed because these would not add any useful information to the analysis. This number was chosen because 7 is considered to be the minimum number of sequences that should be used in a selective pressure analysis using ML methods (Anisimova *et al.* 2001).

In addition to outputting a trimmed version of the gene family alignment, CreateTrimmedAlignment.pl also created a masked alignment in HTML format, such that gap sites were masked by a black background instead of being trimmed. This masked alignment served as a record of both the original alignment and its trimmed counterpart; this is often requested by reviewers of manuscripts. See Figure 2.11 for an example.

III. Selective Pressure Analysis and Likelihood Ratio Tests

A total of eleven models of selective pressure were used in this analysis: the site homogeneous model 0; seven site-heterogeneous models, Model 1a (Neutral), Model 2a (Selection), Model 3 (k=2), Model 3 (k=3), Model 7, Model 8 and Model 8a; and three branch-site heterogeneous models, Model A, Model A Null and Model B. These models were described in detail in Section 1.8.3.

Each individual Codeml run requires its own directory, with its own Codeml control file (stating all parameters and settings to be used, specific to each model) and input gene family phylogeny and nucleotide alignment files. Each Codeml control file specifies settings for a codon model and an initial ω value, as well as file paths to the relevant input files. It was therefore necessary to create a Codeml workspace directory for each gene family using GenerateCodemlWorkspace.pl, which takes as input the gene family alignment file and labelled phylogenies, and creates a Codeml workspace directory for the given gene family containing all the required Codeml control and input files in an ordered directory structure.



```
          Cow AGCTCCTTCTATGACTATGAGTACTACCTGGAGGAC---------ATGATCTTC
          Dog AGCTCCTTCTACGACTATGAGTATTACCTGGACCAG---------GTAGCCTTC
     Marmoset AGC---TTCTATGACTACGACTAC---CTGGATGAG---------ATACCCTTC
   Guinea Pig AGCTTCTACTACGACTATGAATAC---CTGGACTCC---------GTGCTGTTC
        Horse AGCTCCTTCTATGACTACGAGTACTACCTGGAGGAT---------GTGGCCTTC
        Human AGC---TTCTATTACTATGACTAC---CTGGATGAA---------GTGGCCTTC
      Chicken AGT---GACTACCCCTATGAGTAC---CTGAATGAGGAGGATTACATACTTTAT
    Gorilla A AGC---TTCTATTACTATGACTAC---CTGGATGAG---------GTGGCCTTC
    Gorilla B AGC---TTCTATTACTATGACTAC---CTGGATGAG---------GTGGCCTTC
     Elephant TCT---TACTATTACTATGACTAC---TTTGTAGAC---------ATTCCCTTC
          Bat AGCGCCTTCTATGACTACGAGTACTACTTGGACCAG---------GCGGCCTTC
        Mouse AGCTCCATCTACGACTACGACTAC---TTAGATGAT---------ATGACCATC
    Orangutan AGC---TTCTACTACTATGACTAC---CTGGATGAG---------GTGGCCTTC
        Chimp AGC---TTCTATTACTATGACTAC---CTGGATGAG---------GTGGCCTTC
          Rat AGCTCCATCTACGACTATGACTAC---TTGGATGAC---
Zebra Finch A AGT---GAGTACCTGTATGAGTAC---CT
Zebra Finch B AGT---GAGTACCCGTACGAG
```

```
          Cow AGCTCCTTCTATGACTATGAGTACCTGGAGGACATGATCTTC
          Dog AGCTCCTTCTACGACTATGAGTATCTGGACCAGGTAGCCTTC
     Marmoset AGC---TTCTATGACTACGACTACCTGGATGAGATACCCTTC
   Guinea Pig AGCTTCTACTACGACTATGAATACCTGGACTCCGTGCTGTTC
        Horse AGCTCCTTCTATGACTACGAGTACCTGGAGGATGTGGCCTTC
        Human AGC---TTCTATTACTATGACTACCTGGATGAAGTGGCCTTC
      Chicken AGT---GACTACCCCTATGAGTACCTGAATGAGATACTTTAT
    Gorilla A AGC---TTCTATTACTATGACTACCTGGATGAGGTGGCCTTC
    Gorilla B AGC---TTCTATTACTATGACTACCTGGATGAGGTGGCCTTC
     Elephant TCT---TACTATTACTATGACTACTTTGTAGACATTCCCTTC
          Bat AGCGCCTTCTATGACTACGAGTACTTGGACCAGGCGGCCTTC
        Mouse AGCTCCATCTACGACTACGACTACTTAGATGATATGACCATC
    Orangutan AGC---TTCTACTACTATGACTACCTGGATGAGGTGGCCTTC
        Chimp AGC---TTCTATTACTATGACTACCTGGATGAGGTGGCCTTC
          Rat AGCTCCATCTACGACTATGACTACTTGGAT
Zebra Finch A AGT---GAGTACCTGTATGAG
Zebra Finch B AGT---GAGTACC
```

```
          Cow AGCTCCTTCTATGACTATGAGTAC   CTGGAGGAC         ATGATCTTC
          Dog AGCTCCTTCTACGACTATGAGTAT   CTGGACCAG         GTAGCCTTC
     Marmoset AGC---TTCTATGACTACGACTAC   CTGGATGAG         ATACCCTTC
   Guinea Pig AGCTTCTACTACGACTATGAATAC   CTGGACTCC         GTGCTGTTC
        Horse AGCTCCTTCTATGACTACGAGTAC   CTGGAGGAT         GTGGCCTTC
        Human AGC---TTCTATTACTATGACTAC   CTGGATGAA         GTGGCCTTC
      Chicken AGT---GACTACCCCTATGAGTAC   CTGAATGAG         ATACTTTAT
    Gorilla A AGC---TTCTATTACTATGACTAC   CTGGATGAG         GTGGCCTTC
    Gorilla B AGC---TTCTATTACTATGACTAC   CTGGATGAG         GTGGCCTTC
     Elephant TCT---TACTATTACTATGACTAC   TTTGTAGAC         ATTCCCTTC
          Bat AGCGCCTTCTATGACTACGAGTAC   TTGGACCAG         GCGGCCTTC
        Mouse AGCTCCATCTACGACTACGACTAC   TTAGATGAT         ATGACCATC
    Orangutan AGC---TTCTACTACTATGACTAC   CTGGATGAG         GTGGCCTTC
        Chimp AGC---TTCTATTACTATGACTAC   CTGGATGAG         GTGGCCTTC
          Rat AGCTCCATCTACGACTATGACTAC   TTGGATGAC
Zebra Finch A AGT---GAGTACCTGTATGAGTAC   CT
Zebra Finch B AGT---GAGTACCCGTACGAG
```

Figure 2.11: Alignment trimming example.

Shown is an example of the process of trimming gapped sites from a codon alignment. At left is an excerpt of the original alignment, containing 18 sites from 17 sequences. Trimming sites with gaps in more than 10 sequences results in the trimmed alignment excerpt at top right, which is identical to the original alignment except for the removal of unwanted codon sites 9, 13, 14, and 15. A masked alignment can also be produced (bottom right), in which the unwanted codon sites are masked instead of removed. Note that codon site 2 has exactly 10 sequences with gaps and so remains in the output alignments in this example.

Codeml from PAML version 4.4e (Yang 2007) was run for 86 gene families. This was carried out on the ICHEC Stokes HPC cluster, which took an estimated 13,500 core hours to complete 8,408 Codeml runs. As with PRANK, Codeml currently lacks a parallel implementation, so in order to make optimal use of the parallel architecture of the Stokes cluster, Codeml tasks were run using the ICHEC Taskfarm utility. Three taskfarms were run on Stokes, each using 7-8 nodes (with 12 processors per node), in an embarrasingly parallel manner (i.e. serial tasks run separately on parallel architecture). Codeml tasks were sorted in each taskfarm in order of decreasing estimated execution time — based on the model used, as well as the sequence count and sequence length of the alignment — so that as each taskfarm neared completion, smaller tasks were executed on the processors left idle by differences in the execution time of larger tasks.

The set of raw Codeml output files was processed with the Perl script CreateCodemlReports.pl, with 'FOREGROUND' specified as the sequence of interest. This script processed raw Codeml output, performed the necessary likelihood ratio tests (LRTs), and created a set of reports for each gene family. A summary report and LRT table file were output for all gene families. For those gene families in which positive selection was inferred, the following files were output: a positively selected site (PSS) alignment — an alignment highlighting the inferred positively selected sites; and one or more positively selected site (PSS) reports — complete reports on the positively selected sites in each foreground sequence.

A summary of the overall Codeml output was created using the Perl script CreateCodemlAggregateReport.pl. This script processes the raw Codeml output files, performs all necessary LRTs and identifies the model of best fit for each specific dataset. It then outputs an aggregate report showing a summary of results across all the input Codeml workspace directories; this aggregate report allows the overall results of a Codeml selective pressure analysis to be seen at a glance.

For each gene family with positive selection in one or more lineages, the Perl script GetPosSiteOverview.pl was used to create a positively selected site (PSS) overview diagram of each gene in the foreground lineage gene that had a corresponding Swiss-Prot entry. This script takes as input the relevant positively selected site (PSS) report and positively selected site (PSS) alignment, a gene-transcript mapping file and functional annotation data from a local copy of the Swiss-Prot data file (UniProt Consortium 2011). If a corresponding Swiss-Prot entry exists for the gene of interest, this script produces a diagram showing the Swiss-Prot annotation for the cognate protein, overlaid by positively selected sites, as inferred by Codeml (see Figure 2.13 for an example).

Each PSS overview thus produced showed the inferred, positively selected sites overlaid on a representation of the gene/protein showing Swiss-Prot annotation of functionally relevant features. These PSS overviews were reviewed and for those genes in which one or more positively selected sites were in close proximity to a Swiss-Prot annotated feature, possible functional shifts were assessed.

2.4 Results

Following the various filtering and quality control steps outlined in Section 2.3.2, a total of 93 gene families remained for selective pressure analysis — this total includes each of the individual pregnancy specific beta-1-glycoprotein (PSG) gene families, although these were analysed as a superfamily. These 93 gene families met the specific criteria of alignment length, alignment quality, phylogenetic informativeness, taxon number and suitable outgroup data. Positive selection was detected in the Eutherian lineage in 41 of these 93 gene families, or around 44% of cases. Table 2.5 shows the overall results of selective pressure analysis in each gene family and lineage of interest, while Table 2.6 shows functional information for the 41 genes found to be positively selected in the ancestral Eutherian. Figure 2.12 summarises the results of selective pressure analysis with respect to specific lineages. Signals of positive selection were also detected within the Eutheria, with 17 and 24 gene families undergoing positive selection in apes and primates, respectively; 15 and 22 gene families showing signals of positive selection in Murinae and rodents, respectively; and 18 and 28 gene families exhibiting evidence of positive selection in the canid and fereuungulate lineages, respectively. The full set of selective pressure analysis results is accessible from the electronic appendix at the following file path: Appendix / home / projects / placenta / pipelines / selection-analysis / selection-analysis-results.html.

Table 2.5 Legend:

Shown on pages 142 to 144 inclusive are tables of the overall results of the selective pressure analysis of placental genes conducted in this chapter. The first table, with single column **S** (for Sites), shows results for each gene family under site-specific models of selective pressure. The second table, with columns **A**, **P**, **M**, **R**, **D**, **F** and **E**, shows the results for each gene family under a specific lineage: apes, primates, Murinae, rodents, dog, Fereuungulata and Eutheria, respectively. Each row shows the results for a particular gene. A plus sign (+) indicates that positive selection was detected in the lineage in question, while a minus sign (-) indicates that positive selection was not inferred. Note that in the latter case, this is not a positive statement that such positive selection has not occurred, but rather indicates a failure to detect positive selection, such that this failure may or may not be due to the absence of positive selection. Thus, for example, site-specific models failed to infer positive selection in the case of ADAM12, even though branch-site models did infer positive selection in Murinae. For the lineage specific results, an underscore (_) is used for cases in which the given gene family lacked either foreground or background taxa in the given lineage.

Table 2.5: Results of selective pressure analysis.

	S	A	P	M	R	D	F	E
ADAM12	-	-	-	+	-	+	-	-
ADM	-	-	-	-	+	-	-	-
ADRA2A	-	-	+	-	-	+	+	-
AGT	-	+	+	-	+	-	-	-
ALB	-	-	+	+	-	-	-	-
ALPP	+	-	-	-	-	+	+	+
ALPPL2	+	-	-	-	+	+	+	+
AQPEP	+	-	-	-	+	+	-	-
ASCL2	-	-	-	-	-	-	-	-
BIRC6	-	-	-	-	-	-	-	-
CAPN6	-	+	-	-	-	-	-	+
CCBP2	+	-	-	-	-	+	+	+
CCNE1	-	-	-	-	-	-	-	+
CCNE2	-	-	-	-	-	+	-	-
CGA	-	+	+	-	-	-	-	-
COMT	-	-	+	-	+	+	+	-
CRH	-	-	-	-	-	_	+	+
CSF3R	+	-	-	+	-	+	+	+
CSH1	+	+	-	-	+	-	-	-
CSH2	+	+	-	-	+	-	-	-
CSHL1	+	+	-	-	+	-	-	-
CUL7	+	-	-	+	-	-	-	+
CYP19A1	-	-	-	-	+	-	+	-
CYR61	-	-	-	-	-	-	-	-
EBI3	-	-	+	+	-	-	-	-
EDN1	-	-	-	-	-	-	-	+
EGFL6	-	-	-	-	+	-	-	-
EOMES	-	-	-	-	-	-	-	-
ERF	-	-	-	-	+	-	-	+
ERN1	-	-	-	-	-	+	-	-
ESX1	+	-	-	+	+	-	-	_
ETS2	-	-	-	-	-	-	-	-
FBN2	-	-	-	-	-	-	-	+
FOSL1	-	-	-	-	-	-	-	+
GCM1	-	-	-	-	-	-	-	-

Table 2.5: Results of selective pressure analysis. (continued)

	S	A	P	M	R	D	F	E
GH1	+	-	-	-	+	-	-	-
GH2	+	-	-	-	+	-	-	-
GINS1	-	-	-	-	-	-	-	-
GSTP1	-	-	-	-	+	-	-	+
HGF	-	-	-	-	-	-	-	-
HSD11B2	-	-	-	-	-	-	-	+
ISM2	-	+	+	-	-	-	+	+
ITGAV	-	-	-	-	-	-	-	-
KRT8	-	+	-	-	-	+	-	+
LGALS13	+	-	-	-	-	-	+	+
LGALS14	+	-	-	-	-	-	-	+
MAP3K3	-	-	-	-	-	-	-	-
MET	-	-	+	-	-	-	-	-
MFN2	-	+	+	-	-	-	-	-
MTHFD1	-	-	-	-	-	-	-	+
NCOA2	-	-	-	-	-	-	-	+
NCOA6	-	-	-	+	-	-	-	+
NLRP7	+	-	+	+	+	+	+	+
NOS3	-	+	+	-	-	-	-	-
PAGE4	+	-	-	_	_	_	-	_
PAPPA	-	+	-	-	-	-	-	+
PAPPA2	+	-	-	-	-	-	+	+
PBRM1	-	-	-	-	-	+	-	-
PCSK4	-	+	+	+	-	+	-	+
PEG10	-	-	-	+	+	-	+	+
PHLDA2	-	-	-	+	-	_	-	+
PISD	-	-	-	-	-	-	-	-
PLAC1	+	-	-	+	+	-	+	_
PNPLA6	-	-	-	-	-	-	-	+
POU5F1	-	-	-	-	-	-	-	-
PPARG	-	-	-	-	-	-	-	-
PROCR	+	-	-	-	+	-	-	-
PSG1	+	-	+	_	_	_	+	+
PSG11	+	+	+	_	_	_	+	+
PSG2	+	-	+	_	_	_	+	+
PSG3	+	-	+	_	_	_	+	+

Table 2.5: Results of selective pressure analysis. (continued)

	S	A	P	M	R	D	F	E
PSG4	+	-	+	_	_	_	+	+
PSG5	+	+	+	_	_	_	+	+
PSG6	+	-	+	_	_	_	+	+
PSG9	+	-	+	_	_	_	+	+
ROCK2	-	-	-	-	-	-	-	-
RTL1	+	-	+	+	+	-	+	_
S100P	-	-	-	_	+	+	-	-
SENP1	+	-	-	-	-	-	-	-
SERPINE1	-	-	+	-	-	-	-	+
SLC2A1	-	+	-	-	-	+	+	+
SLC40A1	-	-	-	-	-	-	-	-
SOCS3	-	-	-	-	-	+	-	-
STK11	-	-	+	-	-	-	-	-
SULT1E1	-	-	-	-	-	-	+	+
TAC3	+	-	+	-	-	+	-	-
TFAP2C	-	-	-	-	-	-	-	-
TFPI2	+	-	-	+	-	-	+	-
THBD	-	+	-	-	-	-	-	+
TJP1	-	-	-	-	-	-	+	-
VEGFA	-	+	-	-	-	-	-	-
VGLL1	+	-	-	-	+	-	+	-
ZFP36L1	-	-	-	+	-	-	-	+

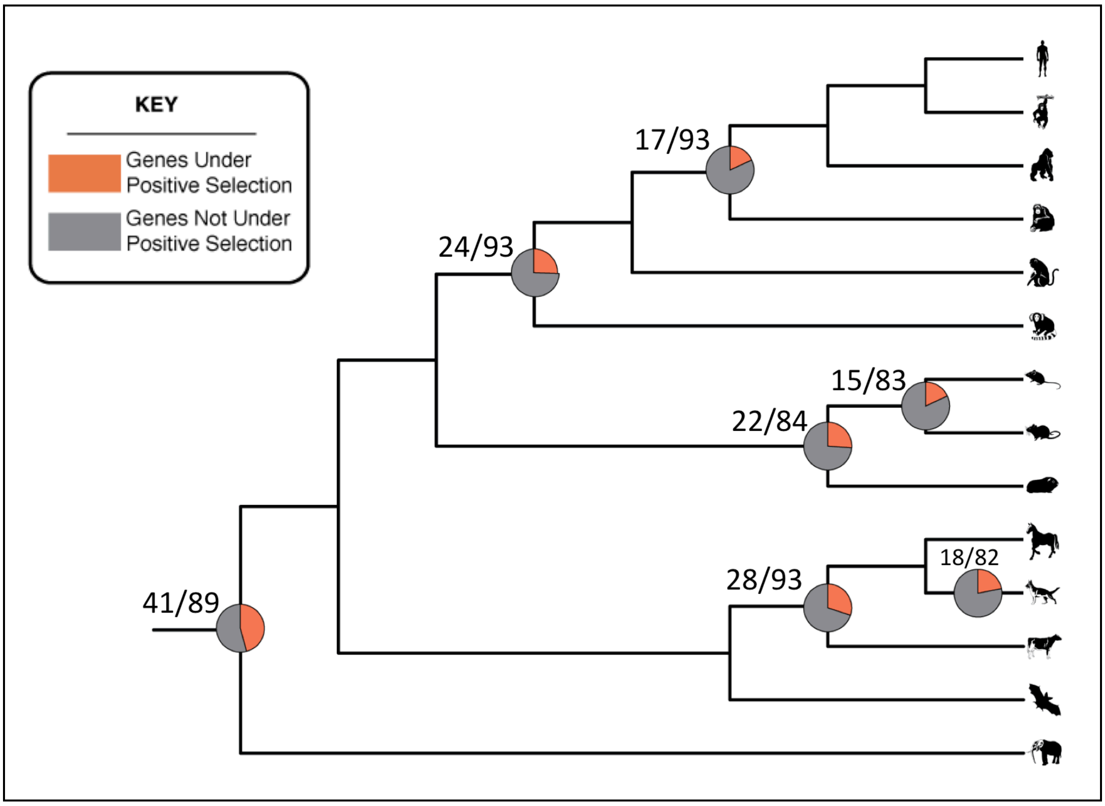

KEY
Genes Under Positive Selection
Genes Not Under Positive Selection
17/93
24/93
15/83
22/84
18/82
28/93
41/89

Figure 2.12: Lineage specific positive selection.

Shown is a phylogeny of Eutheria, overlaid by pie charts on specific lineages showing the proportion of gene families in which positive selection was found in the given lineage. For example, in the Eutherian stem lineage (on the left hand side at the root of the tree), positive selection was identified in 41 of the 89 gene families (i.e. 46.1%) for which selective pressure analysis was performed. For information on species, see Figure 1.6.

Table 2.6: Functions of positively selected genes.

	Cellular Component	Molecular Function
ALPP	plasma membrane, cell surface	phosphatase activity, hydrolase activity, metal ion binding
ALPPL2	plasma membrane, cell surface	phosphatase activity, hydrolase activity, metal ion binding
CAPN6	intracellular, cytoplasm, cytoskeleton, microtubule	endopeptidase activity, protein binding, microtubule binding
CCBP2	plasma membrane	signal transducer activity, receptor activity, protein binding
CCNE1	nucleus, cytoplasm, centrosome	transcription coactivator activity, protein binding, kinase activity, androgen receptor binding
CRH	extracellular region	receptor binding, hormone activity, protein binding
CSF3R	extracellular region, plasma membrane	receptor activity, protein binding
CUL7	nucleus, cytoplasm, mitochondrion	protein binding
EDN1	extracellular region, cytoplasm	cytokine activity, hormone activity, receptor binding, protein binding
ERF	nucleus	transcription factor activity
FBN2	microfibril, extracellular region	extracellular matrix structural constituent, calcium ion binding, protein binding
FOSL1	nucleus, microsome, cytosol, presynaptic membrane	transcription factor activity, protein binding, protein dimerization activity
GSTP1	nucleus, cytoplasm, plasma membrane	protein binding, transferase activity, kinase regulator activity, glutathione binding, nitric oxide binding
HSD11B2	cytoplasm, microsome, endoplasmic reticulum	steroid binding, oxidoreductase activity, NAD binding
ISM2	extracellular region	
KRT8	keratin filament, nucleus, cytoplasm, nuclear matrix	structural molecule activity, protein binding
LGALS13		carboxylesterase activity, sugar binding, lysophospholipase activity, hydrolase activity
LGALS14	nucleus	sugar binding
MTHFD1	mitochondrion	formate-tetrahydrofolate ligase activity, formyltetrahydrofolate dehydrogenase activity, catalytic activity
NCOA2	nucleus, cytoplasm, rough endoplasmic reticulum, Golgi apparatus	transcription coactivator activity, protein binding, nuclear hormone receptor binding, signal transducer activity

Table 2.6: Functions of positively selected genes. (continued)

	Cellular Component	Molecular Function
NCOA6	nucleus, intracellular membrane-bounded organelle	transcription coactivator activity, protein binding, nuclear hormone receptor binding, signal transducer activity
NLRP7		nucleotide binding, ATP binding
PAPPA	extracellular region, cytoplasm, membrane	peptidase activity, metallopeptidase activity, metal ion binding
PAPPA2	extracellular region, cytoplasm, membrane	peptidase activity, metallopeptidase activity, metal ion binding
PCSK4	acrosomal membrane	peptidase activity
PEG10		nucleic acid binding, zinc ion binding
PHLDA2	cytoplasm, membrane	protein binding
PNPLA6	endoplasmic reticulum	lysophospholipase activity, hydrolase activity
PSG1	extracellular region	protein binding
PSG2	extracellular region	protein binding
PSG3	extracellular region	protein binding
PSG4	extracellular region	protein binding
PSG5	extracellular region	protein binding
PSG6	extracellular region	protein binding
PSG9	extracellular region	protein binding
PSG11	extracellular region	protein binding
SERPINE1	extracellular region, plasma membrane, extracellular matrix	protease binding, receptor binding, protein binding, peptidase inhibitor activity
SLC2A1	female pronucleus, cytoplasm, plasma membrane	protein binding, kinase binding, transmembrane transporter activity
SULT1E1	nucleus, cytoplasm, nuclear membrane	steroid binding, transferase activity
THBD	extracellular space, cell surface	transmembrane signaling receptor activity, calcium ion binding, protein binding
ZFP36L1	nucleus, cytoplasm	transcription factor activity, DNA binding, protein binding, zinc ion binding, metal ion binding

Table 2.6 Legend:

Shown here and on page 146 overleaf are tables listing the set of genes that are positively selected in the Eutherian stem lineage, along with functional information for each gene. Gene Ontology (GO) annotation for both cellular component and molecular function were obtained from Ensembl (Flicek *et al.* 2012) through BioMart (Smedley *et al.* 2009).

To assess the functional implications of positive selection, it is necessary to review the positively selected gene in the context of its available functional information. In summary, functional annotation from Swiss-Prot (UniProt Consortium 2011) was integrated with Codeml output on positively selected sites. Of the 75 gene families in which positive selection was inferred on at least one lineage, 61 had a corresponding Swiss-Prot entry for at least one gene family member. Of these 61 gene families, there were 21 gene families in which one or more positively selected sites were in close proximity to a Swiss-Prot annotated functional site or feature. There were 18 such gene families among those positively selected in the ancestral Eutherian lineage. Six of these gene families are discussed further in the following sections, reflecting a spectrum of differing degrees of functional impact in different cases.

2.4.1 Positive Selection of Genes in the Eutherian Lineage

As examples of the 41 gene families under positive selection in the ancestral Eutherian lineage, Table 2.7, Table 2.8 and Table 2.9 each show a summary of the Codeml model parameters and LRT results for the genes chemokine binding protein 2 (CCBP2), sulfotransferase family 1E, estrogen-preferring, member 1 (SULT1E1), and colony stimulating factor 3 receptor (CSF3R), respectively. For CCBP2 and SULT1E1, the results shown are those for which the Eutherian lineage was the foreground, while for CSF3R, results are shown for selective pressure analyses with Murinae and Eutheria respectively labelled as foreground. In all three cases, positive selection was detected in the Eutherian lineage under both branch-site models A and B, with a value of ω much greater than 1. For CSF3R, positive selection was also inferred along the lineage leading to mouse and rat. These results are indicative of strong positive selection, or at the very least a relaxation of selective constraint, on these proteins in the ancestral placental mammal. These three genes and several others are discussed in more detail in the next section.

Table 2.7 Legend:

Shown on page 150 overleaf is a table summarising the results of selective pressure analysis on CCBP2 with the Eutheria as foreground lineage. The **Model** column shows the name of the model used. The **p** column shows the number of free parameters in the ω distribution that are estimated under the given model. The $\boldsymbol{\omega}$ **(t=0)** column shows the initial ω value used in the Codeml run from which results were taken. The **lnL** column shows the log-likelihood of the given model. The **LRT Result** column shows the result of LRTs (if any) for the given model — this shows the null model unless the null was rejected, in which case the alternative model is shown. For example, the **LRT Result** column shows Model A because likelihood ratio tests rejected both Model 1a and Model A Null in favour of their alternative hypothesis (i.e. Model A). Had either of the likelihood ratio tests failed to reject the null hypothesis, this column would instead display the relevant null model(s) (e.g. Model 1a or Model A Null). The **Parameter Estimates** column shows the parameter estimates of each given model for the current dataset. The ω value estimates for each model are as described in Section 1.8.3. The parameters of the form p_i (where i is an integer) indicate the proportion of codon sites in site class i — note that the branch-site model parameters p_2 and p_3 refer to the proportion of codon sites in site classes 2a and 2b, respectively. In Models 7, 8 and 8a, the parameters p and q are the shape parameters of the β distribution used to model sites undergoing neutral evolution and negative selection. The **Positive Selection** column indicates whether positive selection was predicted under the given model. Finally, the rightmost column shows the number of positively selected sites, if any. The **LRT Result** column indicates that both Model 1a and Model A Null were rejected, and that the foreground ω value (ω_2) is much greater than 1, indicating that positive selection has occurred in the Eutherian lineage. Note the very high values for Models A and B of the foreground statistic ω_2. As highlighted by Zhai *et al.* (2012), such high ω estimates may occur at sites with low rates of synonymous substitution (estimates of ω can theoretically range up to infinity, although in practice are typically bounded at a high number, such as 999). However, as these authors point out, such low synonymous substitution rates do not in any case affect the results of the likelihood ratio test (Zhai *et al.* 2012).

Table 2.7: Results of selective pressure analysis on CCBP2 with respect to the Eutherian lineage.

Model	p	ω (t=0)	lnL	LRT Result	Parameter Estimates	Positive Selection	Positive Sites P(ω>1) > 0.5
Model 0	1	2	-8113.608629	N/A	ω=0.23127	No	
Model 1a	2	2	-8006.859619	N/A	p_0=0.74987 p_1=0.25013 ω_0=0.14114 ω_1=1.00000	N/A	
Model 2a	4	2	-8006.859619	Model 1a	p_0=0.74987 p_1=0.17041 p_2=0.07972 ω_0=0.14114 ω_1=1.00000 ω_2=1.00000	No	
Model 3 (k=2)	3	2	-7978.71558	Model 3 (k=2)	p_0=0.49120 p_1=0.50880 ω_0=0.05694 ω_1=0.47667	No	
Model 3 (k=3)	5	2	-7964.470545	Model 3 (k=3)	p_0=0.38201 p_1=0.51736 p_2=0.10063 ω_0=0.03548 ω_1=0.31856 ω_2=1.12623	Yes	36 NEB sites
Model 7	2	2	-7969.190514	N/A	p=0.60548 q=1.55420	N/A	
Model 8	4	2	-7966.59623	Model 7, Model 8a	p=0.70232 q=2.14966 p_0=0.95867 p_1=0.04133 ω=1.42868	No	
Model 8a	4	1	-7967.44171	N/A	p=0.74527 q=2.56519 p_0=0.92276 p_1=0.07724 ω=1.00000	N/A	
Model A	3	2	-7983.917314	Model A	p_0=0.58422 p_1=0.16254 p_2=0.19812 p_3=0.05512 ω_0=0.13840 ω_1=1.00000 ω_2=76.57178	Yes	66 BEB sites
Model A Null	3	1	-7992.406621	N/A	p_0=0.56296 p_1=0.16366 p_2=0.21181 p_3=0.06157 ω_0=0.12948 ω_1=1.00000 ω_2=1.00000	N/A	
Model B	5	2	-7969.955336	Model B	p_0=0.45178 p_1=0.37464 p_2=0.09489 p_3=0.07869 ω_0=0.06506 ω_1=0.52181 ω_2=76.61156	Yes	33 NEB sites

Table 2.8 Legend:

Shown on page 152 overleaf is a table summarising the results of selective pressure analysis on SULT1E1 with the Eutheria as foreground lineage. The **Model** column shows the name of the model used. The **p** column shows the number of free parameters in the ω distribution that are estimated under the given model. The ***ω* (t=0)** column shows the initial ω value used in the Codeml run from which results were taken. The **lnL** column shows the log-likelihood of the given model. The **LRT Result** column shows the result of LRTs (if any) for the given model — this shows the null model unless the null was rejected, in which case the alternative model is shown. For example, the **LRT Result** column shows Model A because likelihood ratio tests rejected both Model 1a and Model A Null in favour of their alternative hypothesis (i.e. Model A). Had either of the likelihood ratio tests failed to reject the null hypothesis, this column would instead display the relevant null model(s) (e.g. Model 1a or Model A Null). The **Parameter Estimates** column shows the parameter estimates of each given model for the current dataset. The ω value estimates for each model are as described in Section 1.8.3. The parameters of the form p_i (where i is an integer) indicate the proportion of codon sites in site class i — note that the branch-site model parameters p_2 and p_3 refer to the proportion of codon sites in site classes 2a and 2b, respectively. In Models 7, 8 and 8a, the parameters p and q are the shape parameters of the β distribution used to model sites undergoing neutral evolution and negative selection. The **Positive Selection** column indicates whether positive selection was predicted under the given model. Finally, the rightmost column shows the number of positively selected sites, if any. The **LRT Result** column indicates that both Model 1a and Model A Null were rejected, and that the foreground ω value (ω_2) is much greater than 1, indicating that positive selection has occurred in the Eutherian lineage. Note the very high values for Models A and B of the foreground statistic ω_2. As highlighted by Zhai *et al.* (2012), such high ω estimates may occur at sites with low rates of synonymous substitution (estimates of ω can theoretically range up to infinity, although in practice are typically bounded at a high number, such as 999). However, as these authors point out, such low synonymous substitution rates do not in any case affect the results of the likelihood ratio test (Zhai *et al.* 2012).

Table 2.8: Results of selective pressure analysis on SULT1E1 with respect to the Eutherian lineage.

Model	p	ω (t=0)	lnL	LRT Result	Parameter Estimates	Positive Selection	Positive Sites P(ω>1) > 0.5
Model 0	1	2	-10075.09816	N/A	ω=0.25900	No	
Model 1a	2	2	-9945.250437	N/A	p_0=0.69879 p_1=0.30121 ω_0=0.17240 ω_1=1.00000	N/A	
Model 2a	4	2	-9945.250437	Model 1a	p_0=0.69879 p_1=0.22096 p_2=0.08025 ω_0=0.17240 ω_1=1.00000 ω_2=1.00000	No	
Model 3 (k=2)	3	2	-9893.906718	Model 3 (k=2)	p_0=0.48245 p_1=0.51755 ω_0=0.07814 ω_1=0.48420	No	
Model 3 (k=3)	5	2	-9874.524571	Model 3 (k=3)	p_0=0.27743 p_1=0.51244 p_2=0.21014 ω_0=0.02808 ω_1=0.26382 ω_2=0.72949	No	
Model 7	2	2	-9872.756688	N/A	p=0.70625 q=1.66399	N/A	
Model 8	4	2	-9872.083947	Model 7, Model 8a	p=0.75471 q=1.91925 p_0=0.97875 p_1=0.02125 ω=1.24002	No	
Model 8a	4	1	-9872.33275	N/A	p=0.76376 q=2.00360 p_0=0.96786 p_1=0.03214 ω=1.00000	N/A	
Model A	3	2	-9926.70281	Model A	p_0=0.59513 p_1=0.19542 p_2=0.15767 p_3=0.05177 ω_0=0.17847 ω_1=1.00000 ω_2=152.19959	Yes	42 BEB sites
Model A Null	3	1	-9939.577474	N/A	p_0=0.60012 p_1=0.23004 p_2=0.12278 p_3=0.04706 ω_0=0.16848 ω_1=1.00000 ω_2=1.00000	N/A	
Model B	5	2	-9883.394406	Model B	p_0=0.41858 p_1=0.41387 p_2=0.08425 p_3=0.08330 ω_0=0.08010 ω_1=0.49403 ω_2=68.60378	Yes	33 NEB sites

Table 2.9 Legend:

Shown on page 154 overleaf is a table summarising the results of selective pressure analysis on CSF3R with Murinae and Eutheria as foreground lineages. The **Model** column shows the name of the model used. The **p** column shows the number of free parameters in the ω distribution that are estimated under the given model. The ***ω* (t=0)** column shows the initial ω value used in the Codeml run from which results were taken. The **lnL** column shows the log-likelihood of the given model. The **LRT Result** column shows the result of LRTs (if any) for the given model — this shows the null model unless the null was rejected, in which case the alternative model is shown. For example, the **LRT Result** column shows Model A for both Murinae and Eutheria because in both cases, likelihood ratio tests rejected both Model 1a and Model A Null in favour of their alternative hypothesis (i.e. Model A). Had any of the likelihood ratio tests failed to reject the null hypothesis, this column would instead display the relevant null model(s) (e.g. Model 1a or Model A Null). The **Parameter Estimates** column shows the parameter estimates of each given model for the current dataset. The ω value estimates for each model are as described in Section 1.8.3. The parameters of the form p_i (where i is an integer) indicate the proportion of codon sites in site class i — note that the branch-site model parameters p_2 and p_3 refer to the proportion of codon sites in site classes 2a and 2b, respectively. In Models 7, 8 and 8a, the parameters p and q are the shape parameters of the β distribution used to model sites undergoing neutral evolution and negative selection. The **Positive Selection** column indicates whether positive selection was predicted under the given model. Finally, the rightmost column shows the number of positively selected sites, if any. The **LRT Result** column indicates that both Model 1a and Model A Null were rejected in both Murinae and Eutheria, and that the foreground ω value (ω_2) is greater than 1, indicating that positive selection has occurred in both Murinae and Eutheria. Note the very high values for Models A and B of the foreground statistic ω_2 when Eutheria is the foreground. As highlighted by Zhai *et al.* (2012), such high ω estimates may occur at sites with low rates of synonymous substitution (estimates of ω can theoretically range up to infinity, although in practice are typically bounded at a high number, such as 999). However, as these authors point out, such low synonymous substitution rates do not in any case affect the results of the likelihood ratio test (Zhai *et al.* 2012).

Table 2.9: Results of selective pressure analysis on CSF3R with respect to Murinae and Eutheria.

	Model	p	ω (t=0)	lnL	LRT Result	Parameter Estimates	Positive Selection	Positive Sites $P(\omega>1) > 0.5$
Sites	Model 0	1	2	-34810.28977	N/A	ω=0.25389	No	
	Model 1a	2	2	-34446.25255	N/A	p_0=0.69403 p_1=0.30597 ω_0=0.19121 ω_1=1.00000	N/A	
	Model 2a	4	10	-34443.296	Model 1a	p_0=0.69 p_1=0.306 p_2=0.0043 ω_0=0.192 ω_1=1.0 ω_2=4.535	No	
	Model 3 (k=2)	3	2	-34296.02441	Model 3 (k=2)	p_0=0.42397 p_1=0.57603 ω_0=0.09420 ω_1=0.43098	No	
	Model 3 (k=3)	5	2	-34227.56711	Model 3 (k=3)	p_0=0.248 p_1=0.526 p_2=0.225 ω_0=0.055 ω_1=0.252 ω_2=0.709	No	
	Model 7	2	2	-34219.04323	N/A	p=1.02382 q=2.37416	N/A	
	Model 8	4	2	-34208.10102	Model 8	p=1.08158 q=2.6637 p_0=0.98619 p_1=0.01381 ω=2.21891	Yes	13 BEB sites
	Model 8a	4	1	-34211.60249	N/A	p=1.18879 q=3.34611 p_0=0.93917 p_1=0.06083 ω=1.00000	N/A	
Murinae	Model A	3	2	-34435.75334	Model A	p_0=0.64990 p_1=0.28453 p_2=0.04561 p_3=0.01997 ω_0=0.18589 ω_1=1.00000 ω_2=3.58641	Yes	16 BEB sites
	Model A Null	3	1	-34438.22543	N/A	p_0=0.58931 p_1=0.25579 p_2=0.10802 p_3=0.04688 ω_0=0.18481 ω_1=1.00000 ω_2=1.00000	N/A	
	Model B	5	2	-34286.27731	Model B	p_0=0.41128 p_1=0.53010 p_2=0.02561 p_3=0.03301 ω_0=0.09402 ω_1=0.42904 ω_2=4.37461	Yes	22 NEB sites
Eutheria	Model A	3	2	-34430.92685	Model A	p_0=0.66857 p_1=0.27940 p_2=0.03670 p_3=0.01534 ω_0=0.19033 ω_1=1.00000 ω_2=53.71444	Yes	11 BEB sites
	Model A Null	3	1	-34445.03703	N/A	p_0=0.65546 p_1=0.28497 p_2=0.04152 p_3=0.01805 ω_0=0.18935 ω_1=1.00000 ω_2=1.00000	N/A	
	Model B	5	2	-34279.40456	Model B	p_0=0.39449 p_1=0.54387 p_2=0.02591 p_3=0.03573 ω_0=0.09238 ω_1=0.42512 ω_2=35.70343	Yes	29 NEB sites

2.4.2 Functional Implications of Positively Selected Sites

This section will focus on three broad categories of positively selected site identified by Codeml across six genes: chemokine binding protein 2 (CCBP2), colony stimulating factor 3 receptor (CSF3R), Ets2 repressor factor (ERF), pleckstrin homology-like domain, family A, member 2 (PHLDA2), sulfotransferase family 1E, estrogen-preferring, member 1 (SULT1E1), and ZFP36 ring finger protein-like 1 (ZFP36L1). These broad categories of positively selected site — (I) sites undergoing functional shift, (II) sites under relaxed selective constraint, and (III) sites erroneously inferred to be positively selected — differ in the degree to which the inferred positive selection could be viewed as being indicative of a functional shift.

I. Signals of Positive Selection: Functional Shift

Four genes are discussed here that have signals of positive selection indicative of functional shift: (i) CCBP2, (ii) CSF3R, (iii) ERF, and (iv) ZFP26L1. In each case, the positively selected sites are discussed with reference to a positively selected site (PSS) overview diagram and alignment excerpt, and in light of the expected effect of observed substitutions, at each positively selected site, on the conformation and/or function of the relevant protein.

(i) CCBP2

The positively selected site (PSS) overview diagram in Figure 2.13 shows the Swiss-Prot annotation for chemokine binding protein 2 (CCBP2) — and in particular for an enlarged region from sites 115-200 — overlaid by codon sites undergoing positive selection in the Eutherian lineage as inferred using Codeml. Under each overview diagram is a breakdown of the Swiss-Prot annotation for CCBP2 (Swiss-Prot accession O00590). The overall structure of the CCBP2 protein can be seen from the alternating topological and transmembrane domains (i.e. `TOPO_DOM` and `TRANSMEM`, respectively); CCBP2 is a transmembrane protein that is threaded repeatedly through the membrane, alternating between cytosolic and extracellular domains.

The enlarged region highlights the cysteine sites at positions 117 and 195 (see sites annotated as `DISFULID` in Figure 2.13), which form an intramolecular disulfide bond between two extracellular domains. Five positively selected sites flank these two cysteines: the first cysteine is adjacent to two positively selected sites at positions 118 and 119, while the second cysteine is flanked by positively selected sites at positions 193, 196 and 197.

It is important to consider the impact of the amino acid substitutions in terms of the physicochemical properties/changes incurred at each positively selected site. Figure 2.14 shows the relevant sites of the nucleotide alignment for CCBP2. Site 118 is a lysine (K) in Eutheria and a proline (P) in the background, while site 119 is either a methionine (M) or valine (V) in Eutheria and either a phenylalanine (F) or valine (V) in the background. The substitutions at site 119 are somewhat neutral, but the substitution of proline with lysine would not be considered neutral and may have altered the local conformation of CCBP2 (Betts and Russell 2003).

With respect to the positively selected sites at 193, 196 and 197: site 193 is an isoleucine (I) in background taxa but a tryptophan (W) in Eutheria; site 196 is an alanine (A) or threonine (T) in background taxa but a histidine (H) or tyrosine (Y) in Eutheria; and site 197 is a histidine (H) or glutamine (Q) in background taxa but an alanine (A) or proline (P) in the Eutherian foreground taxa. Of these, the substitutions at site 196 would likely introduce the biggest change in local conformation of CCBP2, since this site has hydrophobic amino acids in background species and hydrophilic amino acids in Eutherian taxa; this hydrophilic property might locally draw the peptide out of the membrane into extracellular space (Betts and Russell 2003). Taken together, these substitutions and their propensity to change the local physicochemical properties of CCBP2 may indicate that these sites were under positive selection to alter the way that the cysteines at sites 117 and 195 of CCBP2 are presented for the formation of a disulphide bridge.

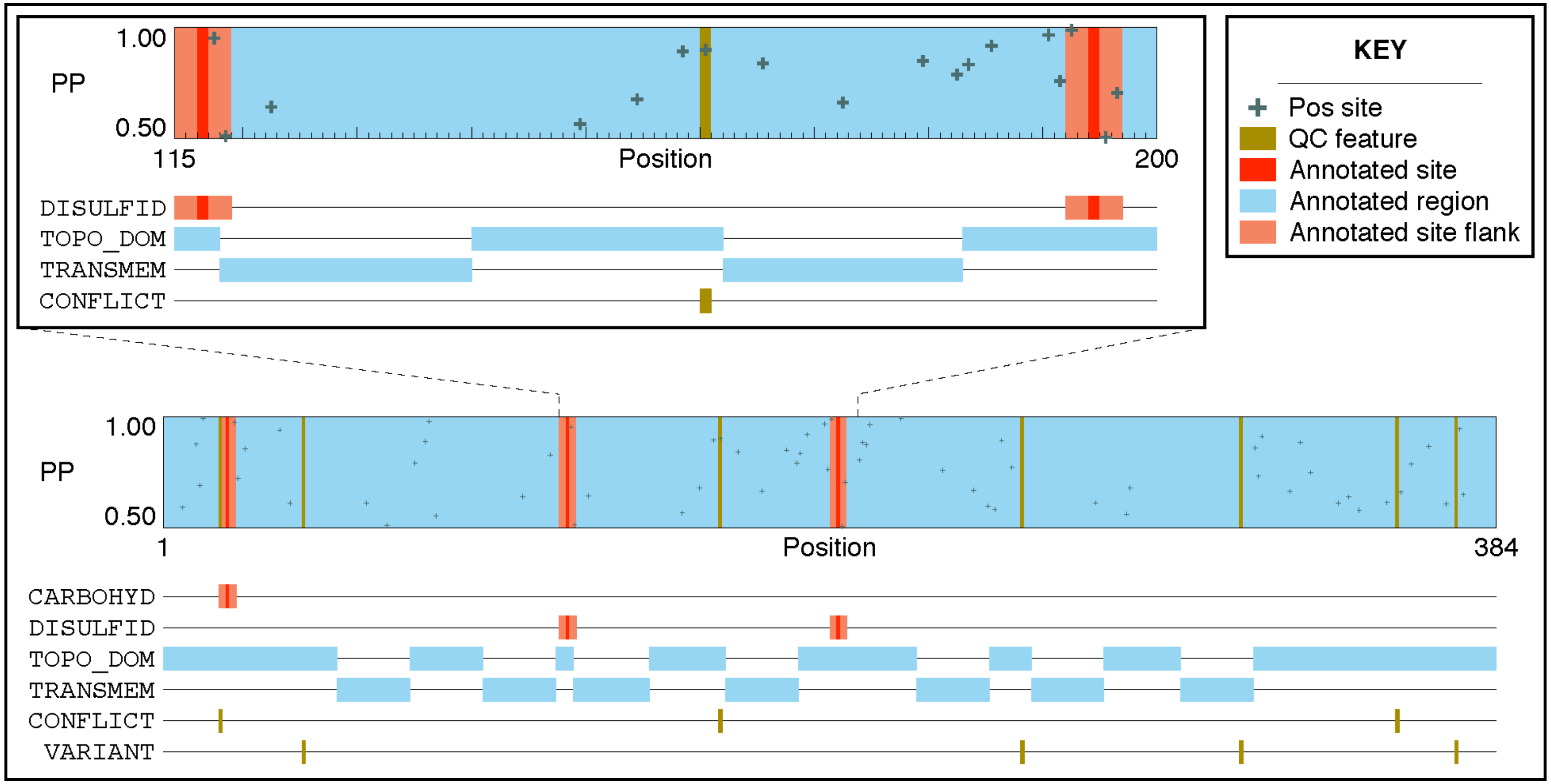
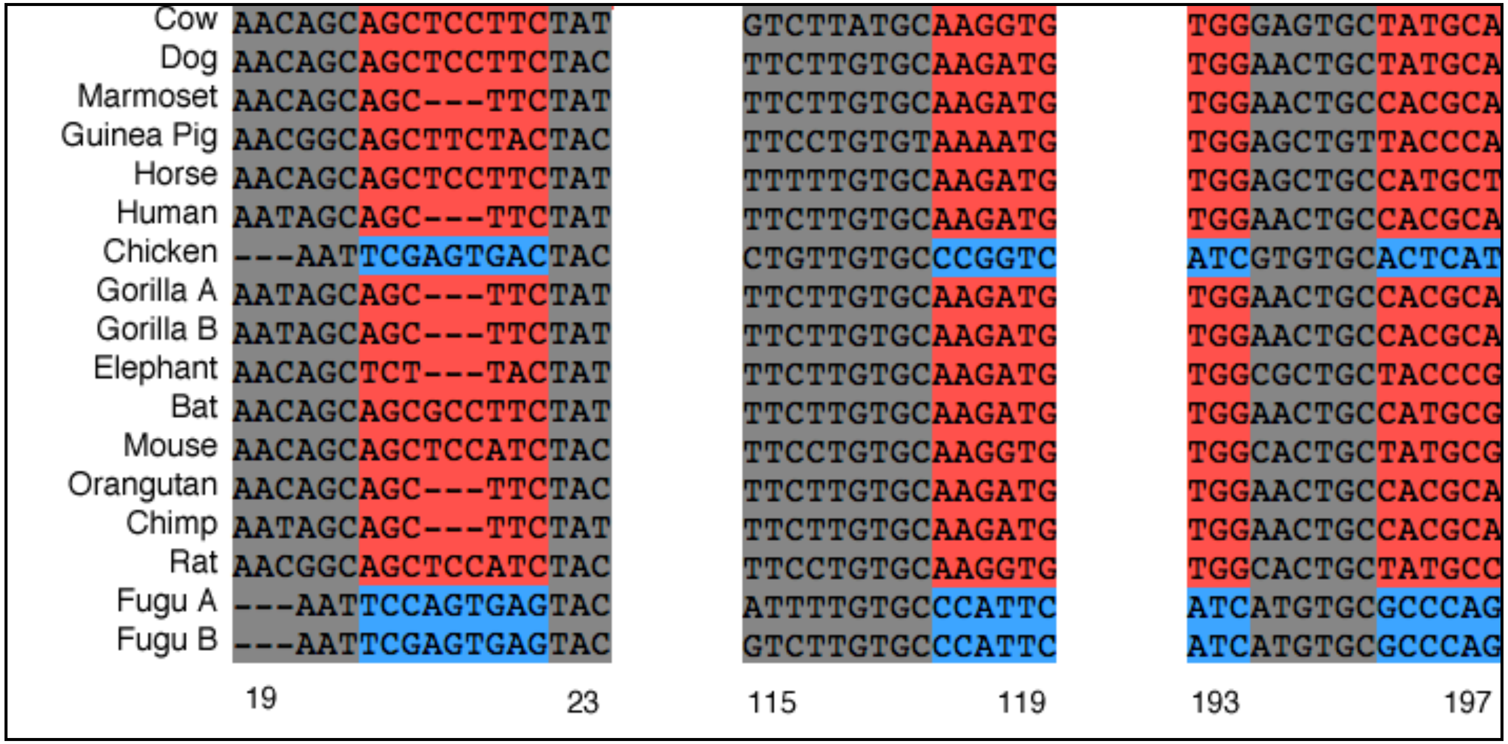


Figure 2.13: Overview of positively selected sites in CCBP2 with respect to the Eutherian lineage.

Above is a cartoon representation of CCBP2, showing sites positively selected with respect to Eutheria overlaid on functional annotation for the CCBP2 protein as obtained from Swiss-Prot (UniProt Consortium 2011). Each positively selected site is represented by a plus sign (+); the horizontal position of each positively selected site in the overview indicates its position within CCBP2, while its vertical position indicates the posterior probability (PP) that the site is undergoing positive selection, as inferred by Codeml (Yang 2007, Yang 1997). Below both the main and enlarged overview is a breakdown of the relevant Swiss-Prot annotation.



Cow	AACAGCAGCTCCTTCTAT	GTCTTATGCAAGGTG	TGGGAGTGCTATGCA
Dog	AACAGCAGCTCCTTCTAC	TTCTTGTGCAAGATG	TGGAACTGCTATGCA
Marmoset	AACAGCAGC---TTCTAT	TTCTTGTGCAAGATG	TGGAACTGCCACGCA
Guinea Pig	AACGGCAGCTTCTACTAC	TTCCTGTGTAAAATG	TGGAGCTGTTACCCA
Horse	AACAGCAGCTCCTTCTAT	TTTTTGTGCAAGATG	TGGAGCTGCCATGCT
Human	AATAGCAGC---TTCTAT	TTCTTGTGCAAGATG	TGGAACTGCCACGCA
Chicken	---AATTCGAGTGACTAC	CTGTTGTGCCCGGTC	ATCGTGTGCACTCAT
Gorilla A	AATAGCAGC---TTCTAT	TTCTTGTGCAAGATG	TGGAACTGCCACGCA
Gorilla B	AATAGCAGC---TTCTAT	TTCTTGTGCAAGATG	TGGAACTGCCACGCA
Elephant	AACAGCTCT---TACTAT	TTCTTGTGCAAGATG	TGGCGCTGCTACCCG
Bat	AACAGCAGCGCCTTCTAT	TTCTTGTGCAAGATG	TGGAACTGCCATGCG
Mouse	AACAGCAGCTCCATCTAC	TTCCTGTGCAAGGTG	TGGCACTGCTATGCG
Orangutan	AACAGCAGC---TTCTAC	TTCTTGTGCAAGATG	TGGAACTGCCACGCA
Chimp	AATAGCAGC---TTCTAT	TTCTTGTGCAAGATG	TGGAACTGCCACGCA
Rat	AACGGCAGCTCCATCTAC	TTCCTGTGCAAGGTG	TGGCACTGCTATGCC
Fugu A	---AATTCCAGTGAGTAC	ATTTTGTGCCCATTC	ATCATGTGCGCCCAG
Fugu B	---AATTCGAGTGAGTAC	GTCTTGTGCCCATTC	ATCATGTGCGCCCAG
	19 23	115 119	193 197

Figure 2.14: Positively selected sites in CCBP2 with respect to the Eutherian lineage.

Excerpts are shown from the positively selected site alignment for CCBP2 with respect to the Eutherian lineage. Positively selected sites are shown in red (for foreground sequences) and blue (for background sequences), while all other sites are shown in grey. Provisional gene names (based on species of origin) for the members of the CCBP2 gene family are shown to the left of the figure, with alignment excerpts shown to the right of these. The numbers below each alignment excerpt show the codon positions within the human CCBP2 gene (i.e. ENSG00000144648). Note that because of the gap in human CCBP2 in the fourth codon of the first alignment excerpt, there is a discrepancy between the number of alignment positions shown and the position of alignment columns relative to human CCBP2.

(ii) CSF3R

The positively selected site (PSS) overview diagrams in Figure 2.15 and Figure 2.16 show the Swiss-Prot annotation for colony stimulating factor 3 receptor (CSF3R) in human and mouse, respectively, overlaid by codon sites undergoing positive selection in the Eutherian lineage as inferred using Codeml. Figure 2.15 shows an overview for human CSF3R with respect to positively selected sites in ancestral Eutheria, focusing on sites 600-620. Figure 2.16 shows an overview for mouse CSF3R with respect to positively selected sites in ancestral Murinae, focusing on sites 405-585. Under each of the overview diagrams is a breakdown of the relevant Swiss-Prot annotation: the corresponding Swiss-Prot accessions are Q99062 for human CSF3R and P40223 for mouse CSF3R.

As can be seen from these overviews, CSF3R is a transmembrane protein, with a transmembrane domain (i.e. `TRANSMEM`) close to its C-terminal end. This transmembrane domain is flanked on the N-terminal side by an extracellular domain and on the C-terminal side by a cytoplasmic domain (i.e. `TOPO_DOM`). Positively selected sites in both human and mouse CSF3R are found close to glycosylation sites in the extracellular domain, pointing to a possible effect on the conformation or binding characeristics of this protein (Varki *et al.* 2009).

The rightmost alignment excerpt in Figure 2.17 shows the site at position 608 in human CSF3R that is positively selected in Eutheria, along with sites 609-612. This Eutherian-specific positively selected site lies close to an N-linked glycosylation site at 610 (see site annotated as `CARBOHYD` in Figure 2.15). The N-glycosylation site itself is a conserved asparagine (N) in Eutherian and most non-Eutherian species, where homologous sequence is present. Where present, this asparagine is followed by a serine (S) or glycine (G), then a threonine (T) or serine (S). This follows a known sequence motif associated with N-linked glycosylation sites (i.e. NX[ST], where N is asparagine, X is any amino acid, and the amino acid at the third position may be a serine or threonine) (Gavel and von Heijne 1990).

In most Eutherian species, site 608 is a hydrophobic alanine (A). In mouse and rat, site 608 is a polar serine (S). In non-Eutherian species, site 608 is one of the polar residues serine (S), cysteine (C) or threonine (T). This may point to a change in the local conformation of CSF3R in the ancestral Eutherian lineage due to the juxtaposition of a hydrophobic amino acid with a glycosylation site, which perhaps underwent a reversion to a polar residue in Murinae.

In addition to the potentially functionally relevant reversion of this positively selected site in Murinae, an instance of directional selection is shown in the central alignment excerpt in Figure 2.17. The codon site at 583 in mouse CSF3R is positively selected in Murinae, and is adjacent to an N-linked glycosylation site at position 582 (see site annotated as `CARBOHYD` in Figure 2.16). The N-glycosylation site itself is a conserved asparagine (N) in Eutherian and most non-Eutherian species, where homologous sequence is present. In mouse and rat, site 583 is an isoleucine (I). In all but one of the other species with the conserved asparagine (N) at 582, site 583 is an alanine. In frog, site 583 is a serine (S). This may indicate that functionally relevant changes in the local conformation of CSF3R have occurred that have affected the local conformation of CSF3R and in turn the positioning of the N-glycosylation site at 582.

Another instance of functionally relevant sequence change in ancestral Murinae is shown in the leftmost alignment excerpt in Figure 2.17. Codon site 408 in mouse CSF3R was positively selected in the mouse-rat ancestor, and coincides with an N-linked glycosylation site (see site annotated as `CARBOHYD` in Figure 2.16). This N-glycosylation site contains an asparagine (N) in mouse and rat only. Both mouse and rat CSF3R have an asparagine (N) at this site, followed by a valine (V) and threonine (T); a sequence motif characteristic of N-glycosylation sites (Gavel and von Heijne 1990). This may indicate that this glycosylation site arose anew in the ancestral Murinae lineage, therefore representing a novel glycosylation site in Murinae, and an instance of functional shift due to sequence change.

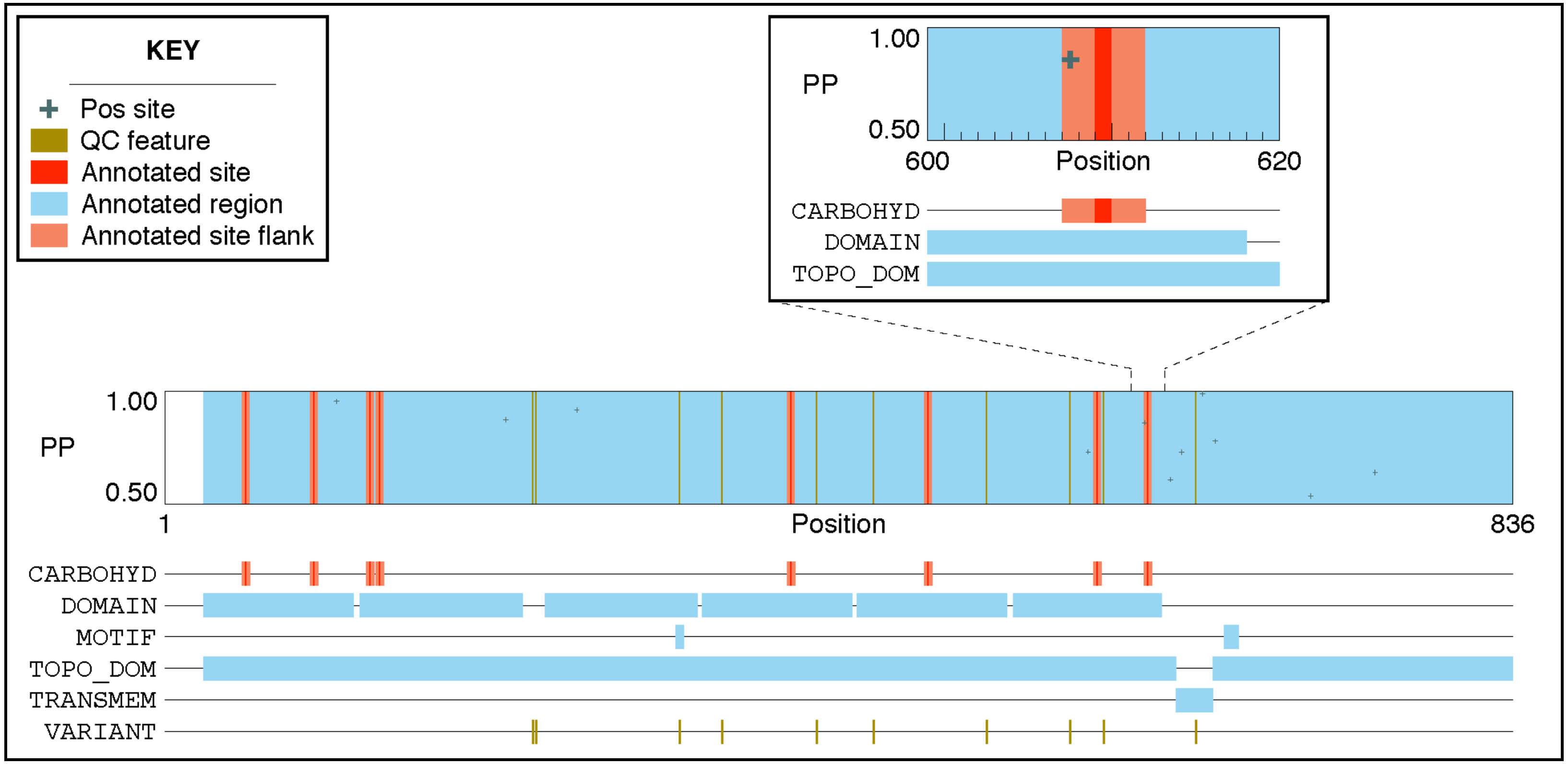


Figure 2.15: Overview of positively selected sites in human CSF3R with respect to the Eutherian lineage.

Above is a cartoon representation of human CSF3R, showing sites positively selected with respect to Eutheria overlaid on functional annotation for the human CSF3R protein as obtained from Swiss-Prot (UniProt Consortium 2011). Each positively selected site is represented by a plus sign (+); the horizontal position of each positively selected site in the overview indicates its position within human CSF3R, while its vertical position indicates the posterior probability (PP) that the site is undergoing positive selection, as inferred by Codeml (Yang 2007, Yang 1997). Below both the main and enlarged overview is a breakdown of the relevant Swiss-Prot annotation.

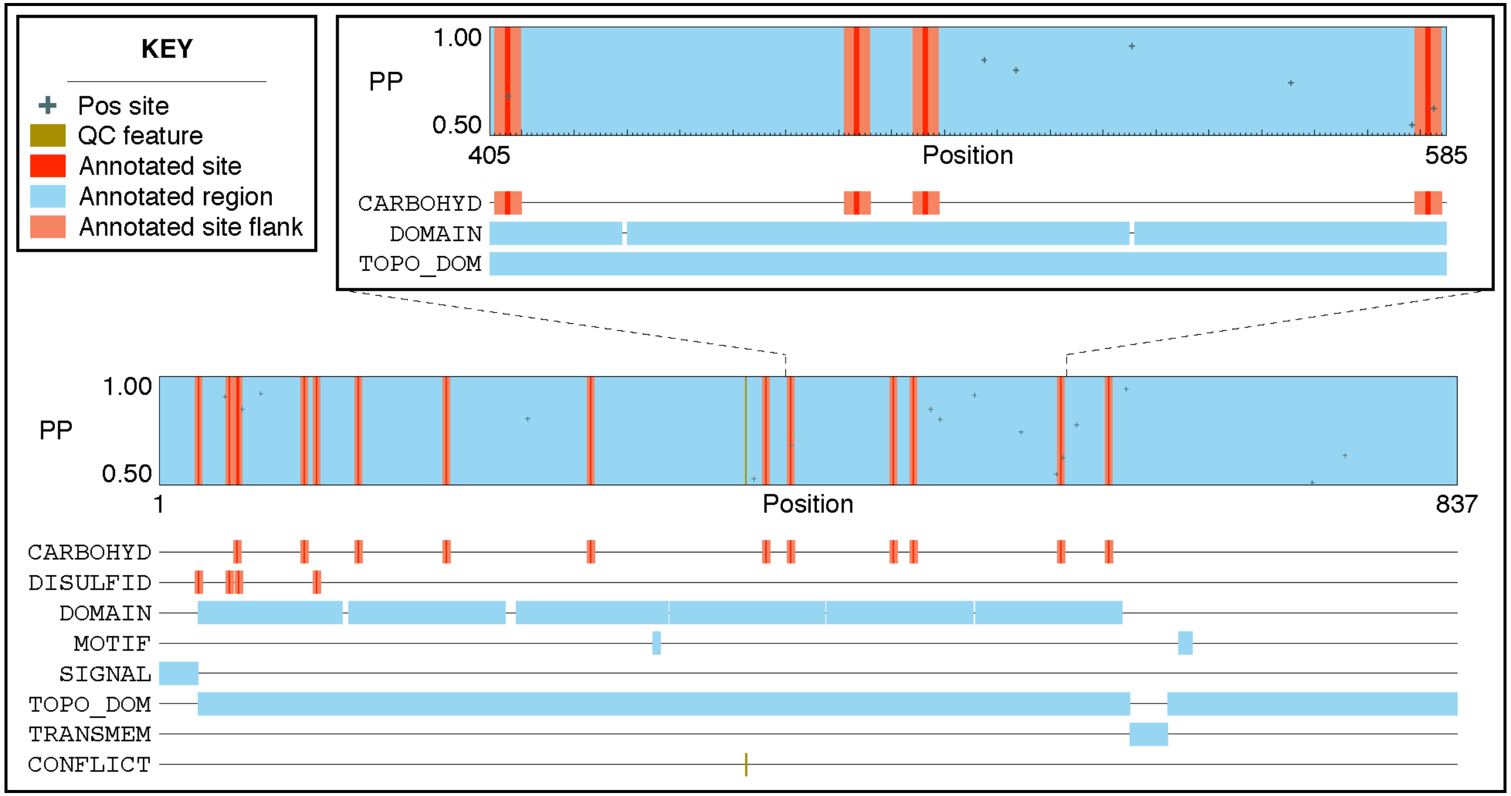
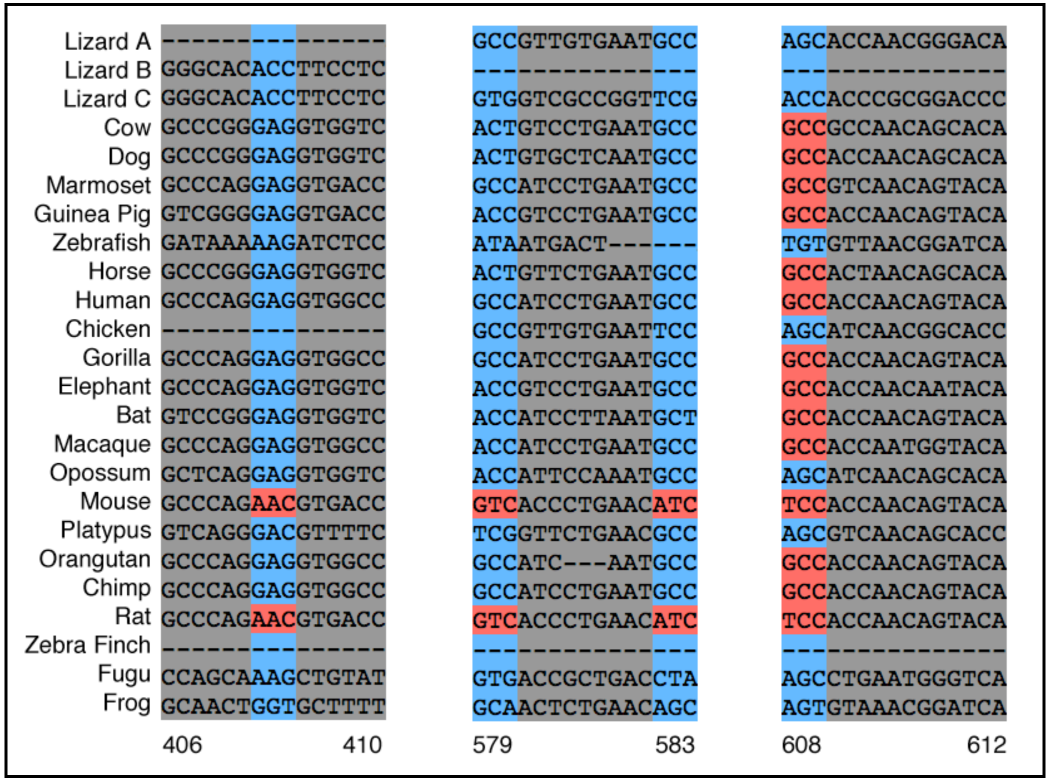


Figure 2.16: Overview of positively selected sites in mouse CSF3R with respect to the lineage Murinae.

Above is a cartoon representation of mouse CSF3R, showing sites positively selected with respect to Eutheria overlaid on functional annotation for the mouse CSF3R protein as obtained from Swiss-Prot (UniProt Consortium 2011). Each positively selected site is represented by a plus sign (+); the horizontal position of each positively selected site in the overview indicates its position within mouse CSF3R, while its vertical position indicates the posterior probability (PP) that the site is undergoing positive selection, as inferred by Codeml (Yang 2007, Yang 1997). Below both the main and enlarged overview is a breakdown of the relevant Swiss-Prot annotation.



Lizard A	---------------	GCCGTTGTGAATGCC	AGCACCAACGGGACA
Lizard B	GGGCACACCTTCCTC	---------------	---------------
Lizard C	GGGCACACCTTCCTC	GTGGTCGCCGGTTCG	ACCACCCGCGGACCC
Cow	GCCCGGGAGGTGGTC	ACTGTCCTGAATGCC	GCCGCCAACAGCACA
Dog	GCCCGGGAGGTGGTC	ACTGTGCTCAATGCC	GCCACCAACAGCACA
Marmoset	GCCCAGGAGGTGACC	GCCATCCTGAATGCC	GCCGTCAACAGTACA
Guinea Pig	GTCGGGGAGGTGACC	ACCGTCCTGAATGCC	GCCACCAACAGTACA
Zebrafish	GATAAAAAGATCTCC	ATAATGACT------	TGTGTTAACGGATCA
Horse	GCCCGGGAGGTGGTC	ACTGTTCTGAATGCC	GCCACTAACAGCACA
Human	GCCCAGGAGGTGGCC	GCCATCCTGAATGCC	GCCACCAACAGTACA
Chicken	---------------	GCCGTTGTGAATTCC	AGCATCAACGGCACC
Gorilla	GCCCAGGAGGTGGCC	GCCATCCTGAATGCC	GCCACCAACAGTACA
Elephant	GCCCAGGAGGTGGTC	ACCGTCCTGAATGCC	GCCACCAACAATACA
Bat	GTCCGGGAGGTGGTC	ACCATCCTTAATGCT	GCCACCAACAGTACA
Macaque	GCCCAGGAGGTGGCC	ACCATCCTGAATGCC	GCCACCAATGGTACA
Opossum	GCTCAGGAGGTGGTC	ACCATTCCAAATGCC	AGCATCAACAGCACA
Mouse	GCCCAGAACGTGACC	GTCACCCTGAACATC	TCCACCAACAGTACA
Platypus	GTCAGGGACGTTTTC	TCGGTTCTGAACGCC	AGCGTCAACAGCACC
Orangutan	GCCCAGGAGGTGGCC	GCCATC---AATGCC	GCCACCAACAGTACA
Chimp	GCCCAGGAGGTGGCC	GCCATCCTGAATGCC	GCCACCAACAGTACA
Rat	GCCCAGAACGTGACC	GTCACCCTGAACATC	TCCACCAACAGTACA
Zebra Finch	---------------	---------------	---------------
Fugu	CCAGCAAAGCTGTAT	GTGACCGCTGACCTA	AGCCTGAATGGGTCA
Frog	GCAACTGGTGCTTTT	GCAACTCTGAACAGC	AGTGTAAACGGATCA
	406 410	579 583	608 612

Figure 2.17: Positively selected sites in CSF3R with respect to stem lineages of Murinae and Eutheria.

Excerpts are shown from the positively selected site alignments for CSF3R with respect to Murinae (at left and centre) and Eutheria (at right). The numbers below the alignment excerpts at left and centre show the codon positions within the mouse CSF3R gene (i.e. ENSMUSG00000028859), while the numbers below the alignment excerpt at right show the codon positions within the human CSF3R gene (i.e. ENSG00000119535). Positively selected sites are shown in red (for foreground sequences) and blue (for background sequences), while all other sites are shown in grey. Provisional gene names (based on species of origin) for the members of the CSF3R gene family are shown to the left of the figure, with the alignment excerpts shown to the right of these.

(iii) ERF

The positively selected site (PSS) overview diagram in Figure 2.18 shows the Swiss-Prot annotation for Ets2 repressor factor (ERF) — in particular for two enlarged regions from sites 140-160 and 425-450 — overlaid by codon sites undergoing positive selection in the Eutherian lineage as inferred using Codeml. Under each overview diagram is a breakdown of the Swiss-Prot annotation for ERF (Swiss-Prot accession P50548). Similarly, Figure 2.19 shows alignment excerpts for two regions of ERF, from sites 146-150 and 433-437, respectively. Both the alignment excerpts and the PSS overviews highlight two particular positively selected sites at positions 147 and 435. Both of these positively selected sites lie close to modified residues: site 147 is adjacent to a phosphothreonine at position 148, while site 435 is annotated as a phosphoserine. (See sites annotated as `MOD_RES` in Figure 2.18.)

Regarding the Eutherian positively selected site at position 147: all homologous sequences have a threonine at site 148, in both Eutherian and non-Eutherian species, while the positively selected site at 147 is a serine (S) in Eutherian species and a cysteine (C) in the non-Eutherian species studied.

The positively selected site 435 of this intracellular protein coincides with a phosphoserine residue. Each Eutherian sequence contains a serine (S) at this site, while the homologous site of non-Eutherian sequences from zebrafish and frog contain a glutamine and glutamic acid, respectively. This may point to the emergence of this phosphorylation site in the ancestral Eutherian lineage, therefore representing a novel phosphorylation site in Eutheria, which may affect the signalling behaviour of this transcription factor.

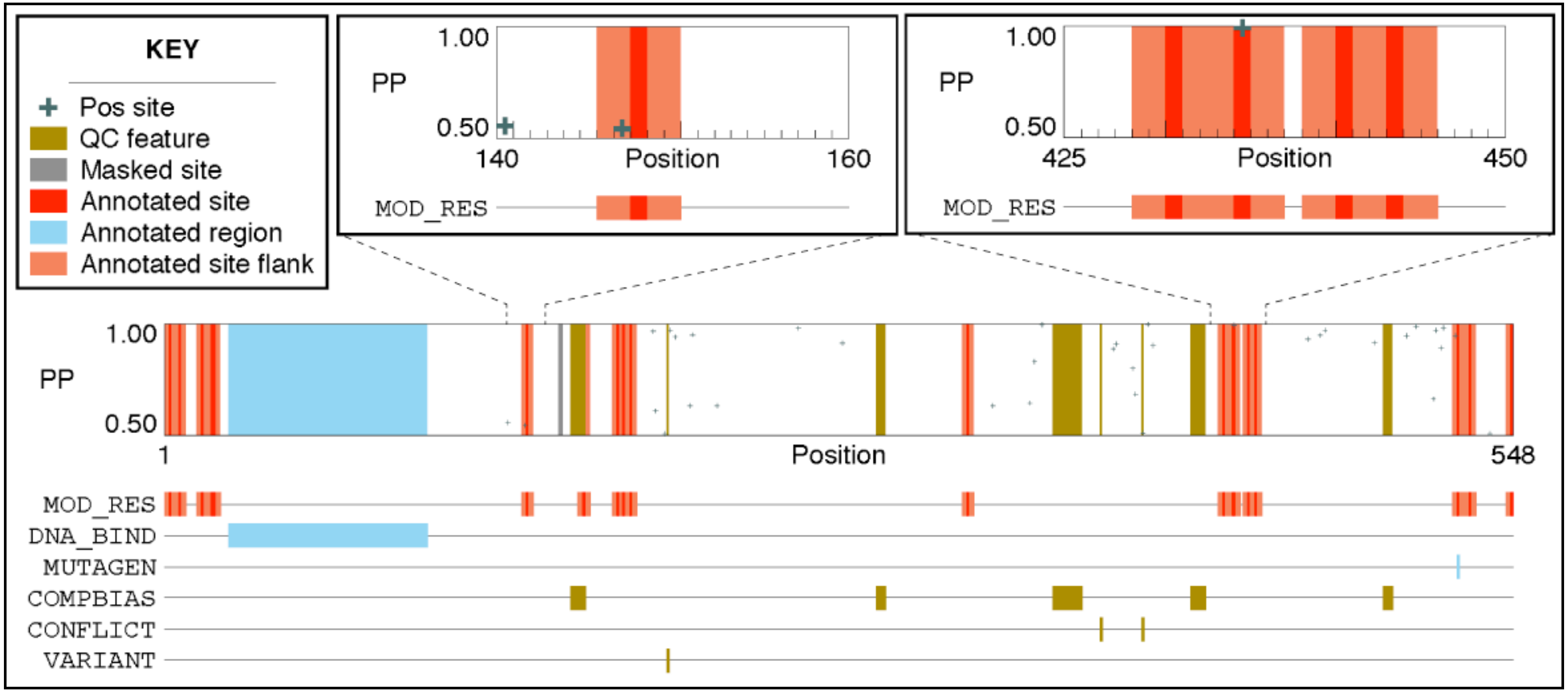


Figure 2.18: Overview of positively selected sites in human ERF with respect to the Eutherian lineage.

Above is a cartoon representation of human ERF, showing sites positively selected with respect to Eutheria overlaid on functional annotation for the human ERF protein as obtained from Swiss-Prot (UniProt Consortium 2011). Each positively selected site is represented by a plus sign (+); the horizontal position of each positively selected site in the overview indicates its position within human ERF, while its vertical position indicates the posterior probability (PP) that the site is undergoing positive selection, as inferred by Codeml (Yang 2007, Yang 1997). Below both the main and enlarged overview is a breakdown of the relevant Swiss-Prot annotation.

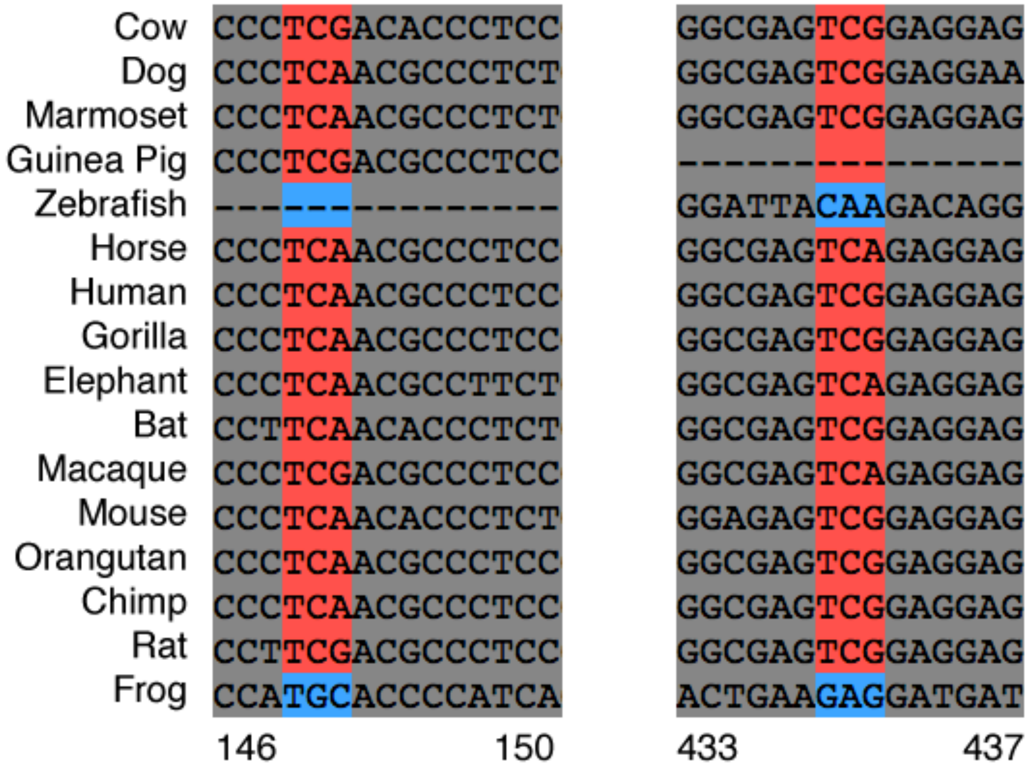

Cow CCCTCGACACCCTCC GGCGAGTCGGAGGAG
Dog CCCTCAACGCCCTCT GGCGAGTCGGAGGAA
Marmoset CCCTCAACGCCCTCT GGCGAGTCGGAGGAG
Guinea Pig CCCTCGACGCCCTCC ---------------
Zebrafish --------------- GGATTACAAGACAGG
Horse CCCTCAACGCCCTCC GGCGAGTCAGAGGAG
Human CCCTCAACGCCCTCC GGCGAGTCGGAGGAG
Gorilla CCCTCAACGCCCTCC GGCGAGTCGGAGGAG
Elephant CCCTCAACGCCTTCT GGCGAGTCAGAGGAG
Bat CCTTCAACACCCTCT GGCGAGTCGGAGGAG
Macaque CCCTCGACGCCCTCC GGCGAGTCAGAGGAG
Mouse CCCTCAACACCCTCT GGAGAGTCGGAGGAG
Orangutan CCCTCAACGCCCTCC GGCGAGTCGGAGGAG
Chimp CCCTCAACGCCCTCC GGCGAGTCGGAGGAG
Rat CCTTCGACGCCCTCC GGCGAGTCGGAGGAG
Frog CCATGCACCCCATCA ACTGAAGAGGATGAT
146
150
433
437

Figure 2.19: Positively selected sites in ERF with respect to the Eutherian lineage. Excerpts are shown from the positively selected site alignment for ERF with respect to the Eutherian lineage. Positively selected sites are shown in red (for foreground sequences) and blue (for background sequences), while all other sites are shown in grey. Provisional gene names (based on species of origin) for the members of the ERF gene family are shown to the left of the figure, with alignment excerpts shown to the right of these. The numbers below each alignment excerpt show the codon positions within the human ERF gene (i.e. ENSG00000105722).

(iv) ZFP36L1

The positively selected site (PSS) overview diagram in Figure 2.20 shows the Swiss-Prot annotation for ZFP36 ring finger protein-like 1 (ZFP36L1) — with a particular focus on the enlarged region at sites 71-107 — overlaid by codon sites undergoing positive selection in the Eutherian lineage as inferred using Codeml. Under each overview diagram is a breakdown of the Swiss-Prot annotation for ZFP36L1 (Swiss-Prot accession Q07352). The region focused on in the PSS overview diagram notably includes all ZFP36L1 sites that are positively selected in the ancestral Eutherian lineage. One positively selected site — at position 94 in human ZFP36L1 — lies close to a modified residue at position 92 (see site annotated as `MOD_RES` in Figure 2.20).

Figure 2.21 shows an alignment excerpt for a region of ZFP36L1 from sites 90-94. The Eutherian-specific positively selected site at position 94 lies two sites away from a phosphoserine site at 92. The phosphoserine site itself appears to be conserved in all homologous sequences, including some non-Eutherian species. Site 94 is a glycine (G) in Eutherian species and a threonine (T) or asparagine (N) in all non-Eutherian species. These substitutions are generally disfavoured and thus unlikely to occur unless they are under positive selective pressure to do so (Betts and Russell 2003). This may therefore indicate a functionally relevant change in the local conformation of ZFP36L1, due to a sequence substitution in the ancestral Eutherian lineage, which affects the positioning — and in turn the function — of the neighbouring phosphoserine.

III. Signals of Positive Selection: Relaxed Selective Constraint

Two genes are discussed here that have signals of positive selection that may be indicative of functional shift, but may also reflect relaxed selective constraint: (i) PHLDA2 and (ii) SULT1E1. In each case, positively selected sites are discussed with reference to a positively selected site (PSS) overview diagram and alignment excerpt, and in light of the expected effect of observed substitutions at each positively selected site.

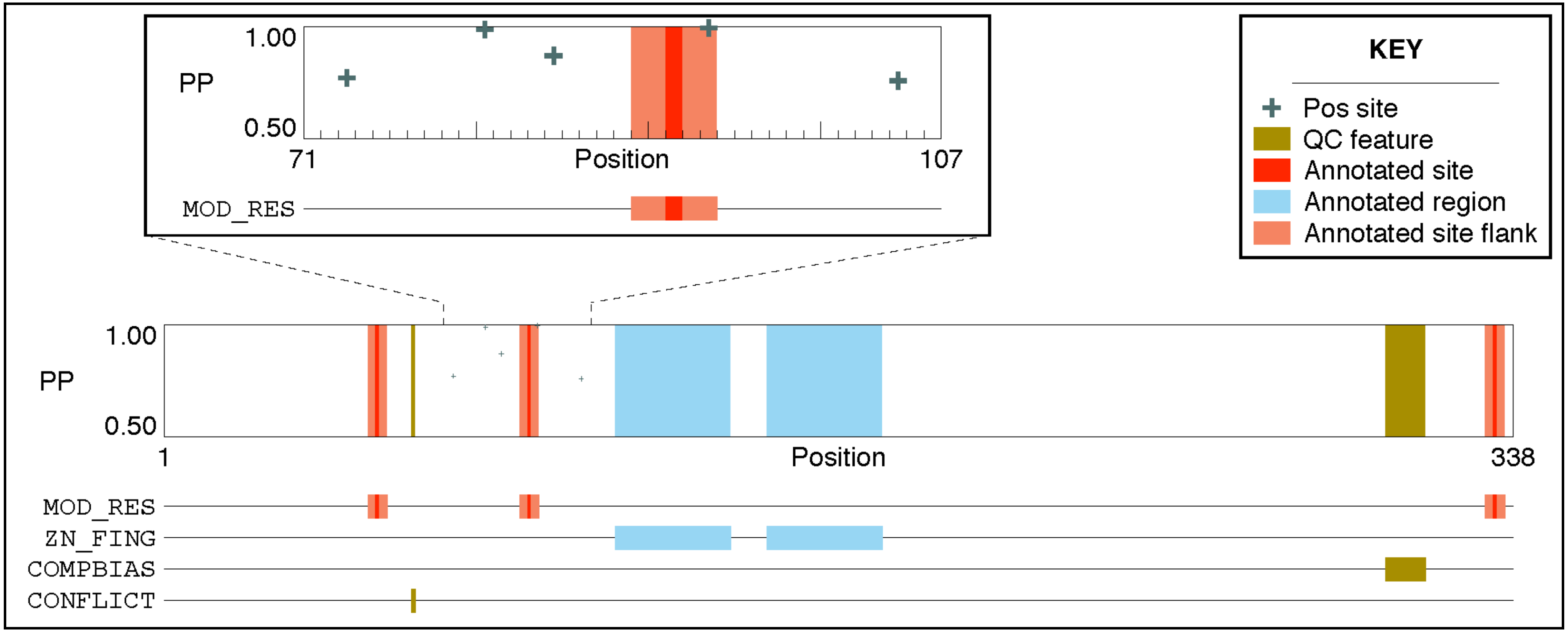


Figure 2.20: Overview of positively selected sites in human ZFP36L1 with respect to the Eutherian lineage.

Above is a cartoon representation of human ZFP36L1, showing sites positively selected with respect to Eutheria overlaid on functional annotation for the human ZFP36L1 protein as obtained from Swiss-Prot (UniProt Consortium 2011). Each positively selected site is represented by a plus sign (+); the horizontal position of each positively selected site in the overview indicates its position within human ZFP36L1, while its vertical position indicates the posterior probability (PP) that the site is undergoing positive selection, as inferred by Codeml (Yang 2007, Yang 1997). Below both the main and enlarged overview is a breakdown of the relevant Swiss-Prot annotation.

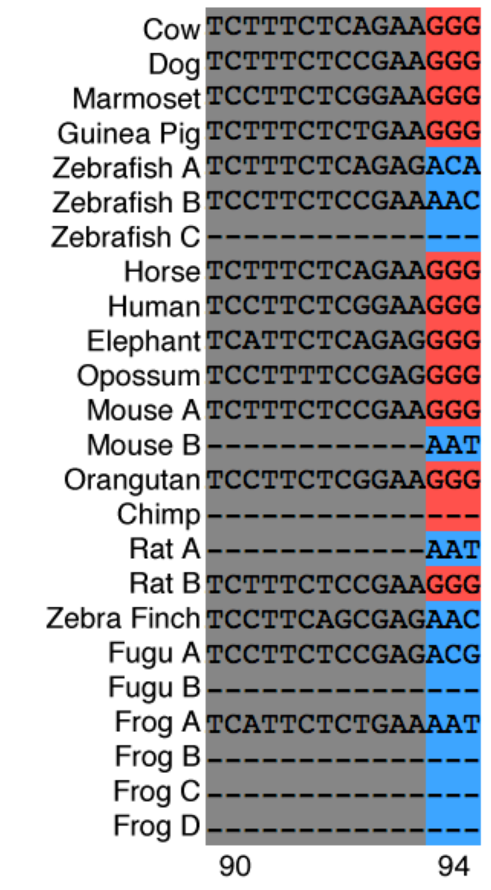


Figure 2.21: Positively selected sites in ZFP36L1 with respect to the Eutherian lineage. Excerpts are shown from the positively selected site alignment for ZFP36L1 with respect to the Eutherian lineage. Positively selected sites are shown in red (for foreground sequences) and blue (for background sequences), while all other sites are shown in grey. Provisional gene names (based on species of origin) for the members of the ZFP36L1 gene family are shown to the left of the figure, with alignment excerpts shown to the right of these. The numbers below each alignment excerpt show the codon positions within the human ZFP36L1 gene (i.e. ENSG00000185650).

(i) PHLDA2

The positively selected site (PSS) overview diagram in Figure 2.22 shows the Swiss-Prot annotation for pleckstrin homology-like domain, family A, member 2 (PHLDA2) — in particular for an enlarged region at sites 1-50 — overlaid by codon sites undergoing positive selection in the Eutherian lineage as inferred using Codeml. Under each overview diagram is a breakdown of the Swiss-Prot annotation for PHLDA2 (Swiss-Prot accession Q53GA4).

As can be seen from the alignment excerpts in Figure 2.23, two Eutherian-specific positively selected sites at 2 and 43 are each adjacent to a phosphoserine residue at sites 3 and 42, respectively. (See sites annotated as `MOD_RES` in Figure 2.22.)

Both phosphoserine residues are serine (S) throughout primates. Site 3 is also a serine in cow and lizard, and is a threonine (T) in the remaining Eutherian species. In Eutherian species, site 2 is a lysine (K) or arginine (R). In non-Eutherian species, site 2 is a serine (S), threonine (T) or glycine (G). The amino acids present at this site in the foreground would be relatively less favoured to substitute the amino acids in the background sequences (Betts and Russell 2003). Site 43 is one of proline (P), lysine (K), asparagine (N) or alanine (A) in Eutherian species, but is either lysine (K) or arginine (R) in non-Eutherian species. The substitutions at this site would be considered to be relatively neutral (Betts and Russell 2003). While the positive selection at site 2 involves somewhat disfavoured substitutions potentially indicative of functional shift, the substitutions at site 43 are relatively neutral — i.e. close to being functionally equivalent (Betts and Russell 2003) — and therefore may simply indicate a relaxation of selective constraint.

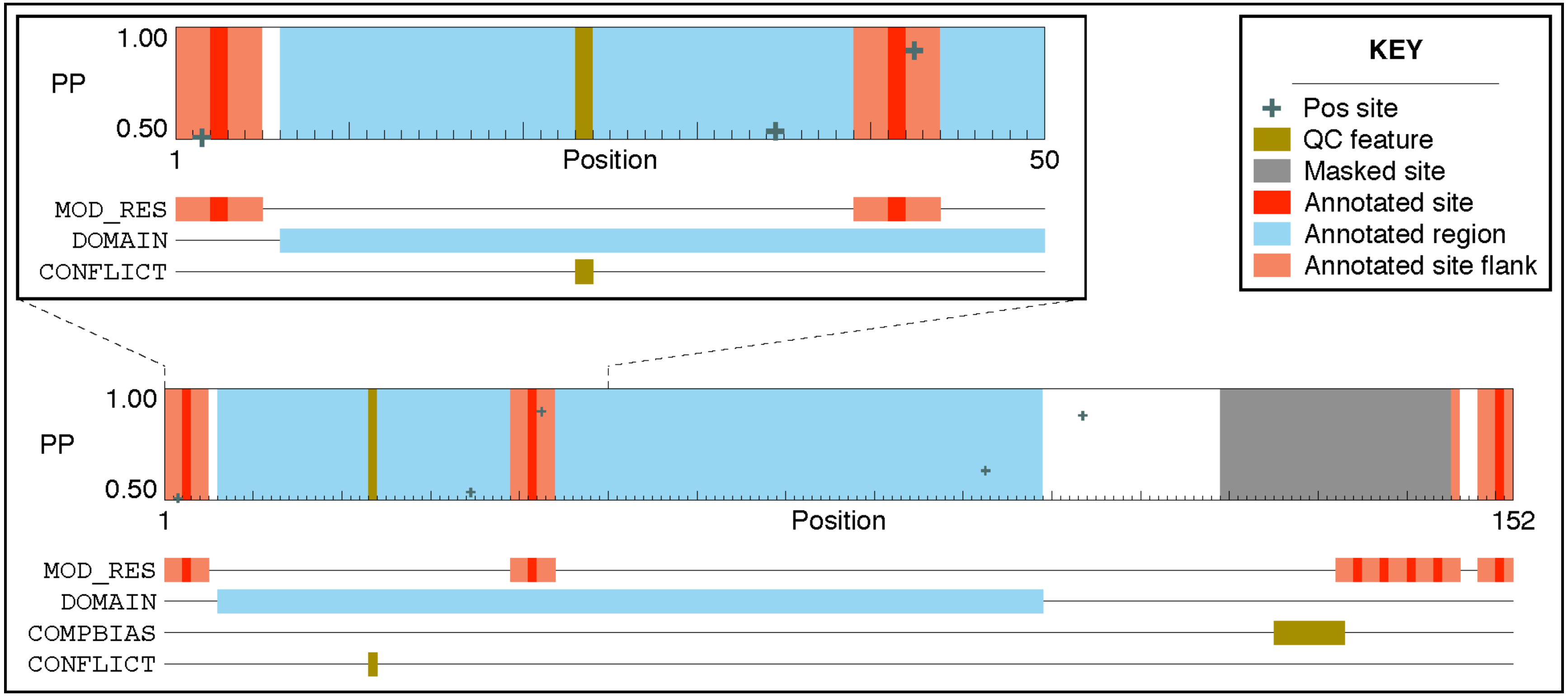
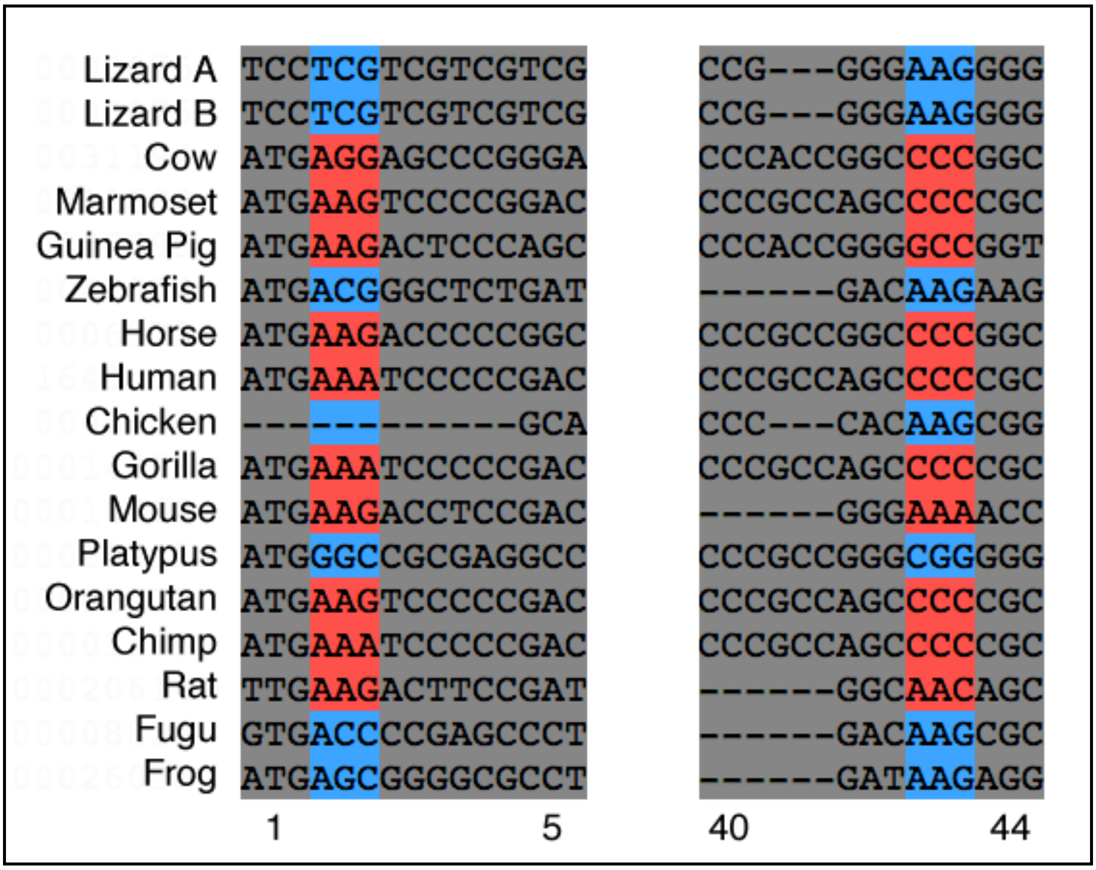


Figure 2.22: Overview of positively selected sites in human PHLDA2 with respect to the Eutherian lineage.

Above is a cartoon representation of human PHLDA2, showing sites positively selected with respect to Eutheria overlaid on functional annotation for the human PHLDA2 protein as obtained from Swiss-Prot (UniProt Consortium 2011). Each positively selected site is represented by a plus sign (+); the horizontal position of each positively selected site in the overview indicates its position within human PHLDA2, while its vertical position indicates the posterior probability (PP) that the site is undergoing positive selection, as inferred by Codeml (Yang 2007, Yang 1997). Below both the main and enlarged overview is a breakdown of the relevant Swiss-Prot annotation.



Lizard A	TCCTCGTCGTCGTCG	CCG---GGGAAGGGG
Lizard B	TCCTCGTCGTCGTCG	CCG---GGGAAGGGG
Cow	ATGAGGAGCCCGGGA	CCCACCGGCCCCGGC
Marmoset	ATGAAGTCCCCGGAC	CCCGCCAGCCCCCGC
Guinea Pig	ATGAAGACTCCCAGC	CCCACCGGGGCCGGT
Zebrafish	ATGACGGGCTCTGAT	------GACAAGAAG
Horse	ATGAAGACCCCCGGC	CCCGCCGGCCCCGGC
Human	ATGAAATCCCCCGAC	CCCGCCAGCCCCCGC
Chicken	------------GCA	CCC---CACAAGCGG
Gorilla	ATGAAATCCCCCGAC	CCCGCCAGCCCCCGC
Mouse	ATGAAGACCTCCGAC	------GGGAAAACC
Platypus	ATGGGCCGCGAGGCC	CCCGCCGGGCGGGGG
Orangutan	ATGAAGTCCCCCGAC	CCCGCCAGCCCCCGC
Chimp	ATGAAATCCCCCGAC	CCCGCCAGCCCCCGC
Rat	TTGAAGACTTCCGAT	------GGCAACAGC
Fugu	GTGACCCCGAGCCCT	------GACAAGCGC
Frog	ATGAGCGGGGCGCCT	------GATAAGAGG
	1 5	40 44

Figure 2.23: Positively selected sites in PHLDA2 with respect to the Eutherian lineage. Excerpts are shown from the positively selected site alignment for PHLDA2 with respect to the Eutherian lineage. Positively selected sites are shown in red (for foreground sequences) and blue (for background sequences), while all other sites are shown in grey. Provisional gene names (based on species of origin) for the members of the PHLDA2 gene family are shown to the left of the figure, with alignment excerpts shown to the right of these. The numbers below each alignment excerpt show the codon positions within the human PHLDA2 gene (i.e. ENSG00000181649).

(ii) SULT1E1

The PSS overview diagram in Figure 2.24 shows the Swiss-Prot annotation for sulfotransferase family 1E, estrogen-preferring, member 1 (SULT1E1), overlaid by codon sites undergoing positive selection in the Eutherian lineage as inferred using Codeml. Under the overview diagram is a breakdown of the Swiss-Prot annotation for SULT1E1 (Swiss-Prot accession P49888). The protein produced by SULT1E1 is an enzyme with three 3′-Phosphoadenosine-5′-phosphosulfate (PAPS) binding sites at positions 129, 137 and 192 (see sites annotated as `BINDING` in Figure 2.24).

Figure 2.25 shows the nucleotide alignment of SULT1E1 with its homologs from site 127 to 129, inclusive. Adjacent to the binding site at position 129 are sites 127 and 128, which are positively selected in the Eutherian lineage. Site 127 is a leucine (L) or isoleucine (I) in Eutheria, but a valine (V) in background taxa, while site 128 is a cysteine (C) in Eutherian taxa but an alanine (A) in non-Eutherian taxa. The proximity of these positively selected sites to the binding site at position 129 may indicate that these sites have been under positive selection to alter the local conformation of SULT1E1 and the presentation of the binding site. However, the substitutions at sites 127 and 128 would be considered neutral or perhaps slightly favourable, because the amino acids before and after the relevant substitutions are close to being functionally equivalent (Betts and Russell 2003), so the prospect of relaxation of selective constraint can not be ruled out in this case.

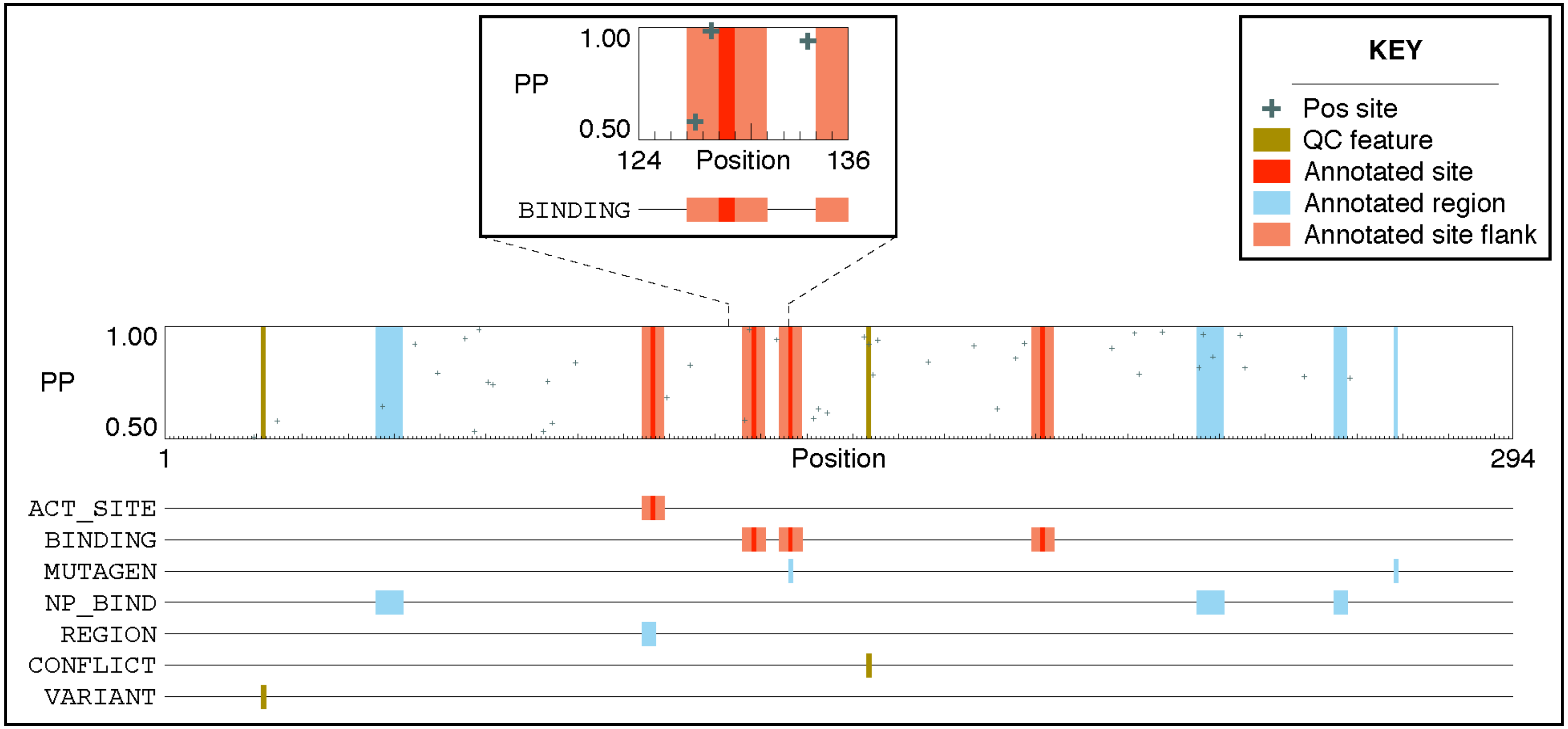
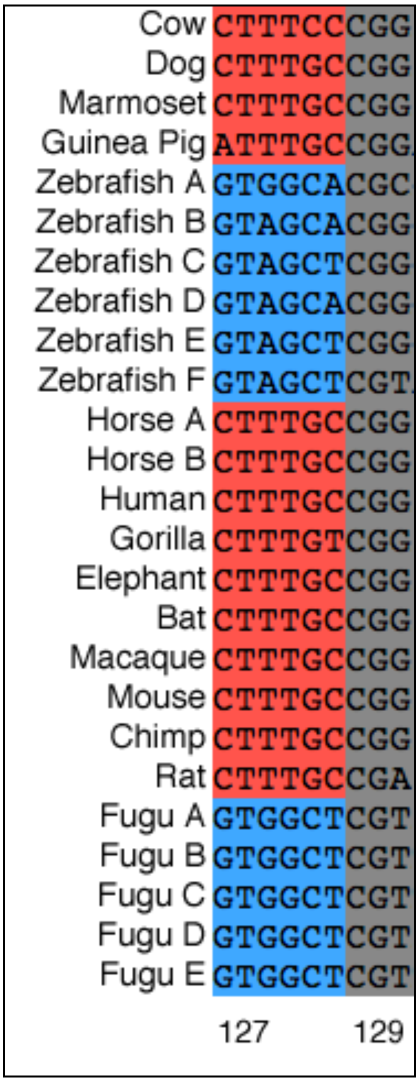


Figure 2.24: Overview of positively selected sites in SULT1E1 with respect to the Eutherian lineage.

Above is a cartoon representation of SULT1E1, showing sites positively selected with respect to Eutheria overlaid on functional annotation for the SULT1E1 protein as obtained from Swiss-Prot (UniProt Consortium 2011). Each positively selected site is represented by a plus sign (+); the horizontal position of each positively selected site in the overview indicates its position within SULT1E1, while its vertical position indicates the posterior probability (PP) that the site is undergoing positive selection, as inferred by Codeml (Yang 2007, Yang 1997). Below both the main and enlarged overview is a breakdown of the relevant Swiss-Prot annotation.



	127		129
Cow	CTT	TCC	CGG
Dog	CTT	TGC	CGG
Marmoset	CTT	TGC	CGG
Guinea Pig	ATT	TGC	CGG
Zebrafish A	GTG	GCA	CGC
Zebrafish B	GTA	GCA	CGG
Zebrafish C	GTA	GCT	CGG
Zebrafish D	GTA	GCA	CGG
Zebrafish E	GTA	GCT	CGG
Zebrafish F	GTA	GCT	CGT
Horse A	CTT	TGC	CGG
Horse B	CTT	TGC	CGG
Human	CTT	TGC	CGG
Gorilla	CTT	TGT	CGG
Elephant	CTT	TGC	CGG
Bat	CTT	TGC	CGG
Macaque	CTT	TGC	CGG
Mouse	CTT	TGC	CGG
Chimp	CTT	TGC	CGG
Rat	CTT	TGC	CGA
Fugu A	GTG	GCT	CGT
Fugu B	GTG	GCT	CGT
Fugu C	GTG	GCT	CGT
Fugu D	GTG	GCT	CGT
Fugu E	GTG	GCT	CGT

Figure 2.25: Positively selected sites in SULT1E1 with respect to the Eutherian lineage. Excerpts are shown from the PSS alignment for SULT1E1 with respect to the Eutherian lineage. Provisional gene names (based on species of origin) for the members of the SULT1E1 gene family are shown to the left of the figure, with alignment excerpt shown to the right of these. The numbers below the alignment excerpt show the codon positions within the human SULT1E1 gene (i.e. ENSG00000101193). Positively selected sites are shown in red (for foreground sequences) and blue (for background sequences), while all other sites are shown in grey.

III. Signals of Positive Selection: Alignment Error

A final point of interest —relevant to the category of erroneously predicted positively selected sites — lies in sites 21-22 of CCBP2, which were inferred as positively selected in the ancestral Eutherian lineage by Codeml under Model A. Alignment sites corresponding to codons 19-23 of CCBP2 are shown in Figure 2.14. It can be seen from this subsection of the CCBP2 alignment that the positive selection inferred for sites 21-22 in human (corresponding to three sites in some other species) is quite likely due to a misalignment of the codon 'AAT', which has been positioned in site 19 of the foreground sequences and site 20 of the background sequences. An alternative alignment of these sites might align the 'AAT' codon in site 19 across all sequences, then place a gap codon in the 3 background sequences at position 22: this would imply that a codon was inserted at site 22 in Eutheria and deleted in primates. In any case, it is quite likely that the inferred positive selection of substitutions at sites 21-22 were precipitated by an alignment error. This confirms the importance of vigilance in the construction of multiple sequence alignments, and the impact of alignment errors on downstream analyses. In any case, the specific proposed alignment error outlined here does not affect the overall result for CCBP2: selective pressure analysis was repeated with the errors corrected, and positive selection was inferred, as before, in the same sites and in the same lineages (i.e. dog, Fereuungulata, Eutheria).

2.5 Discussion

The aim of this chapter was to estimate levels of positive selection that can be detected by comparative sequence analysis of placental genes and their gene families, and to assess the extent to which such inferred positive selection can be attributed to protein functional shift. In addition to this purely biological aim, we also set out to design and implement the associated software and pipelines for the automation of large-scale selective pressure analyses.

While for some placental genes in our carefully assembled dataset it was not possible to address this question with selective pressure analysis using maximum likelihood — due either to an insufficient number of homologs or inadequate resolution of the gene family phylogeny — it was at least possible in the majority of cases (93 gene families) to conduct complete analyses across all models in Codeml (Yang 1997, Yang 2007).

Whether positive selection inferred by a selective pressure analysis *per se* can constitute evidence of functional shift is another question that remains the subject of some debate (Anisimova and Liberles 2007, Hughes 2007, Nozawa *et al.* 2009, Zhai *et al.* 2012). It is really only through the combined efforts of *in silico* prediction and *in vitro* validation that the link between positive selection and functional shift can be confirmed. Studies in the literature that have carried out such cross-disciplinary analyses are relatively rare, but provide strong and conclusive support for the relationship between positive selection and functional shift in those proteins that have been tested (Levasseur *et al.* 2006, Huang *et al.* 2012, Loughran *et al.* 2012).

In the absence of the ability to perform large-scale rational mutagenesis experiments, the consideration of functional annotation through resources such as Swiss-Prot can inform an assessment of the functional implications of positive selection on a gene of interest. However, it is important to note that this information was available for only a third of the genes that were identified in this chapter as positively selected.

Indeed, the increase in experimentally supported information on protein function in databases such as Swiss-Prot will most likely fail to keep pace with the considerable growth in genome sequence data in recent years, so this situation is unlikely to change for the foreseeable future (Koonin 2005).

In any case, those genes for which functional annotation was available illustrated the different degrees to which inferred positive selection can be attributed to functional shift: from cases in which a functional shift has been mediated by non-synonymous substitution, to those situations where the observed non-synonymous substitutions are likely to be selectively neutral or nearly so, to misaligned sites that appear to be positively selected due to erroneous inference of sitewise homology. Where positive selection is inferred, care must be taken to ensure that this is not due to a relaxation of selective constraint (Hughes 2007) or to an alignment error (Schneider *et al.* 2009).

It is worth noting that the three placental genes identified as preferentially expressed in murine placenta by Knox and Baker (2008) — adrenomedullin (ADM), cyclin E1 (CCNE1) and placenta-specific 1 (PLAC1) — were found to be positively selected: ADM and PLAC1 in the rodent lineage, CCNE1 in the Eutherian stem lineage. This is perhaps an indication that these genes have undergone a functional shift in these lineages, becoming critical to development of placenta in mice; indeed, two of the three — ADM and CCNE1 — were initially chosen as placenta-critical genes because they had been found to have lethal null phenotypes in mouse knockout studies (Li *et al.* 2006, Geng *et al.* 2003).

The results of this chapter are in broad agreement with the previous study by Hou *et al.* (2009), in which 222 genes with preferential expression in term placenta were analysed for selective pressure using Codeml. Focusing only on genes with one-to-one orthologs throughout Eutheria, Hou *et al.* (2009) found that about 27.9% of these placenta-specific genes had undergone positive selection in the ancestral Eutherian lineage (62 of 222 gene families analysed). The genes analysed in this chapter have undergone gene duplication in many cases and so the fact that the overall level of positive selection differs somewhat in this analysis is not entirely surprising.

Under the neofunctionalisation model (Lynch and Conery 2000), an increased rate of positive selection would be expected among gene families with gene duplications in the Eutherian lineage. The overall level of positive selection at the Eutherian stem detected in this chapter, across all gene families, is 44.1% (41 of 93 gene families analysed). This is significantly greater than the rate of positive selection inferred by Hou *et al.* (2009) (two-tailed Fisher's exact test, $P < 0.01$).

However, it should be noted that this does not remain the case when the dataset excludes the 11 gene families with a phylogeny sampled from a distribution with a standard deviation of split frequencies greater than 0.01 (see Section 2.3.1), all but one of which were inferred to be positively selected in the Eutherian lineage. Excluding these gene families, positive selection was found in 37.8% of cases (31 of 82 gene families for which Bayesian inference of phylogeny converged); this is still higher than the proportion of positively selected genes identified by Hou *et al.* (2009), but no longer significantly so (two-tailed Fisher's exact test, $P = 0.1225$).

Considering only those gene families with one-to-one orthologs within Eutheria from our dataset, the selective pressure analysis in this chapter identified positive selection in 16 of 44, or 36.4% of cases — somewhat greater than the rate inferred by Hou *et al.* (2009). However, the difference in rate of positive selection between our one-to-one gene families and those of the Hou study was not statistically significant (two-tailed Fisher's exact test, $P = 0.2794$). The discrepancy between the two datasets in terms of the level of positive selection detected on one-to-one orthologs is most likely due to the specific genes in each dataset differing, so we return to those 7 gene families that were in common between the two datasets to determine why this discrepancy may exist.

A comparison of the 7 specific genes analysed here and by Hou *et al.* (2009) follows a similar pattern, with a lower level of positive selection detected in our analysis — 1 gene in our analysis compared to 3 in that of Hou *et al.* (2009). Here we infer the presence of positive selection signatures for nuclear receptor coactivator 6 (NCOA6), while Hou *et al.* (2009) inferred positive selection for NCOA6 as well as ADAM metallopeptidase domain 12 (ADAM12), and kisspeptin (KISS1).

The higher number of genes inferred to be under positive selection by Hou *et al.* (2009) may be due to the multiple sequence alignment method used in that study: ClustalW (Larkin *et al.* 2007, Thompson *et al.* 1994). ClustalW has been found to perform poorly in comparison to other multiple sequence alignment methods — including PRANK, the method used in this chapter — when used to prepare a gene family alignment for use in selective pressure analysis (Fletcher and Yang 2010), see Section 2.3.1.

In any case, when compared to the conservative estimates of positive selection produced by Kosiol *et al.* (2008), the levels of positive selection inferred by both this study and by Hou *et al.* (2009) indicate that a considerable amount of positive selection has occurred in the ancestral placental mammal. With over 40% of placental genes inferred to be undergoing positive selection by this study, it is difficult to escape the conclusion that significant levels of functional shift occurred in these proteins at a time coincident with the emergence of the Eutherian lineage, and it is possible that this may have played a significant role in the emergence of this novel tissue.

Chapter 3: Benchmark of Current MicroRNA-Target Prediction Methods

3.1 Introduction

Since their tentative discovery nearly twenty years ago (Lee *et al.* 1993, Wightman *et al.* 1993) and the realisation of their broader implications (Reinhart *et al.* 2000, Pasquinelli *et al.* 2000), miRNAs have been intensively studied, resulting in an explosion in the number of known miRNAs and miRNA-target interactions (Alexiou *et al.* 2009).

The evolution of methods for miRNA-target prediction (miTP) has been intimately linked with research on miRNAs, and with our growing knowledge of miRNAs has come a concomitant growth in the number of available miRNA-target prediction methods, (Reyes-Herrera and Ficarra 2012). The large number of miTP methods, in itself, presents a challenge to non-expert users who must choose which method to use from a range of options with differing strategies and, (all too often), differing predictions (Alexiou *et al.* 2009).

Treating miRNA-target prediction as a binary classification task (i.e. classifying genes as either a target or non-target of a given miRNA), an ideal miTP method would maximise the following attributes, among others:

- **sensitivity**: proportion of actual miRNA-target genes correctly identified.
- **specificity**: proportion of non-target genes correctly rejected.
- **precision**: proportion of predicted target genes that are actual miRNA targets.

In practice, no miTP method can attain perfection in this regard, and the choice of miTP method and level of stringency will frequently require choosing which of these attributes is more important for the task at hand. For example, if a miRNA-target prediction method were to be used to identify a target gene for a miRNA of interest, and if the costs of validating miRNA-target interactions were such that only one miRNA-target interaction could be tested, then precision would take precedence over sensitivity in that case.

An intuitive strategy is to combine the predictions of two or more miTP methods, and several strategies have been proposed for the optimal integration of miRNA-target predictions from different methods (Zhang and Verbeek 2010). However, this strategy has been shown to perform poorly, with increases in specificity more than offset by decreases in sensitivity, and *vice versa*; the use of a single, accurate miRNA-target prediction method being frequently preferable (Alexiou *et al.* 2009, Ritchie *et al.* 2009, Peter 2010).

For example, in the study by Alexiou *et al.* (2009), sets of predictions from several miTP methods were checked against the results of the protein repression survey conducted by Selbach *et al.* (2008). The prediction sets for different miTP methods were also combined, in both intersection and union sets, so that the performance of method combinations could be compared with individual methods. TargetScan 5.0 was the most precise individual method, with precision and sensitivity of 50.6% and 12.34%, respectively, while miRanda was the most sensitive individual method (albeit at some cost to precision), with precision and sensitivity of 28.77% and 19.83%, respectively. While the union set of all methods surveyed achieved sensitivity of 51.66%, this was at some cost to precision, with only 25.17% of predictions corresponding to actual targets. At the other extreme, the intersection set of predictions by several methods — DIANA-microT, EIMMo, miRanda, miRBase Targets and RNA22 — achieved precision of 75%, but at a sensitivity of 0.37% (Alexiou *et al.* 2009).

This chapter will compare the performance of the 9 miRNA-target prediction programs described in Section 3.1.2. In this benchmark comparison, each miRNA-target prediction program was run on a set of miRNAs and target genes and their predictions compared to a set of known miRNA-target interactions. Because of the predominance of miRNA-target interactions mediated through target sites in the 3′ UTR (Grimson *et al.* 2007, Baek *et al.* 2008, Vergoulis *et al.* 2012) miRNA-target predictions were focused on the 3′ UTR region of the target mRNA sequences for the purposes of this benchmark study.

3.1.1 MicroRNA-Target Prediction Benchmark Studies

Several studies have sought to compare the performance of contemporary miRNA-target prediction methods in terms of their ability to identify genuine miRNA-target interactions (Sethupathy *et al.* 2006, Baek *et al.* 2008, Selbach *et al.* 2008, Alexiou *et al.* 2009). Typically, these studies compare miRNA-target predictions to a set of experimentally validated miRNA-target interactions. Not all studies take the same approach to comparison of miRNA-target prediction methods, and so the results of these studies are not necessarily directly comparable. Four such studies are outlined below.

The first such benchmark was performed by Sethupathy *et al.* (2006), which compared five methods available at that time: DIANA-microT (Kiriakidou *et al.* 2004), miRanda (Enright *et al.* 2003), PicTar (Krek *et al.* 2005), TargetScan (Lewis *et al.* 2003) and TargetScanS — a revised and simplified form of TargetScan that was the foundation for TargetScan as it is now known (Lewis *et al.* 2005). Using a combination of *de novo* and pre-compiled miRNA-target predictions, Sethupathy *et al.* (2006) compared these with 84 validated miRNA-target interactions from the nascent TarBase (Sethupathy *et al.* 2006). TargetScanS, PicTar and miRanda — recently developed methods at that time — achieved sensitivity levels of 47.6%, 47.6% and 48.8%, respectively. This was an order of magnitude more sensitive than DIANA-microT and the original TargetScan, which had sensitivity levels of 9.5% and 20.8%, respectively. Although it had the highest sensitivity, miRanda also made the largest number of miRNA-target predictions (i.e. 18,289); while TargetScanS and PicTar achieved a comparable level of sensitivity to miRanda despite making relatively fewer predictions (i.e. 10,351 and 11,259, respectively) (Sethupathy *et al.* 2006).

Baek *et al.* (2008) used stable isotope labelling with amino acids in cell culture (SILAC) (Ong *et al.* 2002) to measure the effect on protein expression caused by the introduction of miR-124, miR-1 or miR-181 into HeLa cells in separate experiments, and a gene knockout of mir-223 in mouse neutrophils. Using the results of these experiments, Baek *et al.* (2008) compared sets of miRNA-target predictions against the observed repressive effects of miRNA-target interactions on protein expression in their SILAC experiment. The prediction sets compared were those of: miRanda (Betel *et al.* 2008, John *et al.* 2004, Enright *et al.* 2003), PicTar (Lall *et al.* 2006, Krek *et al.* 2005), PITA (Kertesz *et al.* 2007), miRBase Targets, now known as MicroCosm Targets (Griffiths-Jones *et al.* 2008) and TargetScan, which had initially been known as TargetScanS to distinguish it from earlier iterations of the program (Grimson *et al.* 2007, Lewis *et al.* 2005). Unlike Sethupathy *et al.* (2006), which assessed the performance of miTP methods as binary classifiers of miRNA targets and non-targets, this study assessed the degree to which genes predicted as miRNA targets were downregulated at protein level.

TargetScan and PicTar were noted as having the best performance in this benchmark: on knockout of mir-223, the observed mean $\log_2$ fold change of protein expression was 0.32 for the protein products of genes predicted to be miRNA targets by TargetScan, 0.27 for PicTar, and 0.15 for the next best performing method, miRanda. TargetScan was the only method for which the ranking of predictions was significantly correlated with protein downregulation (Mann-Whitney *U* test, $P < 0.01$), with a lower context score indicating greater protein repression. Nevertheless, with two thirds of targets predicted by TargetScan and PicTar being false-positive predictions whose expression remained unresponsive to mir-223 knockout, these methods could be estimated to have a precision of about one third (Baek *et al.* 2008).

A similar survey of protein repression was performed by Selbach *et al.* (2008), involving the transfection and overexpression of let-7b, miR-1, miR-16, miR-30a or miR-155 in separate cultures of HeLa cells. Pulsed SILAC (pSILAC) was used to measure the effect of overexpression of each individual miRNA on protein expression. As part of their study, Selbach *et al.* (2008) tested the correlation of miRNA-target predictions with their miRNA repression data for the following pre-compiled miTP datasets: DIANA-microT v3.0 (Maragkakis *et al.* 2009a), TargetScan, then still known as TargetScanS (Grimson *et al.* 2007, Lewis *et al.* 2005), PicTar (Lall *et al.* 2006, Krek *et al.* 2005), RNA22 (Miranda *et al.* 2006), PITA (Kertesz *et al.* 2007), miRanda (Betel *et al.* 2008, John *et al.* 2004, Enright *et al.* 2003) and miRBase Targets (Griffiths-Jones *et al.* 2008).

For each miTP method, Selbach *et al.* (2008) calculated the proportion of predicted targets for which protein fold change exceeded -0.1 on a $\log_2$ scale (i.e. approximately 10% or greater reduction in expression from basal levels). About 27% of proteins in the survey by Selbach *et al.* (2008) were downregulated by at least this amount, so a miTP method that predicts miRNA-target interactions at random would be expected to identify ~27% of the downregulated proteins simply by chance. In addition, Selbach *et al.* (2008) considered the performance of a simple seed match heuristic — by which the presence in the target mRNA of a site complementary to the miRNA seed constitutes a miRNA-target prediction — and found that 44% of the downregulated proteins had such a seed match in their mRNA sequence. TargetScan, PicTar and DIANA-microT had the greatest precision in identifying genes whose protein product was downregulated, with precision levels of 61.3%, 61.4% and 66%, respectively. Other miTP datasets had lower precision, ranging in performance between those of the random predictor and seed match heuristic (i.e. 27% and 44%, respectively). The performance of DIANA-microT v3.0 was particularly noteworthy, since its precision exceeded that of TargetScan and PicTar while making 294 miRNA-target predictions — less than half the number predicted by either TargetScan (622 predictions) or PicTar (629 predictions) (Selbach *et al.* 2008).

Alexiou *et al.* (2009) performed a benchmark that made use of validated miRNA-target interactions from three sources: the aforementioned pSILAC study by Selbach *et al.* (2008), a microarray study of the effects on mRNA expression of transfection of miR-1 and miR-124 in separate cultures of HeLa cells (Lim *et al.* 2005) and a set of over 1300 validated miRNA-target interactions from the most recent update of TarBase at that time, TarBase v5.0 (Papadopoulos *et al.* 2009).

This study compared a broader range of methods than previous benchmarks: DIANA-microT v3.0 (Maragkakis *et al.* 2009a), TargetScan v5.0 (Friedman *et al.* 2009, Grimson *et al.* 2007, Lewis *et al.* 2005), EIMMo (Gaidatzis *et al.* 2007), PicTar (Lall *et al.* 2006, Krek *et al.* 2005), miRanda (Betel *et al.* 2008, John *et al.* 2004, Enright *et al.* 2003), PITA (Kertesz *et al.* 2007), RNA22 (Miranda *et al.* 2006) and miRBase Targets (Griffiths-Jones *et al.* 2008). As with Selbach *et al.* (2008), Alexiou *et al.* (2009) compared these methods with a simple seed match heuristic.

For all three validated miTI datasets, four miTP methods performed comparably better: DIANA-microT v3.0, TargetScan v5.0, EIMMo and PicTar. In the comparison based on the proteomic data of Selbach *et al.* (2008), these methods had estimated precision ranging from 48-51%, while their estimated sensitivity ranged from 8-12%. (For comparison, a simple seed match heuristic attained precision of 30% and sensitivity of 45%.) The remaining four methods had precision about equal to the seed match heuristic, although miRanda was notable among these for attaining the highest sensitivity of all the programs tested (i.e. 20%), while also attaining a reasonably high level of precision (i.e. 29%) (Alexiou *et al.* 2009).

From these studies, it would be reasonable to select as a miRNA-target prediction method one of TargetScan, PicTar, DIANA-microT v3.0, EIMMo or possibly miRanda. However, miRNA-target prediction software continues to be actively developed, including: SVMicrO (Liu *et al.* 2010), TargetSpy (Sturm *et al.* 2010), miREE (Reyes-Herrera *et al.* 2011), TargetMiner (Bandyopadhyay and Mitra 2009) and its successor MultiMiTar (Mitra and Bandyopadhyay 2011), PACCMIT (Marín and Vaníček 2012) (Marín and Vaníček 2011) and the miRmap Python package (Vejnar and Zdobnov 2012). In addition, new versions of three existing methods have been published since the benchmark studies discussed above: miRanda v3.3 (Betel *et al.* 2010), TargetScan 6 (Garcia *et al.* 2011) and DIANA-microT v4.0 (Reczko *et al.* 2012, Reczko *et al.* 2011). A comparison of the miTP methods now available would therefore be timely and useful for those seeking to identify candidate miRNA target genes *in silico*.

Such a comparison would nevertheless be challenging, as it is difficult to compare, in a fair way, miRNA-target prediction methods that are based on different principles (Sturm *et al.* 2010), have different input data (Sethupathy *et al.* 2006) and different miTP output (Betel *et al.* 2010).

Because of their reliance on pre-compiled miRNA-target predictions that would inevitably be based on different input data, all of the benchmarks outlined above could potentially have been affected by data heterogeneity (Sethupathy *et al.* 2006, Baek *et al.* 2008, Selbach *et al.* 2008, Alexiou *et al.* 2009). Running each method on consistent data would mitigate this issue and could allow for a more consistent comparison of different miTP methods. A standalone version of the miTP program would be required for a miRNA-target prediction method to be compared in this way, so that the program can be run with the test data on a local computer. Therefore, in this chapter we compare the miRNA-target prediction performance of only those 9 currently available methods that have a standalone version. These are described in detail in the following section.

3.1.2 Standalone MicroRNA-Target Prediction Methods

There are 9 currently available miRNA-target prediction (miTP) programs with a standalone program: (I) RNAhybrid, (II) MicroTar, (III) PITA, (IV) miRanda, (V) TargetSpy, (VI) Hitsensor, (VII) MultiMiTar, (VIII) TargetScan, and (IX) miRmap. The features of each are described in turn below and summarised in Table 3.1.

I. RNAhybrid

RNAhybrid extends the RNA secondary structure algorithm proposed by Zuker and Stiegler (1981), obtaining the optimal energy of hybridisation between a miRNA and its potential target sequence (Krüger and Rehmsmeier 2006, Rehmsmeier *et al.* 2004). Because of the short sequence length of the miRNA relative to its target and RNAhybrid's constraint on the formation of intramolecular loops in either miRNA or target mRNA, the execution time of RNAhybrid is estimated to be linearly proportional to the target RNA sequence length. The hybridisation energy calculation can also be constrained so that the miRNA seed sites are bound to the corresponding target site on the target. RNAhybrid outputs two metrics for each predicted miTI (Krüger and Rehmsmeier 2006, Rehmsmeier *et al.* 2004):

- the RNAhybrid **Energy** metric reflects the minimum free energy (MFE) of hybridisation between the miRNA and its target.

- the **p-value** of the miTI: the probability that the given miRNA-target interaction would have that MFE or better by chance.

Table 3.1 Legend:

Shown on page 190 overleaf is a table of the key features used by the 9 miTP methods being compared in this chapter. A plus sign (i.e. +) denotes that a feature is incorporated into a miTP method; if it is followed by a question mark (i.e. +?), then that feature is optional for the given miRNA-target prediction program. Three miRNA binding features are shown: the **miRNA** column indicates whether overall miRNA-target binding information is used, the **Seed** column indicates whether seed binding information is used, and the **3′** column indicates whether binding of the 3′ region of the miRNA to its target sequence is a feature of the miTP method. Two thermodynamics features are shown: the **Hybridisation** column shows whether the give method estimates the free energy of hybridisation of the miRNA to its target, while the **Accessibility** column indicates whether thermodynamic accessibility of the target site is accounted for. (Note that for TargetScan, assessment of hybridisation is limited to the seed sequence.) Six target site features are included: use of conservation to identify conserved functional miRNA-target sites (**Conservation**), assessment of sequence composition (e.g. AU content) in the target site (**Composition**), position of miRNA-target sites relative to each other (**Relative Position**), position of miRNA-target sites within the target sequence (**Absolute Position**), the proportion of base pair bonds between the miRNA and its target site (**Compactness**), and the probability of chance occurrence of a given target site sequence (**Probability**). The rightmost column shows a feature used only by TargetScan: target abundance, the estimated quantity of competing target sites in the transcriptome (**TA**).

Table 3.1: Key features of microRNA-target prediction benchmark methods.

miTP Method	miRNA Binding			Thermodynamics		Target Site						TA
	miRNA	Seed	3′	Hybridisation	Accessibility	Conservation	Composition	Relative Position	Absolute Position	Compactness	Probability	
RNAhybrid		+?		+								
MicroTar		+		+								
PITA		+		+	+							
miRanda	+	+?		+								
TargetSpy	+	+?	+				+			+		
Hitsensor		+?	+				+	+	+			
MultiMiTar		+					+					
TargetScan		+	+	+		+	+		+			+
miRmap		+	+	+	+	+	+		+		+	

II. MicroTar

MicroTar was developed by Thadani and Tammi (2006), and is similar to RNAhybrid, in that it calculates the binding energy of the miRNA and its target sequence using the algorithm described by Zuker and Stiegler (1981). However, MicroTar takes a more thorough approach — the secondary structure and MFE of the entire target mRNA is first estimated using RNAfold from the ViennaRNA package (Gruber *et al.* 2008, Hofacker 2003, Hofacker *et al.* 1994); the 3′ UTR of the target mRNA is then checked for miRNA seed matches; and for each seed match site, the secondary structure and MFE of the mRNA and its bound miRNA are estimated using the ViennaRNA cofold program. Predicted miTIs are output with an associated negative normalized MFE of the bound miRNA-target pair (Thadani and Tammi 2006).

III. PITA

PITA focuses on the miRNA-target accessibility within the folded 3′ UTR (Kertesz *et al.* 2007). Sequences in the target 3′ UTR that are complementary to the miRNA seed site are taken to be miRNA-target sites and are each given an accessibility score $\Delta\Delta G$ (see Equation 3.1).

$$\Delta\Delta G = \Delta G_{duplex} - \Delta G_{open}$$

Equation 3.1: Accessibility score in PITA.

The accessibility score ($\Delta\Delta G$) used by PITA is shown in terms of the free energy gained by miRNA-target hybridisation (ΔG_{duplex}) and the energetic cost of unpairing the miRNA-target site nucleotides (ΔG_{open}). Both are calculated by PITA using the ViennaRNA package (Gruber *et al.* 2008, Hofacker 2003, Hofacker *et al.* 1994).

Where there is more than one target site for a given miRNA-target pair, the accessibility scores are integrated into one target score as in Equation 3.2.

$$TS = -\log\left(\sum_{i=1}^{n} e^{-\Delta\Delta G_i}\right)$$

Equation 3.2: Target score in PITA.

The PITA target score (*TS*) is shown in terms of the individual ΔΔ*G* values of *n* putative miRNA-target sites for the given miRNA-target pair.

The lower the value of the target score for a given miTI is, the more accessible the miRNA-target sites are considered to be for the given gene. PITA also includes the coding DNA sequence and target site flanking sequence in its energy calculations for each miRNA-target site, if these are specified by the user (Kertesz *et al.* 2007).

IV. miRanda

miRanda calculates the optimal sequence complementarity between the miRNA and the target sequence using a modified form of the Smith-Waterman algorithm (Smith and Waterman 1981), which gives greater weight to complementarity in the region of the target site complementary to the miRNA seed, allows for G-U 'wobble' pairs, and applies empirical rules constraining mismatches between the miRNA and its target (Betel *et al.* 2010, Betel *et al.* 2008, John *et al.* 2004, Enright *et al.* 2003). For target regions of high sequence complementarity, the ViennaRNA package (Gruber *et al.* 2008, Hofacker 2003, Hofacker *et al.* 1994) is then used to calculate the minimum free energy (MFE) of formation of the miRNA-target duplex. miRanda therefore outputs two metrics for each miTI:

- the miRanda **Score** metric reflects the weighted complementarity of the miRNA and target sequences, estimated according to its dynamic programming algorithm.

- the miRanda **Energy** metric reflects the MFE of formation of the hybridised miRNA-target duplex.

Note that miRanda miRNA-target predictions are available to download from the website microRNA.org (Betel *et al.* 2008). Here the miRNA-target predictions are ranked using mirSVR (Betel *et al.* 2010) — a support vector regression trained using the miRNA transfection experiments conducted by Grimson *et al.* (2007). mirSVR ranks miTPs by their estimated downregulation level, and was reported to perform well by Betel *et al.* (2010): mirSVR-ranked predictions — along with those of TargetScan's standard Context score (Grimson *et al.* 2007), as opposed to the Context+ score (Garcia *et al.* 2011) — were checked against the observed downregulation of gene expression in a miRNA transfection experiment by Linsley et al. (2007). In a ROC analysis of their relative performance, mirSVR had a higher area under the ROC curve in 21 out of 25 test sets, indicating a statistically significant improvement on the standard Context score (signed rank test, $P < 0.002$). However, mirSVR is not currently available either as standalone software or integrated in miRanda.

V. TargetSpy

TargetSpy was developed by Sturm *et al.* (2010) with machine learning methods, using the Ago HITS-CLIP data set of Chi *et al.* (2009) as a training data set. The learning scheme MultiBoost (Webb 2000) — as implemented in WEKA (Witten and Frank 2005) — was used with a wide range of miTI features including the extent of miRNA-target binding, G:U base pairing, duplex bulges, position-specific features, compositional features, 'compactness', and the accessibility of the target site. After training, the set of features were ranked using the ReliefF algorithm (Kononenko 1994) and the most discriminative features were selected by Correlation-Based Feature Selection. (Hall and Smith 1997). A novel feature introduced by these authors is **compactness**, which is defined as the mean of two ratios: the number of base pairings divided by the length of the miRNA, and the number of base pairings divided by the length of the target site.

TargetSpy uses RNAduplex from the ViennaRNA package (Gruber *et al.* 2008, Hofacker 2003, Hofacker *et al.* 1994) to identify all possible duplex structures for the miRNA-target pair. Regions of the target that have a large number of energetically favourable miRNA-target duplexes are selected as candidate zones.

TargetSpy outputs a ranked list of candidate miRNA-target sites, each with two metrics:

- the TargetSpy **Score** metric, which takes account of discriminative features found by machine learning methods to be associated with known miRNA-target sites.

- the TargetSpy **Energy** metric: the estimated Gibbs free energy of hybridisation of the miRNA-target duplex (i.e. ΔG_{duplex}).

VI. Hitsensor

The Hitsensor algorithm was developed by Zheng and Zhang (2010). This program first uses a modified form of the Smith-Waterman algorithm (Smith and Waterman 1981) that, similarly to the dynamic programming algorithm incorporated in miRanda, finds regions of high complementarity in the target sequence, allowing for G-U pairs and mismatches. These highly complementary regions are then scored according to five specific features of known miRNA-target sites: local AU content, base pairing of miRNA nucleotides 12–17, optimal intersite spacing, miRNA-target site position in the 3′ UTR sequence, and seed pairing in the 3′ part of the miRNA. In this benchmark, for each miRNA-target gene with one or more miRNA-target sites, the Hitsensor score is taken from that of the highest scoring miRNA-target site. For example, Hitsensor predicts 5 miRNA-target sites of cel-lin-4 in the 3′ UTR of the *C. elegans* gene lin-14, with Hitsensor scores ranging from 111.4 to 291.3. In this benchmark, the highest scoring miRNA-target site is taken as representative for a given miRNA-target pair; in the example given, this would be the miRNA-target site with Hitsensor score 291.3.

VII. MultiMiTar

MultiMiTar (Mitra and Bandyopadhyay 2011, Bandyopadhyay and Mitra 2009) is also based on machine learning — in this case using an SVM integrated with 'a simulated annealing based multi-objective optimization algorithm' (AMOSA) (Bandyopadhyay *et al.* 2008). This is an integration of a support vector machine (SVM) used as a classifier, and a simulated annealing based multi-objective tool, AMOSA, used for feature selection. Features assessed by the SVM include the presence of a miRNA seed match and local AU content, frequency of single nucleotides and dinucleotides in the seed match and in the remainder of the target site, miRNA-target base-pairing in the seed region and dinucleotide base-pairing in the seed region. MultiMiTar outputs a result for each miRNA-target pair with an associated score — known as a decision value — taking account of all predicted miRNA-target sites in that miRNA-target pair. Predictions of miRNA-target interactions with a higher decision value may be considered to be more reliable. For example, the MultiMiTar decision value for a miTI involving hsa-miR-16-5p and cyclin T2 (CCNT2) is 0.869, indicating that such an interaction is highly probable.

VIII. TargetScan

TargetScan is a package of Perl scripts that identify miRNA-target sites by searching for discriminative features of experimentally verified miTIs (Garcia *et al.* 2011, Friedman *et al.* 2009, Grimson *et al.* 2007, Lewis *et al.* 2005). The standard TargetScan script searches for canonical miRNA seed matches in the target sequence. The set of seed matches can then be further analysed by the complementary TargetScan context and conservation scripts.

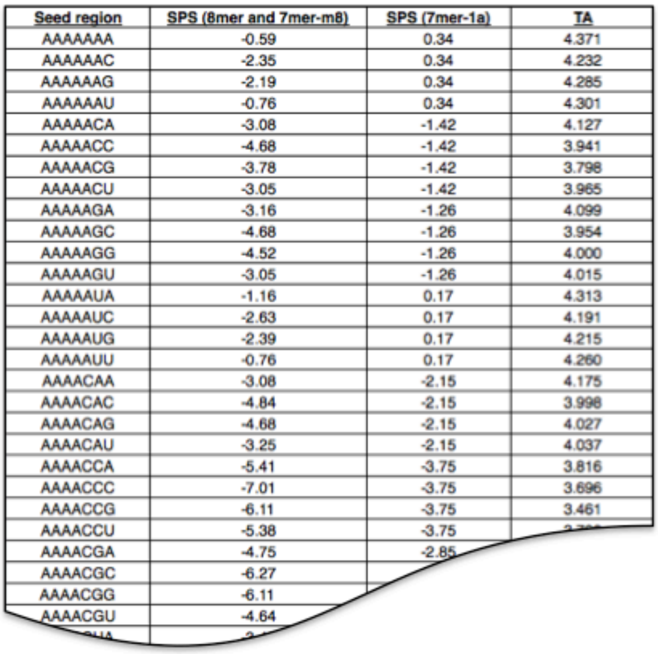
The TargetScan context script checks for contextual features of the target site that are associated with functional sites: seed match site type, base pairing at the 3′ end of the miRNA, local AU content, position of the target site in the target sequence (Grimson *et al.* 2007). Garcia *et al.* (2011) added two further context features: the seed-pairing stability and the target abundance within the genome under study. The Context+ score returned by TargetScan 6.2 takes account of all these features and tends to decrease from zero, with a lower Context+ score being indicative of a stronger miRNA-target prediction; the Context+ score for a miRNA and its target gene is obtained by summing all negative Context+ scores for that miRNA-target pair (Garcia *et al.* 2011, Grimson *et al.* 2007). For example, TargetScan predicts a miRNA-target interaction between hsa-let-7a and LIN28B, involving 5 miRNA-target sites. The individual miRNA-target sites have the following Context+ scores: -0.442, -0.184, -0.152, -0.075 and -0.056; the total Context+ score for hsa-let-7a and LIN28B is the sum of these: -0.909.

The TargetScan conservation script compares putative target sites across multiple orthologous target sequences, and returns a probability of preferentially conserved targeting. Rather than require conservation throughout all species, the TargetScan conservation script estimates the branch length across which a predicted target site is conserved, on a phylogeny of 28 species. This "TargetScan phylogeny" was created by Friedman *et al.* (2009) using the topology given by Miller *et al.* (2007), estimating branch lengths from comparison of the orthologous 3′ UTRs aligned by the same study (Miller *et al.* 2007). In assessing the conservation of a miRNA-target site, the background conservation of the target 3′ UTR is accounted for, so that, other things being equal, a conserved target site is scored more highly if it's located in a poorly conserved 3′ UTR (Friedman *et al.* 2009).

A slightly modified version of TargetScan was used in the analysis in this chapter and in Chapter 4, such that the following two changes were made to TargetScan by this author:

- in the conservation script, the reference species (from which branch lengths are calculated) can be set at the command line.

- in the context script, a user-specified TA/SPS file can be set at the command line, and if this is done, variables within the context script tracking the minimum and maximum TA values are updated using the TA values in the specified TA/SPS file.



Seed region	SPS (8mer and 7mer-m8)	SPS (7mer-1a)	TA
AAAAAAA	-0.59	0.34	4.371
AAAAAAC	-2.35	0.34	4.232
AAAAAAG	-2.19	0.34	4.285
AAAAAAU	-0.76	0.34	4.301
AAAAACA	-3.08	-1.42	4.127
AAAAACC	-4.68	-1.42	3.941
AAAAACG	-3.78	-1.42	3.798
AAAAACU	-3.05	-1.42	3.965
AAAAAGA	-3.16	-1.26	4.099
AAAAAGC	-4.68	-1.26	3.954
AAAAAGG	-4.52	-1.26	4.000
AAAAAGU	-3.05	-1.26	4.015
AAAAAUA	-1.16	0.17	4.313
AAAAAUC	-2.63	0.17	4.191
AAAAAUG	-2.39	0.17	4.215
AAAAAUU	-0.76	0.17	4.260
AAAACAA	-3.08	-2.15	4.175
AAAACAC	-4.84	-2.15	3.998
AAAACAG	-4.68	-2.15	4.027
AAAACAU	-3.25	-2.15	4.037
AAAACCA	-5.41	-3.75	3.816
AAAACCC	-7.01	-3.75	3.696
AAAACCG	-6.11	-3.75	3.461
AAAACCU	-5.38	-3.75	[illegible]
AAAACGA	-4.75	-2.85	
AAAACGC	-6.27		
AAAACGG	-6.11		
AAAACGU	-4.64		
[illegible]	[illegible]		

Figure 3.1: TargetScan TA/SPS file sample.

Shown is a sample of a TA/SPS file used by TargetScan to calculate the Context+ score (Garcia *et al.* 2011). Four columns are included: a **Seed Region** column that includes every possible 7mer seed region, an **SPS (8mer and 7mer-m8)** column that contains the estimated seed-pairing stability (SPS) for seed match sites of type 8mer and 7mer-m8, an **SPS (7mer-1a)** column that shows SPS estimates for seed match sites of type 7mer-1a, and a **TA** column showing the estimated target abundance for the given miRNA seed region.

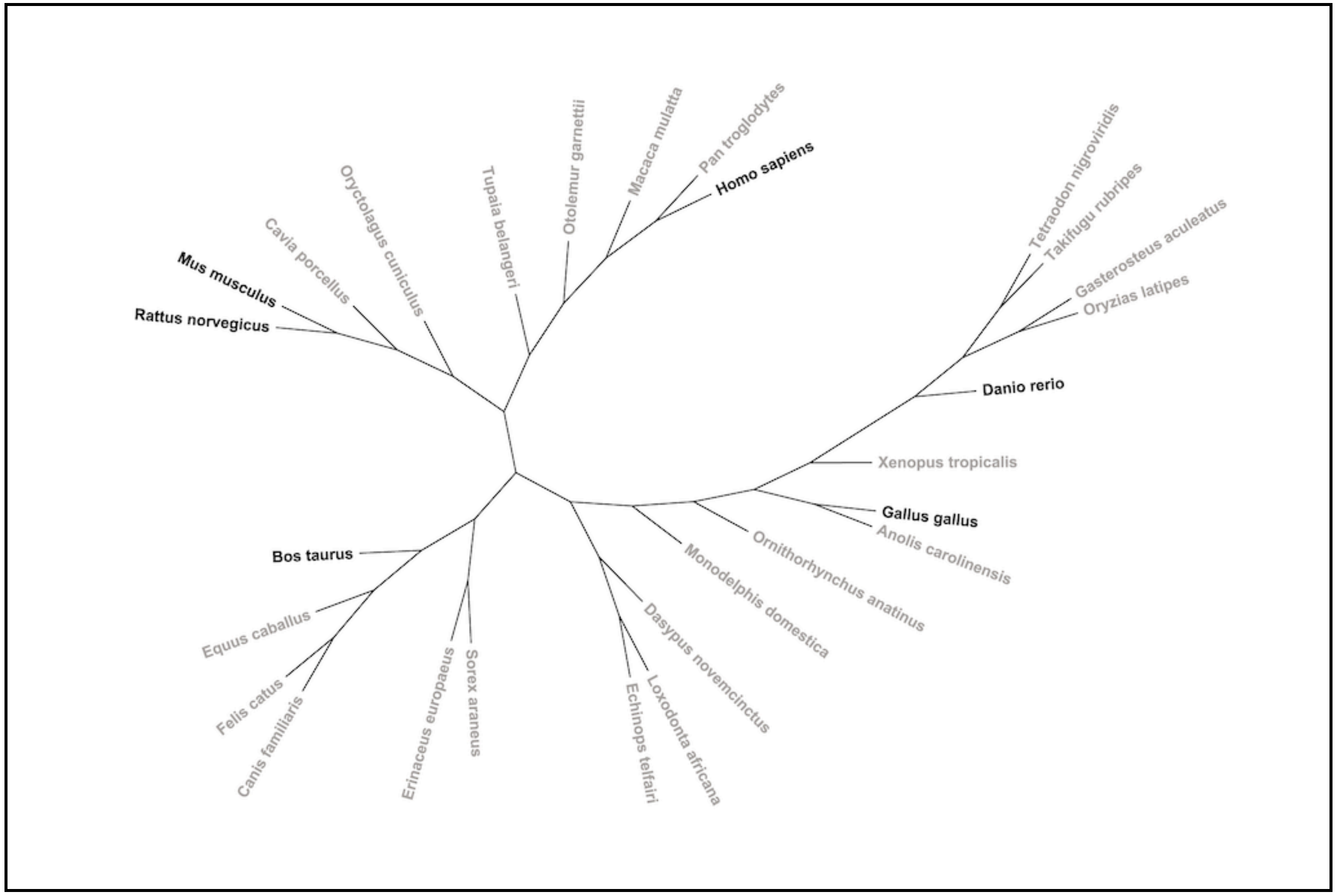


Figure 3.2: TargetScan phylogeny.

Shown is the "TargetScan phylogeny" created by Friedman *et al.* (2009) based on the topology and aligned 3′ UTR sequences of Miller *et al.* (2007). TargetScan infers preferentially conserved targeting based on the estimated branch length on this phylogeny that a miRNA-target site is conserved (Friedman *et al.* 2009). The 6 species in bold are the benchmark test species for this chapter.

IX. miRmap

miRmap is an open-source Python library that integrates features previously used only by different miTP methods (Vejnar and Zdobnov 2012). Using the ViennaRNA package (Gruber *et al.* 2008, Hofacker 2003, Hofacker *et al.* 1994), miRmap assesses the thermodynamic stability of the miRNA-target duplex (i.e. Δ*G*), taking account of the accessibility of the target site (i.e. ΔΔ*G*) in a manner similar to PITA (see above). It estimates the probability of chance occurrence of the proposed target site sequence (e.g. for miRNA hsa-let-7b-5p, with seed sequence 'GAGGUAG', this might be 'CUACCUC'), using an approximation following the binomial distribution as in Marín and Vaníček (2011), as well as an exact probability calculated in the manner described by Nuel *et al.* (2010). It estimates target site conservation by summing over the branch lengths of a species phylogeny (Stark *et al.* 2007, Friedman *et al.* 2009), using DendroPy (Sukumaran and Holder 2010) for tree manipulations; it can also test for conservation due to negative selection using the Siepel, Pollard and Haussler (SPH) test (Pollard *et al.* 2010) as implemented in the PhyloP program of the PHAST suite (Hubisz *et al.* 2011). In addition, it takes account of the miRNA-target site features outlined by Grimson *et al.* (2007): seed match type, local AU content, proximity of target sites to either end of the 3′ UTR sequence and binding of miRNA bases 12-17 to the putative target site.

These features were incorporated into miRmap and tested against a number of validated miTI datasets: Grimson *et al.* (2007), Linsley *et al.* (2007), Selbach *et al.* (2008), and Hendrickson *et al.* (2009). The relative importance of these features were computed by the CAR method — with the strained acronym Correlation Adjusted marginal coRrelation — which helps determine which variables of a model account most for the observed data (Zuber and Strimmer 2011). For each miRNA-target prediction, miRmap outputs a score combining these features. Like the TargetScan context score, this tends to decrease from zero so that lower scores are considered to be better.

3.2 Materials

A set of validated miRNA-target interactions (miTIs) was assembled using the TarBase 6.0 database. For each of the miRNAs with at least one miTI validation, the miRNA sequence was obtained from miRBase release 18[7] (Kozomara and Griffiths-Jones 2011, Griffiths-Jones 2004, Griffiths-Jones 2006, Griffiths-Jones *et al.* 2008). For each benchmark test species, all annotated 3′ UTR sequences were downloaded from the Ensembl server (Flicek *et al.* 2012) through the BioMart interface (Smedley *et al.* 2009); from these, a representative 3′ UTR was selected for each gene, against which miRNA-target prediction was to be performed. All scripts used during this process can be accessed from the following location in the electronic appendix: Appendix / home / code / scripts.html.

3.2.1 Validated MicroRNA-Target Interaction Data Assembly

In order to compare the performance of different miRNA-target prediction methods, it is necessary to have validated miRNA-target interactions with which to compare their predictions. This involved several stages as follows: (I) identifying a good quality source of validated miRNA-target interaction data, (II) downloading validated miRNA-target interaction data, (III) curating that data to eliminate redundancy and remove any errors, (IV) preparing the validated miRNA-target interaction data by resolving miRNA-target validations and interactions, (V) filtering the set of validated miTIs for quality purposes, and (VI) dividing the set of miRNA-target validations into two different data subsets more appropriate to their use in the subsequent miRNA-target prediction benchmark process. This process is outlined below, and summarised in the bioinformatics pipeline in Figure 3.3.

[7] Please note that throughout this thesis, unless otherwise specified, Ensembl data has been obtained from Ensembl release 65 and miRBase data has been obtained from miRBase release 18.

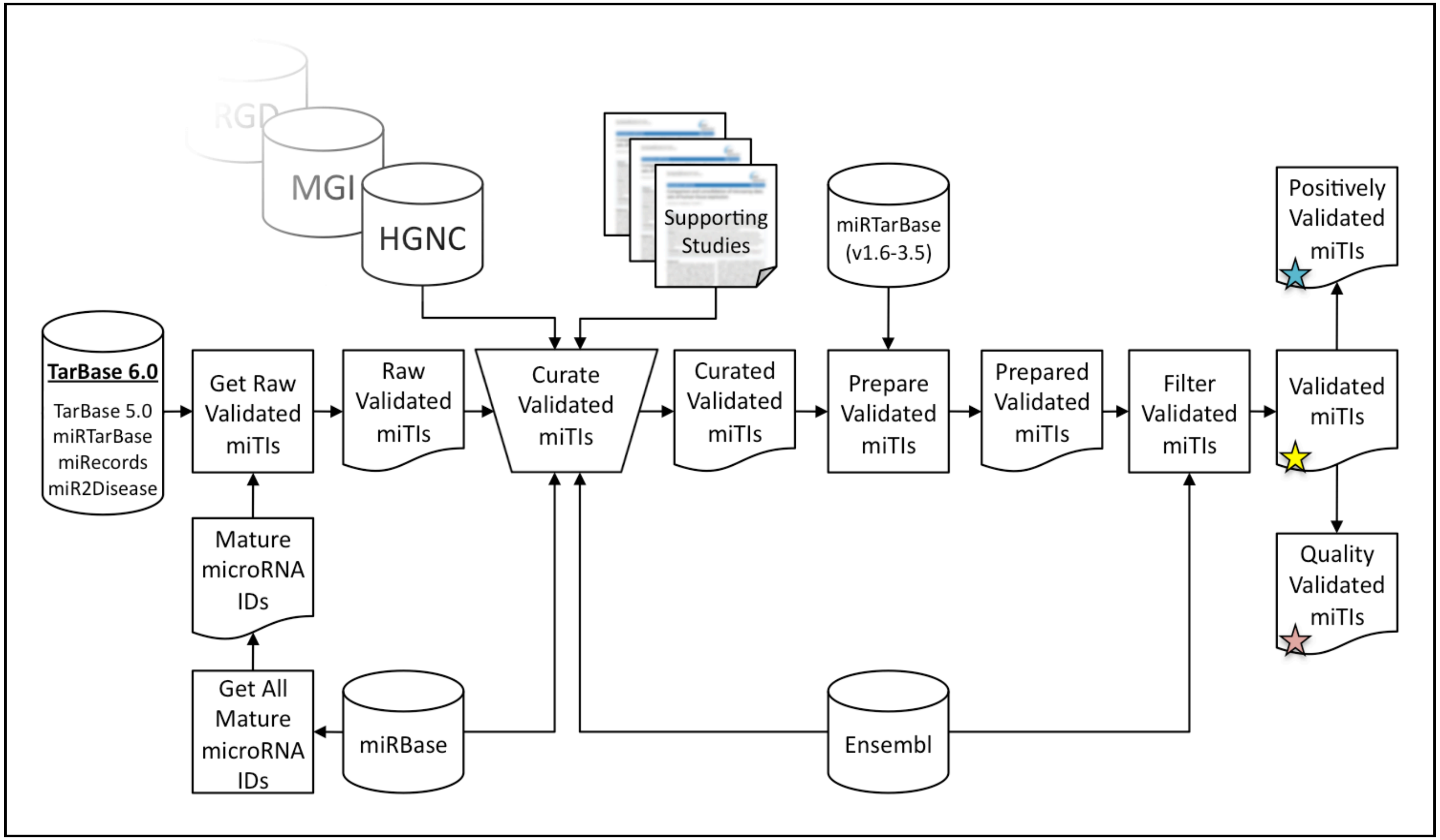


Figure 3.3: Validated microRNA-target interaction data assembly pipeline.

Shown is a bioinformatics pipeline depicting the steps involved in validated miRNA-target interaction data assembly. Arrows indicate the direction of process flow. For information on symbols used, see the pipeline key in Figure 2.2.

I. Selecting a Validated MicroRNA-Target Interaction Database

Several databases of validated miRNA-target interactions were reviewed. These were: miR2Disease (Jiang *et al.* 2009), miRecords (Xiao *et al.* 2009), miRTarBase (Hsu *et al.* 2010), miRWalk (Dweep *et al.* 2011) and TarBase 6.0 (Vergoulis *et al.* 2012). These were reviewed with regard to the breadth of species coverage within vertebrates and the number and quality of validated miRNA-target interactions. From this review, it was apparent that TarBase 6.0 was the best currently available source of miRNA-target interactions, so this database was selected as a source of validated miRNA-target interactions (see Table 3.2). TarBase 6.0 has been recently updated and includes the highest number of validated miTIs across the broadest range of vertebrate species (Vergoulis *et al.* 2012). TarBase 6.0 also integrates entries from three other databases: miR2Disease (Jiang *et al.* 2009), miRecords (Xiao *et al.* 2009) and miRTarBase (Hsu *et al.* 2010).

II. Downloading Validated MicroRNA-Target Interaction Data

Validated miRNA-target interaction data is available from TarBase 6.0 for a total of 9 vertebrate species (Vergoulis *et al.* 2012). The vertebrate species given as available in TarBase 6.0 are: Human (*Homo sapiens*), Mouse (*Mus musculus*), Rat (*Rattus norvegicus*), Cow (*Bos taurus*), Sheep (*Ovis aries*), Chicken (*Gallus gallus*), African Clawed Frog (*Xenopus laevis*), Western Clawed Frog (*Xenopus tropicalis*) and Zebrafish (*Danio rerio*). Each validated miRNA-target interaction (miTI) is associated with one or more supporting studies, and validation data are given both for the miTI and for each supporting study. Validations are grouped by experiment type (see Table 3.3). Validation outcomes are given as either positive or negative. A positive validation indicates that the miTI was validated by the given experiment, while a negative validation indicates that the miTI was not validated by the given experiment. With respect to the integrated external databases, both miRecords and miR2Disease report only positive validations, while miRTarBase reports both positive and negative miTI validations.

Table 3.2: Databases of validated miRNA-target interactions.

Database Name	Number of Validations	Availability	References
miRecords	6653	mirecords.umn.edu/miRecords	Xiao *et al.* (2009)
miR2Disease	809	www.mir2disease.org	Jiang *et al.* (2009)
miRTarBase	4867	mirtarbase.mbc.nctu.edu.tw	Hsu *et al.* (2010)
miRWalk	Unknown	www.umm.uni-heidelberg.de/apps/zmf/mirwalk/	Dweep *et al.*(2011)
TarBase 6.0	65814	diana.cslab.ece.ntua.gr/DianaToolsNew/index.php?r=tarbase/index	Vergoulis *et al.*(2012)

Table 3.3: Categories of miTI validation experiment used in TarBase 6.0.

Symbol	Name	Description
R	Reporter Gene Assay	A reporter gene (e.g. luciferase) is used to test for changes in expression of the putative target mRNA in response to miRNA transfection or knockout.
N	Northern Blot	Tests changes in expression of the putative target mRNA in response to miRNA transfection or knockout.
W	Western Blot	Tests changes in expression of the putative target protein in response to miRNA transfection or knockout.
Q	PCR	Quantitative PCR is used to test for changes in expression of the putative target mRNA in response to miRNA transfection or knockout.
P	Proteomics	A high-throughput proteomics method (e.g. pSILAC) is used to test changes in expression of multiple proteins in response to miRNA transfection or knockout.
M	Microarray	A microarray is used to test changes in expression of multiple mRNAs in response to miRNA transfection or knockout.
A	Sequencing	Sequencing of miRNA-target sites, often associated with Argonaute immunoprecipitation (e.g. HITS-CLIP).
D	Degradome	MicroRNA cleavage sites in a putative target mRNA are identified using degradome sequencing.
O	Other	An experimental method was used that does not fit into any of the other categories, or the experimental method used is unknown.

For each of these vertebrate species given as available in TarBase 6.0 (Vergoulis *et al.* 2012), all available mature miRNA IDs were obtained from a local copy of miRBase release 13 (Griffiths-Jones 2004, Griffiths-Jones 2006, Griffiths-Jones *et al.* 2008), using the Perl script GetMirbaseInfo.pl with the given species name as a filter and with the attribute ‘Mature miRNA ID'. The set of mature miRNA IDs for each species were then input to the Perl script TarBaseQueryAgent.pl — a web query robot script that makes use of the web user agent Perl module LWP::UserAgent (Aas 2012) — which queried the TarBase 6.0 web interface and output a TSV file with a set of validated miRNA-target interactions for the given species, each with its own supporting study. No results were returned for the set of query miRNAs submitted for the Western Clawed Frog (*Xenopus tropicalis*), so this species could not be included in the dataset. The raw validated miTI data files were then curated, prepared and filtered as described below.

III. Curating Validated MicroRNA-Target Interaction Data

The raw miTI validation data downloaded from TarBase 6.0 required manual curation due to issues with some entries: (a) a miRNA being derived from a different species to that in which its target gene is present; (b) a miRNA lacking a corresponding miRBase accession; (c) inconsistent and occasionally uninformative gene identifiers — the same gene could be given different identifiers depending on the supporting study and source database. The last issue presented the greatest challenge: because a gene could have multiple identifiers, redundant entries exist for the same miTI validation. This redundancy — especially apparent between entries sourced from different databases — was eliminated by mapping each miRNA target to a unique Ensembl gene entry by automated processes and, in cases where these failed, manual curation. See Figure 3.4 for an overview of this process.

For those TarBase entries in which the species of the target gene was different from that of the miRNA: if the supporting study indicated clearly that the target gene was in the same species as the given miRNA, the entry was edited to reflect this; otherwise, the entry was removed from the dataset.

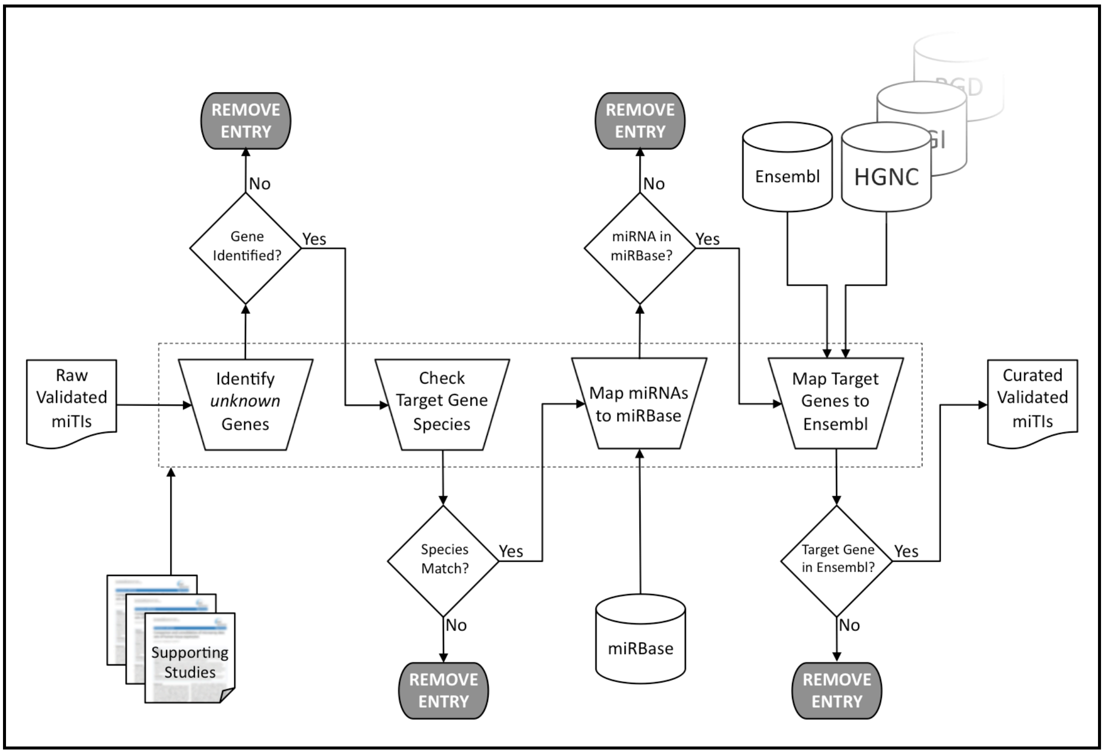


Figure 3.4: Validated microRNA-target interaction curation pipeline.

Shown is a bioinformatics pipeline depicting the steps involved in curating validated miRNA-target interaction data. Arrows indicate the direction of process flow. For information on symbols, see the pipeline key in Figure 2.2.

For those TarBase entries in which the miRNA was listed without a miRBase accession: if the relevant supporting study identified the miRNA unambiguously, the miRBase accession was obtained and the entry was edited to included this. Otherwise, the entry was removed from the dataset.

For those TarBase entries in which the gene identifier was given as 'unknown' or a similarly uninformative name: if possible, the name of the gene was obtained from the supporting study. Otherwise, the entry was removed.

Because the gene symbol used in each TarBase entry may differ depending on the symbol used in the supporting study and on the source database, the gene identifiers used in the relevant TarBase 6.0 entries were each mapped to a corresponding Ensembl entry to eliminate redundancy. This was done as follows: a non-redundant list of the gene identifiers in TarBase 6.0 was obtained for each species. This non-redundant gene symbol set was separated into two: one containing RefSeq symbols and one containing all other symbols. Both sets of symbols were mapped to their corresponding Ensembl Gene ID using an Ensembl BioMart query with filters of type 'RefSeq mRNA' and 'HGNC symbol' (or equivalent for that species), respectively.

With those genes for which the BioMart query failed to return a result, a corresponding Ensembl Gene ID was obtained manually, where possible. This was done with reference to authoritative databases — such as Ensembl (Flicek *et al.* 2012), HGNC (Seal *et al.* 2011) and MGI (Eppig *et al.* 2012) — and if necessary, by consulting the relevant supporting study. In cases where there was conflicting information about a target gene, the original supporting study took precedence. For example, the miR2Disease-sourced TarBase entry validating an interaction between miRNA hsa-miR-17 and human gene 'FBX031' lists Tan *et al.* (2009) as its supporting study. Unfortunately, there is no record of a human gene with symbol 'FBX031' in either HGNC or Ensembl. However, the original supporting study (Tan *et al.* 2009) refers to this target gene by the symbol 'FBXO31' as well as 'FBX031'; giving its GenBank accession as NM_024735. The gene 'FBXO31', with gene name 'F-box only protein 31' has Ensembl Gene ID 'ENSG00000103264'. Indeed, a miRTarBase-sourced TarBase entry exists for a validated interaction between hsa-miR-17 and human 'FBXO31', listing Tan *et al.* (2009) as its supporting study. Because of their redundancy, these two TarBase entries — one sourced from miR2Disease and one from miRTarBase — were subsequently merged into one entry.

For genes that are part of a genome fix patch — indicating an error in the reference sequence of the genome — the patched version was chosen if there was no reference sequence version of the gene, or if the annotated 3′ UTR sequences of the gene in question differed between the patch and the reference sequence. Otherwise, the reference sequence entry was given preference. For miRNA target symbols that referred to a clone or EST as a validated miRNA target, the alignment of the clone/EST to the relevant genome assembly in Ensembl release 65 was used to identify a specific target gene, where that alignment contained a single annotated gene.

If a target symbol — for whatever reason — could not be mapped unambiguously to a single Ensembl Gene ID, all TarBase entries relating to that validated miRNA target were removed from the dataset. Table 3.4 shows the number of target symbols in the raw TarBase data, the number of such symbols mapped unambiguously to an Ensembl Gene ID, and a breakdown of the numbers of target symbols that could not be mapped to Ensembl for different reasons. These reasons could include unrecognised target identifiers without an associated gene symbol (e.g. Contig10441_RC), an invalid GenBank ID, an unsuccessful query of all relevant databases, the failure of a clone or EST to align to a unique gene, or the target symbol referring to a conjoined gene. Athough a considerable number of miRNA targets could not be mapped successfully to Ensembl, it is unlikely that their exclusion would bias the resulting set of validated miRNA-target interactions: their exclusion was due in many cases to the lack of information about validated miRNA targets (e.g. unavailable symbol, invalid GenBank ID, unsuccessful query), and in most cases to the inability to determine, from the available evidence, which of a set of genes is the miRNA target (e.g. failure to align clone sequences to a unique gene, conjoined genes).

The end result of this curation process was a set of raw TarBase 6.0 entries for which each entry had a corresponding miRNA accession in miRBase and a corresponding target gene accession in Ensembl. (Table 3.5 shows the number of TarBase entries modified, inserted and deleted in this processs, as well as the total number of curated TarBase entries for each species.) However, redundant entries remained that referred to the same miRNA-target interaction — these entries had different gene identifiers but mapped to the same Ensembl Gene ID. It was necessary to eliminate this redundancy and to shift the miRNA-target validation data from the perspective of miRNA-target validation to the perspective of the miRNA-target interaction. This preparation of the TarBase 6.0 data is described in the following section.

IV. Preparing Validated MicroRNA-Target Interaction Data

The set of curated, validated miRNA-target interactions for each species were prepared for use with the Perl script PrepTarBaseData.pl. Taking as input a raw TarBase data file as output by TarBaseQueryAgent.pl, this script output two TSV files: one containing the non-redundant set of miRNA-target interaction validations (miTVs) and one containing the non-redundant set of validated miRNA-target interactions (miTIs). The number of miTVs and miTIs for each species is shown in Table 3.5. See Figure 3.5, part A for an overview of this process.

The files output by PrepTarBaseData.pl contain much the same data as the input, but with redundancies resolved and conflicting evidence or interpretations noted, and reoriented to the perspective of a miRNA-target validation and of a miRNA-target interaction, respectively. In this context, a miRNA-target validation is regarded as a validation of a miRNA-target interaction supported by a specific study, while a validated miRNA-target interaction refers specifically to the interaction between the given miRNA and target, which may be supported by more than one study. Each miRNA and target gene was referred to by its miRBase accession and Ensembl Gene ID, respectively, so as to eliminate redundancy caused by the use of different gene symbols in different databases and supporting studies.

Table 3.4: Mapping target genes to the Ensembl database.

Species	Targets in TarBase	Symbol Unavailable	GenBank ID Invalid	Query Failure	Alignment Failure	Conjoined Gene	Targets in Ensembl
Bos taurus	8	0	0	0	0	0	8
Danio rerio	111	0	0	0	0	0	111
Gallus gallus	11	0	0	0	0	0	11
Homo sapiens	12,919	138	65	287	457	5	11,967
Mus musculus	4,589	0	0	62	48	0	4,479
Rattus norvegicus	186	0	1	6	0	0	179

Table 3.4 Legend:

Shown is a summary of the process of mapping TarBase miRNA-target gene identifiers to their corresponding Ensembl gene entry, across 6 species of interest. The leftmost column but one, **Targets in TarBase**, shows the total number of unique miRNA-target gene identifiers in the raw TarBase data for the given species. The rightmost column, **Targets in Ensembl**, shows the number of such miRNA-target gene identifiers for which a corresponding Ensembl Gene ID was obtained. Each of the intermediate columns shows the number of miRNA-target identifiers that could not be mapped to an Ensembl Gene ID for a particular reason. The column **Symbol Unavailable** shows the number of miRNA targets for which the supporting study provided an identifier for the target, but not a corresponding gene symbol (e.g. Contig10441_RC). The column **GenBank ID Invalid** shows the number of miRNA targets for which the given GenBank ID is under review, suppressed, replaced or removed. The column **Query Failure** shows the number of miRNA targets for which a query to a database failed to return a unique gene identifier. The column **Alignment Failure** shows the number of miRNA-target identifiers referring to a clone or EST, for which alignment to Ensembl genomic sequence encompassed multiple genes, or none. The **Conjoined Gene** column shows the number of conjoined genes in each species — these were removed, since it was unknown which of the two component genes harboured target sites for the cognate miRNA.

Table 3.5: Assembly of validated microRNA-target interactions from TarBase data.

Species	Raw TarBase Entries	Curation			Curated TarBase Entries	miTVs	miTIs	Filters		Filtered Validated miTIs
		E	I	D				Conflict	3′ UTR	
Bos taurus	8	0	0	0	8	4	4	0	0	4
Danio rerio	143	6	1	4	140	106	106	1	3	102
Gallus gallus	12	0	0	0	12	11	11	0	1	10
Homo sapiens	22,828	81	5	1,397	21,436	19,345	18,424	167	324	17,933
Mus musculus	9,475	79	0	264	9,211	9,034	8,950	10	85	8,855
Rattus norvegicus	421	9	0	55	366	269	267	0	16	251

Table 3.5 Legend:

Shown is a summary of the process of curating, preparing and filtering TarBase-sourced validated miRNA-target interaction data for 6 species of interest. The leftmost column but one, **Raw TarBase Entries**, shows the number of raw miRNA-target validation entries obtained from TarBase. Each such raw TarBase entry corresponds to an individual validation of a given miRNA-target interaction, as supported by a specific study. The subsequent three columns summarise the semi-automated curation of the raw TarBase data. The columns **E**, **I** and **D** respectively show the number of edits, insertions and deletions performed on the TarBase data, while the column **Curated TarBase Entries** shows the number of TarBase entries remaining after the curation process was complete. The columns **miTVs** and **miTIs** show the output of TarBase data preparation, and contain the number of non-redundant miRNA-target validations and the number of non-redundant validated miRNA-target interactions, respectively. The filtering process is summarised by two columns, **Conflict** and **3′ UTR**, which show the number of miRNA-target interactions removed by the 'Conflicting Evidence' and 'Annotated 3′ UTR' filters, respectively. The rightmost column, **Filtered Validated miTIs**, shows the final tally of miRNA-target interactions in each species.

Where multiple source databases referred to the same miTI validation, the outcome reported by each source database was checked against the others. Each validation sourced from miR2Disease (Jiang *et al.* 2009) and miRecords (Xiao *et al.* 2009) was taken to be a positive validation: the outcome of the validation experiment supported the cognate miRNA-target interaction. If a validation was sourced from miRTarBase (Hsu *et al.* 2010), the outcome of the validation was taken directly from the miRTarBase data files. If the same validation outcome was reported by all source databases, the miTI validation was marked as positive or negative as appropriate. If different outcomes were reported, the miTI validation was marked as having a conflicting interpretation.

The set of miTI validations were then used to create the set of validated miTIs. For each miRNA-target pair, all relevant miTI validations were checked against each other. If the outcome of every validation was positive, the miRNA-target interaction was marked as positively validated. If the outcome of every validation was negative, the miRNA-target interaction was marked as negatively validated. If differing validation outcomes were reported for a given miRNA-target interaction, the miRNA-target interaction validation status was set to conflicting evidence. The only exception to this was where the set of validations appeared to support a miRNA-target interaction mediated by translational repression: where validations based on mRNA expression (i.e. N, Q or M in Table 3.3) were negative, validations based on protein expression (i.e. W or P in Table 3.3) were positive, and no other validation outcome contradicted this interpretation.

The key result of this preparation process was a non-redundant set of validated miRNA-target interactions. However, miRNA-target interactions with conflicting evidence or conflicting interpretations of the evidence remained. Additionally, target genes lacking an annotated 3′ UTR in Ensembl release 65 could not be used in the miTP benchmark, so validated miRNA-target interactions involving such unannotated genes would have to be removed.

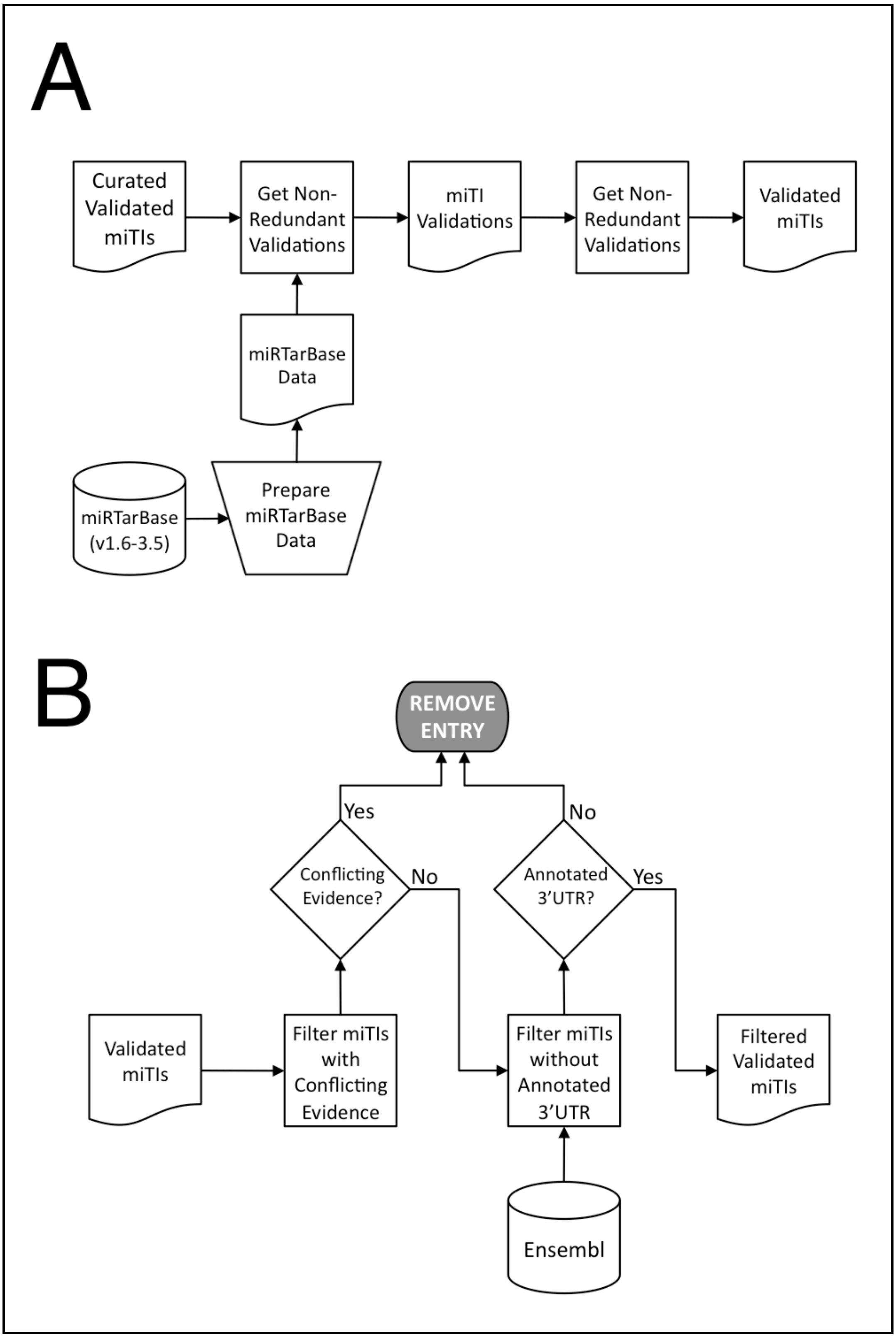


Figure 3.5: Validated microRNA-target interaction preparation and filtering pipelines. Shown are two bioinformatics pipelines depicting the steps involved in (A) preparing and (B) filtering validated miRNA-target interaction data. Arrows indicate the direction of process flow. For information on symbols used, see the pipeline key in Figure 2.2.

Both types of problematic miRNA-target interaction were filtered from this non-redundant set of validated miRNA-target interactions, as described in the following section. The number of miTIs removed by these filters in each species is shown in Table 3.5, as well as the number of filtered, validated miTIs remaining after this process. It is not expected that this filtering process would bias the set of validated miRNA-target interactions considerably, since they were removed due to either conflicting experimental results or lack of annotation; neither factor is known to be associated with any feature of miRNA-target interactions to the extent that their removal would bias the resulting dataset.

V. Filtering Validated MicroRNA-Target Interaction Data

Non-redundant validated miRNA-target interactions were filtered using the Perl script FilterTable.pl to remove all miRNA-target interactions subject to conflicting evidence or conflicting interpretations of the evidence, then filtered to retain only miTIs for which the target gene had an annotated 3′ UTR in Ensembl. See Figure 3.5, part B for an overview of this process.

To filter for Ensembl 3′ UTR annotation, the target gene of each miRNA-target interaction was submitted in a query of Ensembl BioMart (Flicek *et al.* 2012, Smedley *et al.* 2009) to obtain all available 3′ UTR sequences. For each gene, a representative 3′ UTR sequence was selected using the Perl script GetRepresentativeUTRs.pl. The representative 3′ UTR sequence was taken to be the longest annotated 3′ UTR sequence for the given gene. If a target gene lacked a representative 3′ UTR, or had a representative 3′ UTR shorter than 37 nucleotides, all miRNA-target interactions involving that target gene were removed from the dataset of validated miTIs. The minimum 3′ UTR length was set to 37 bases as this is the sum of the average length of a mature miRNA (i.e. 22 bases) and the distance downstream of the open reading frame (i.e. 15 bases) in which miRNA-target interactions have been found to have reduced effectiveness, comparable with that found in coding regions (Grimson *et al.* 2007).

No annotated 3′ UTR sequences could be obtained for the *Ovis aries* target gene Rtl1 (the only target gene in that species), or for any of the 4 target genes in *Xenopus laevis*, so all miRNA-target interactions for both of these species were removed from the dataset. This left 6 remaining species of interest for the benchmark test: Human (*Homo sapiens*), Mouse (*Mus musculus*), Rat (*Rattus norvegicus*), Cow (*Bos taurus*), Chicken (*Gallus gallus*) and Zebrafish (*Danio rerio*).

In addition to the aforementioned filters, to control for bias introduced by the presence of validated miRNA-target interactions that were used as training data — directly or indirectly — by any of the miTP methods being benchmarked, the validated miTI dataset was filtered for any studies or specific miRNA-target interactions that had been used as training data by any of the 9 miTP methods described in Section 3.1.2. A total of 36 papers and 521 specific miRNA-target interactions were identified as possible sources of training data bias. TarBase entries supported by these 36 studies, or involving one of the 521 specific miRNA-target interactions, were removed from the validated miTI dataset using the Perl script FilterTarBaseData.pl. The removed miTI validations were not analysed further, due to the possible bias introduced by their use as training data. The effect of filtering training data is outlined briefly during the next section.

VI. Dividing Validated MicroRNA-Target Interaction Data

The dataset of validated miRNA-target interactions was split into two subsets for the benchmark: (i) a set of positively validated miRNA-target interactions, and (ii) a high-quality subset of positively and negatively validated miRNA-target interactions.

(i) Positively Validated MicroRNA-Target Interactions

The positively validated miTIs were to be used to compare the recovery rates of each miTP method in a graph similar to that used by Alexiou *et al.* (2009), showing the number recovered miRNA-target interactions per miRNA against the number of predicted miRNA-target interactions per miRNA. This would involve testing every miRNA of interest against every 3′ UTR in the cognate species. These were obtained by filtering the main validated miTI dataset to remove all negatively validated miRNA-target interactions. There were 17,699 positively validated human miTIs; exclusion of training data removed 3,384 of these, leaving a total of 14,315 validated miRNA-target interactions in *Homo sapiens*. This set of positively validated miTIs is accessible in the electronic appendix at this file path: Appendix / home / projects / benchmark / pipelines / mitp-benchmark / files / Benchmark-POSITIVE-XTD.zip.

(ii) High-Quality Validated MicroRNA-Target Interactions

The high-quality validated miRNA-target interactions were to be used to construct a **receiver operating characteristic** (ROC) graph (Egan 1975). In general terms, a ROC curve graphs the true-positive rate against the false-positive rate of a binary classifier at different levels of stringency. This was to be used to compare the ability of each miTP method to distinguish known positively validated miRNA-target interactions from known negatively validated miRNA-target interactions. Unlike the estimation of recovery rates, the construction of a ROC graph would only involve testing the miRNA-target pairs in the set of high-quality validated miTIs.

High-quality *positively* validated miRNA-target interactions were taken to be those validated miTIs that had been annotated in TarBase 6.0 as having been directly validated (Vergoulis *et al.* 2012). There were 479 such directly and positively validated miRNA-target interactions in *Homo sapiens*.

High-quality *negatively* validated miRNA-target interactions were taken to be those validated miTIs that had been annotated in TarBase 6.0 as having been directly validated or validated in a gene reporter assay (Vergoulis *et al.* 2012). These high-quality negatively validated miTIs were augmented by those miRNA-target interactions annotated as non-functional in the most recent release 3.5 of miRTarBase (Hsu *et al.* 2010). These validated miTIs were manually reviewed: if there was reason to suspect that the miRNA-target interaction was functional, the validation was removed from consideration. At the end of this process, there were 45 high-quality negatively validated miRNA-target interactions in *Homo sapiens*.

Due to the low number of high-quality validations in species other than *Homo sapiens*, a high-quality validated miTI dataset could only be assembled for human. This set of high-quality miTI validations originally included 1092 positively validated miTIs and 46 negative validations. Exclusion of training data reduced these to 479 positive and 45 negative validations. The resulting set of high-quality miTI validations is accessible in the electronic appendix at the path: Appendix / home / projects / benchmark / pipelines / mitp-benchmark / files / Benchmark-DIRECT-END2.zip.

3.2.2 MicroRNA Sequence Download

The Perl script GetMirbaseSequences.pl was used to access a local copy of miRBase release 18 (Griffiths-Jones 2004, Griffiths-Jones 2006, Griffiths-Jones *et al.* 2008, Kozomara and Griffiths-Jones 2011), obtaining sequences for all 586 mature miRNAs involved in one or more validated miRNA-target interactions. The result for each species was a FASTA sequence file containing all mature miRNA sequences involved in a validated miRNA-target interaction.

3.2.3 Target Sequence Download

For the miRNA-target prediction method benchmark, a set of representative 3′ UTRs was required for each of the 6 benchmark test species: Human (*Homo sapiens*), Mouse (*Mus musculus*), Rat (*Rattus norvegicus*), Cow (*Bos taurus*), Chicken (*Gallus gallus*) and Zebrafish (*Danio rerio*). For each of these species, the 3′ UTR of every transcript was downloaded from Ensembl BioMart (Flicek *et al.* 2012, Smedley *et al.* 2009), using a 'Sequences' query with the attributes 'Ensembl Gene ID', 'Ensembl Transcript ID' and '3′ UTR'. A set of representative transcripts was obtained using the script GetRepresentativeUTRs.pl, which took the longest annotated 3′ UTR for each gene, and removed genes that lacked an annotated 3′ UTR longer than the minimum 37 nucleotides. This was set as the minimum 3′ UTR length as it is the sum of the average length of a mature miRNA (i.e. 22 bases) and 15 bases downstream of the open reading frame, a region in which miRNA-target interactions have been found to have reduced effectiveness (Grimson *et al.* 2007). The result was a FASTA file containing all representative 3′ UTRs for the given species.

3.3 Methods

This section describes the procedures followed in a benchmark study of 9 current miRNA-target prediction methods. Each miTP method was tested against a common dataset, with the results then compared to a set of known miRNA-target interactions. All scripts used during this process can be accessed from the following location in the electronic appendix: Appendix / home / code / scripts.html.

3.3.1 MicroRNA-Target Prediction Benchmark Run

A miRNA-target prediction benchmark run was performed for all 9 miTP methods on a common dataset, in three main stages: (I) preparation of benchmark run input data files, (II) the benchmark run itself, and (III) processing of benchmark run output files. See the bioinformatics pipeline in Figure 3.7 for an overview.

I. Benchmark Run Input Preparation

All miTP methods made use of the miRNA and target 3′ UTR sequences in some way. However, only miRanda, Hitsensor and TargetSpy took the standard FASTA miRNA and target sequence files as input. Specific input data preparation measures were required for the remaining programs: (i) MicroTar, (ii) MultiMiTar, (iii) PITA, (iv) RNAhybrid, (v) TargetScan, and (vi) mirMap. These were as follows.

(i) MicroTar

Because it calculates the folding energy of the entire messenger RNA molecule, MicroTar requires the full mRNA sequence as input, with the 3′ UTR region set to lowercase. The Ensembl Transcript IDs of the representative 3′ UTRs were submitted to an Ensembl BioMart 'Sequences' query with sequence attribute 'cDNA'. With the resulting mRNA sequence and the previously obtained 3′ UTR sequence as input, the Perl script SetCaseFASTA.pl identified the 3′ UTR of the full mRNA sequence and set it to lowercase. For ENSGALG00000013959 and ENSGALG00000022194 (*Gallus gallus* genes), the 3′ UTR matched multiple regions in their mRNA sequences. Both were separated from the main dataset, used to set the mRNA sequence case manually, and then returned to the dataset.

(ii) MultiMiTar

MultiMiTar has very specific requirements for its data input (Mitra and Bandyopadhyay 2011). In consultation with the authors (Mitra 2012, personal correspondence), the following changes were made to the input data prior to running MultiMiTar. These are illustrated in Figure 3.6.

- A 'test' file was compiled, listing each miRNA-target pair to be tested.

- All miRNA headers were changed to follow the format '>miRNA_ID miRBase18', where 'miRNA_ID' was the ID of the miRNA and 'miRBase18' was a required non-whitespace string.

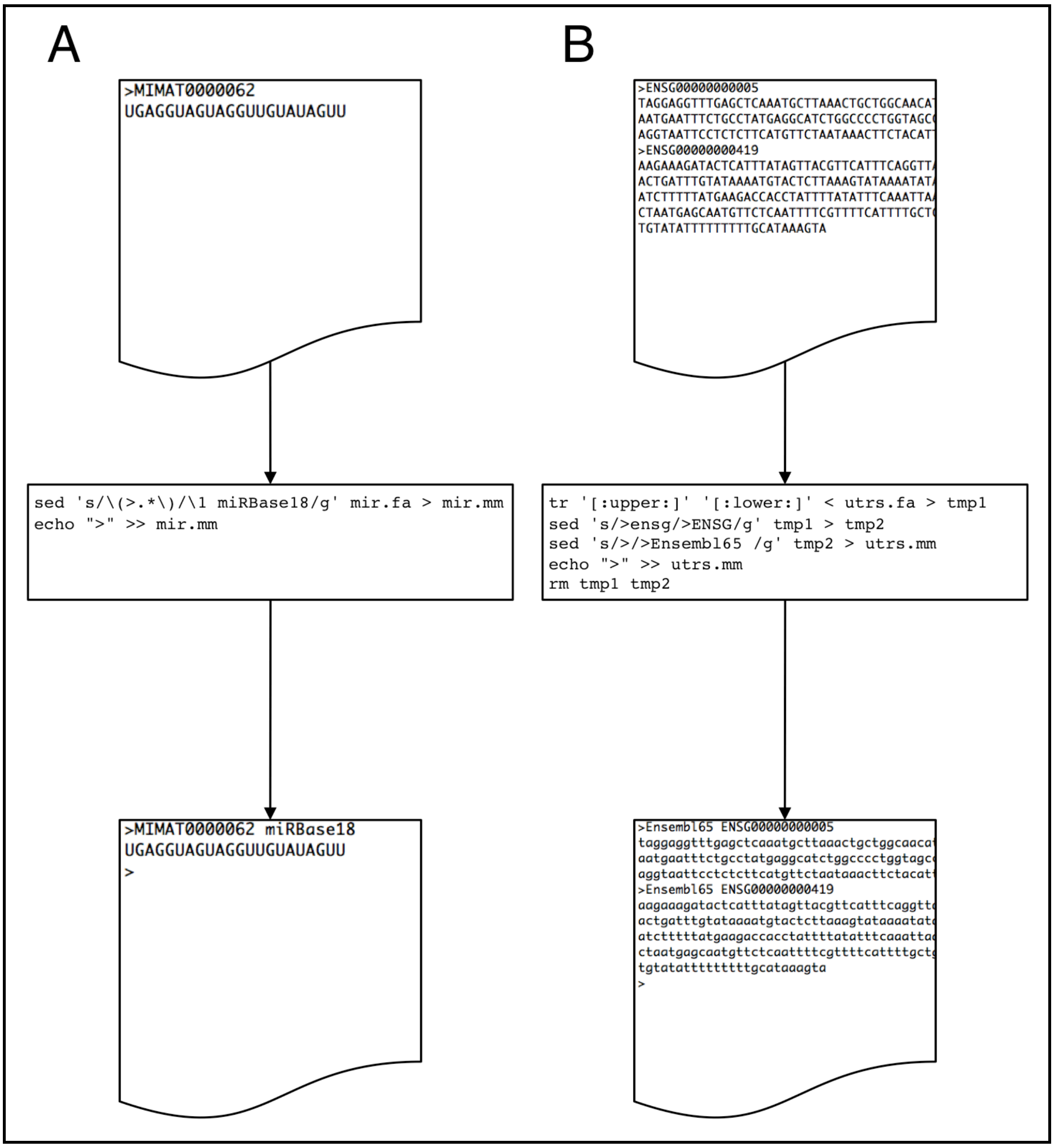
- All target 3′ UTR headers must follow the format '>Ensembl65 target_ID', where 'Ensembl65' was a required non-whitespace string and 'target_ID' was the Ensembl ID of the target 3′ UTR.

- All target 3′ UTR sequences were set to lowercase.

- All input files, whether miRNA sequence, target sequence or test list, were appended with a line containing nothing but a greater-than symbol (i.e. '>').

(iii) PITA

PITA requires that for each 3′ UTR sequence tested, the corresponding coding sequence (CDS) is also input (Kertesz *et al.* 2007). To obtain these, the Ensembl Transcript IDs of the representative 3′ UTRs were submitted to an Ensembl BioMart 'Sequences' query with sequence attribute 'CDS'. Each target 3′ UTR was then input to PITA along with its corresponding CDS.



```
>MIMAT0000062
UGAGGUAGUAGGUUGUAUAGUU
```

```
sed 's/\(>.*\)/\1 miRBase18/g' mir.fa > mir.mm
echo ">" >> mir.mm
```

```
>MIMAT0000062 miRBase18
UGAGGUAGUAGGUUGUAUAGUU
>
```

```
>ENSG00000000005
TAGGAGGTTTGAGCTCAAATGCTTAAACTGCTGGCAACAT
AATGAATTTCTGCCTATGAGGCATCTGGCCCCTGGTAGCC
AGGTAATTCCTCTCTTCATGTTCTAATAAACTTCTACATT
>ENSG00000000419
AAGAAAGATACTCATTTATAGTTACGTTCATTTCAGGTTA
ACTGATTTGTATAAAATGTACTCTTAAAGTATAAAATATA
ATCTTTTTATGAAGACCACCTATTTTATATTTCAAATTAA
CTAATGAGCAATGTTCTCAATTTTCGTTTTCATTTTGCTG
TGTATATTTTTTTTTGCATAAAGTA
```

```
tr '[:upper:]' '[:lower:]' < utrs.fa > tmp1
sed 's/>ensg/>ENSG/g' tmp1 > tmp2
sed 's/>/>Ensembl65 /g' tmp2 > utrs.mm
echo ">" >> utrs.mm
rm tmp1 tmp2
```

```
>Ensembl65 ENSG00000000005
taggaggtttgagctcaaatgcttaaactgctggcaacat
aatgaatttctgcctatgaggcatctggcccctggtagcc
aggtaattcctctcttcatgttctaataaacttctacatt
>Ensembl65 ENSG00000000419
aagaaagatactcatttatagttacgttcatttcaggtta
actgatttgtataaaatgtactcttaaagtataaaatata
atctttttatgaagaccacctattttatatttcaaattaa
ctaatgagcaatgttctcaattttcgttttcattttgctg
tgtatattttttttttgcataaagta
>
```

Figure 3.6: MultiMiTar input preparation.

Shown is an outline of the process of preparing input files for MultiMiTar. (A) An example conversion of a miRNA FASTA sequence file to the format required by MultiMiTar. At top is the original FASTA file and at bottom the resulting MultiMiTar input miRNA file. At centre are the bash commands used to insert the text 'miRBase18' into each FASTA header and to append a greater-than symbol to the end of the file. (B) An example conversion of a target 3′ UTR FASTA file to the format required by MultiMiTar. At top is the original FASTA file and at bottom the resulting MultiMiTar input target sequence file. At centre are the bash commands used to convert all sequences to lowercase, restore Ensembl Gene IDs to uppercase, insert the text 'Ensembl65' into each FASTA header and to append a greater-than symbol to the end of the file. Note that for the purposes of this example, FASTA files are given the extension '.fa', while files ready for input to MultiMiTar are given the extension '.mm'.

(iv) RNAhybrid

The input parameters of RNAhybrid must be calibrated using RNAcalibrate, which takes as input a file listing dinucleotide frequency (DiF) information for the target genome (Rehmsmeier *et al.* 2004, Krüger and Rehmsmeier 2006). For each benchmark test species, the dinucleotide frequencies of its representative 3′ UTRs were obtained using the Perl script GetDinucleotideFrequencies.pl. These dinucleotide frequencies were used by RNAcalibrate to estimate parameter values for RNAhybrid (Rehmsmeier *et al.* 2004, Krüger and Rehmsmeier 2006).

(v) TargetScan

TargetScan requires an orthologous 3′ UTR alignment (where possible) for each 3′ UTR sequence (Garcia *et al.* 2011, Friedman *et al.* 2009, Grimson *et al.* 2007, Lewis *et al.* 2005). TargetScan 6 also requires a TA/SPS file containing a table of target abundance (TA) and seed-pairing stability (SPS) estimates for all possible 7mer miRNA seed sequences (Garcia *et al.* 2011).

A human TA/SPS file is bundled with TargetScan: SPS estimates can be used across species, but TA estimates are species-specific and were obtained separately for each benchmark test species. The set of representative 3′ UTRs for each genome were input to the Perl script GetTargetAbundance.pl, which estimates the miRNA target abundance (TA) of the sequences in an input FASTA file and outputs these to a target abundance table file. Following the procedure described in Garcia *et al.* (2011), this script searches for non-overlapping canonical miRNA target sites in the set of input sequences and estimates target abundance based on this. TargetScan 6 expects, and GetTargetAbundance.pl produces, log-transformed (to base 10) TA estimates. These were then included with the SPS data in the 'TA' column of a species-specific TA/SPS file.

To assess miRNA-target site conservation, TargetScan requires an alignment of a 3′ UTR across two or more species. Other than the 6 benchmark test species, there are 22 species in the TargetScan phylogeny (the phylogeny used by TargetScan to assess the extent of conservation). Of these, 14 have annotated 3′ UTR sequences in Ensembl release 65. A group of representative transcripts were obtained for these 14 species. These were then incorporated in orthologous 3′ UTR alignments as follows. For each of the six benchmark test species, a list was compiled of the Ensembl Gene IDs of all representative 3′ UTRs. This was used as a filter in a series of Ensembl BioMart 'Homologs' queries — one for each other species in the TargetScan phylogeny — with the 'Ortholog' attribute. The resulting two-column ortholog tables were merged into one raw ortholog table for the given species using the Perl script MergeTables.pl.

This raw ortholog table contained the Ensembl Gene IDs of all Ensembl orthologs for each representative 3′ UTR in the given species. In the alignment of orthologous 3′ UTR sequences, TargetScan only allows for one ortholog per species, so it was necessary to choose a representative ortholog in each species for each gene. Given a gene of interest in one species that has many co-orthologs in a second species, a **representative ortholog** is a single ortholog chosen as representative from among those co-orthologs.

The Perl script GetRepresentativeOrthologs.pl was used for each key species to remove orthologs lacking a representative 3′ UTR and to select a representative ortholog in cases where multiple orthologs exist for a given gene in the same species. The representative ortholog was taken to be the orthologous 3′ UTR that aligned best to the 3′ UTR of the given gene. The pairwise alignment between each gene and its orthologs was performed by Needle from the EMBOSS package (Rice *et al.* 2000), an implementation of the Needleman-Wunsch algorithm (Needleman and Wunsch 1970).

The resulting representative ortholog table was used to create an ortholog sequence file for each representative 3′ UTR in the six benchmark test species: Human (*Homo sapiens*), Mouse (*Mus musculus*), Rat (*Rattus norvegicus*), Cow (*Bos taurus*), Chicken (*Gallus gallus*) and Zebrafish (*Danio rerio*). For each representative 3′ UTR with one or more orthologs, a guide tree was created using the Perl script CreateGuideTree.pl with the TargetScan phylogeny as a template tree (see Figure 3.2). The ortholog sequence file and guide tree were then input to PRANK v.100802.

TargetScan does not accept miRNA or target gene input sequence files in FASTA format, but in a custom TSV format. Given the set of miRNA and target 3′ UTR alignment FASTA files, the TargetScan input files were created using the Perl scripts PrepTargetScanMirFiles.pl and PrepTargetScanTarFile.pl, respectively. Once these TargetScan input TSV files were prepared, TargetScan input was ready.

(vi) miRmap

The input requirements for miRmap are similar to those of TargetScan: an orthologous alignment is required where ortholog sequences are available. Because of this, the 3′ UTR ortholog sequence/alignment files created for TargetScan were used again for miRmap. To allow conservation to be estimated for orthologous 3′ UTR ortholog sequences, miRmap also requires a phylogenetic model fit to the 3′ UTR ortholog alignment. This was created for each 3′ UTR ortholog alignment using phyloFit (Siepel and Haussler 2004, Yang 1994, Yang *et al.* 1994) with the TargetScan phylogeny (Friedman *et al.* 2009), using the GTR model (see Section 1.6.2). With the evolutionary model prepared, miRmap input was ready.

II. Benchmark Run

For each of the benchmark test species, the miRNA-target prediction benchmark itself involved running each miRNA-target prediction program on all miRNAs against every representative 3′ UTR in the given species.

All 9 miRNA-target prediction methods were run with default settings. Hitsensor, miRanda, TargetSpy and RNAhybrid all have a miRNA seed binding filter or feature option. These were all run both with and without a seed option. PITA was additionally run both with and without its site flank option. Where used, the PITA site flank was set to 15 nucleotides upstream and downstream. In total, 15 miTP benchmark runs were performed (see Table 3.6).

As the benchmark run for each miRNA-target prediction method was being run in an embarrassingly parallel manner, the set of tasks for each program was run using the ICHEC Taskfarm utility on the Stokes HPC cluster.

Every miTP method except MicroTar performed all miRNA-target predictions within the expected time frame of less than one second per miRNA-target prediction. Possibly because of its more thorough structural analysis, MicroTar failed to perform the full miRNA-target prediction analysis in a timely manner, in some cases taking more than 12 hours to test a single miRNA-target interaction. This called into question the applicability of this method for large-scale miRNA-target predictions, so it was only tested on the comparatively smaller number of high-quality validated miTIs.

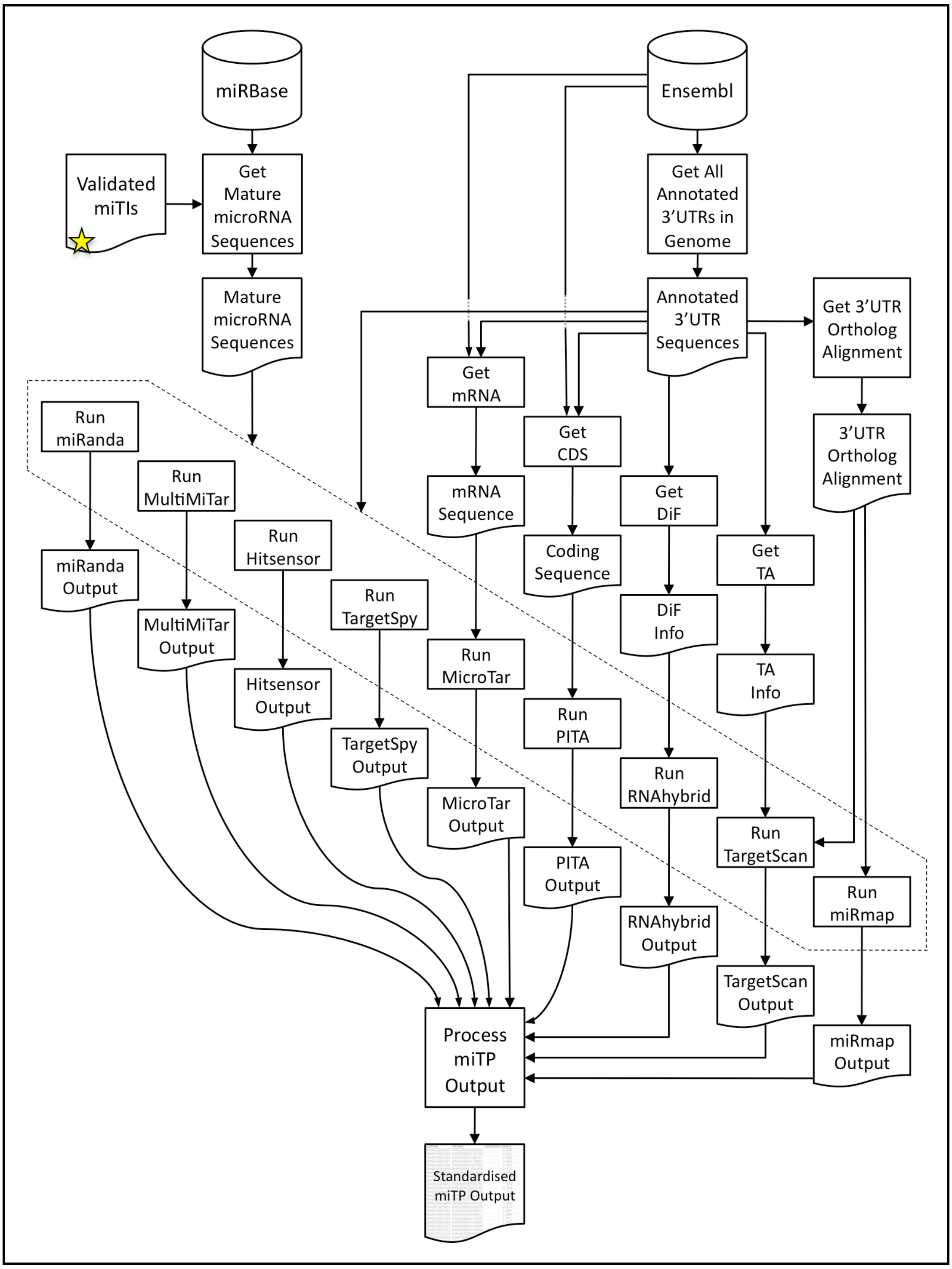


Figure 3.7: MicroRNA-target prediction benchmark run pipeline.

Shown is a bioinformatics pipeline depicting the steps involved in running each of the 9 miTP methods to be compared in the miRNA-target prediction benchmark. Arrows indicate the direction of process flow. For information on symbols used, see the bioinformatics pipeline key in Figure 2.2.

Table 3.6: MicroRNA-target prediction benchmark runs.

Name	Description
HITSENSOR	Hitsensor with default settings
HITSENSORs	Hitsensor with increased weighting for seed match sites
MICROTAR	MicroTar
MIRANDA	miRanda without strict seed
MIRANDAs	miRanda with strict seed binding
MIRMAP	miRmap
MULTIMITAR	MultiMiTar
PITA	PITA with default settings
PITAf	PITA accounting for target site flanking sequence
RNAHYBRID	RNAhybrid with default settings
RNAHYBRIDs	RNAhybrid constrained by seed binding
TARGETSPY	TargetSpy with default settings
TARGETSPYs	TargetSpy with seed match requirement
TSCONSERV	TargetScan conservation script
TSCONTEXT	TargetScan context script

Table 3.6 Legend:

Shown are names and descriptions of the 15 miRNA-target prediction methods and variants run in the miTP benchmark.

III. Benchmark Run Output Processing

Having run each miTP program on its benchmark data, the raw output of each program was processed by the Perl script ProcessMicrornaTargetPredictions.pl. This script takes as input the name of a miRNA target prediction program and a set of raw output files from that program, then outputs a single processed output TSV file, which shows for each miRNA-target prediction the miRNA, target gene and associated prediction score value; these predictions are ranked in the output file from best to worst score value. For an example of this process, see Figure 3.8.

Some miTP programs output predictions at the level of individual miRNA-target sites (e.g. TargetScan, MicroTar), while others output prediction predictions for both each individual target site and for the target gene as a whole (e.g. miRanda, PITA). The script ProcessMicrornaTargetPredictions.pl follows the convention of the previous benchmark studies discussed above, such that miRNA-target site predictions are aggregated to miRNA-target gene predictions where necessary (Sethupathy *et al.* 2006, Baek *et al.* 2008, Selbach *et al.* 2008, Alexiou *et al.* 2009).

TargetSpy, RNAhybrid and miRanda all give two metrics for each miRNA-target prediction. TargetSpy gives a hybridisation energy and score value. RNAhybrid gives a hybridisation energy and p-value. miRanda gives a hybridisation energy and score value. The miTP output files for all three programs were processed for each score type in turn. The score type was then chosen in the manner described in Section 3.3.2.

3.3.2 MicroRNA-Target Prediction Benchmark Comparison

The different methods were assessed by comparing the miRNA-target predictions of each method with known validated miRNA-target interactions in a miRNA-target prediction benchmark comparison (see Figure 3.9). There were three stages: (I) score types were selected for those methods with more than one, (II) a ROC graph was created to compare true- and false-positive rates of all 9 methods, and (III) a recovery graph was made to compare the recovery rates of 8 of the miTP methods.

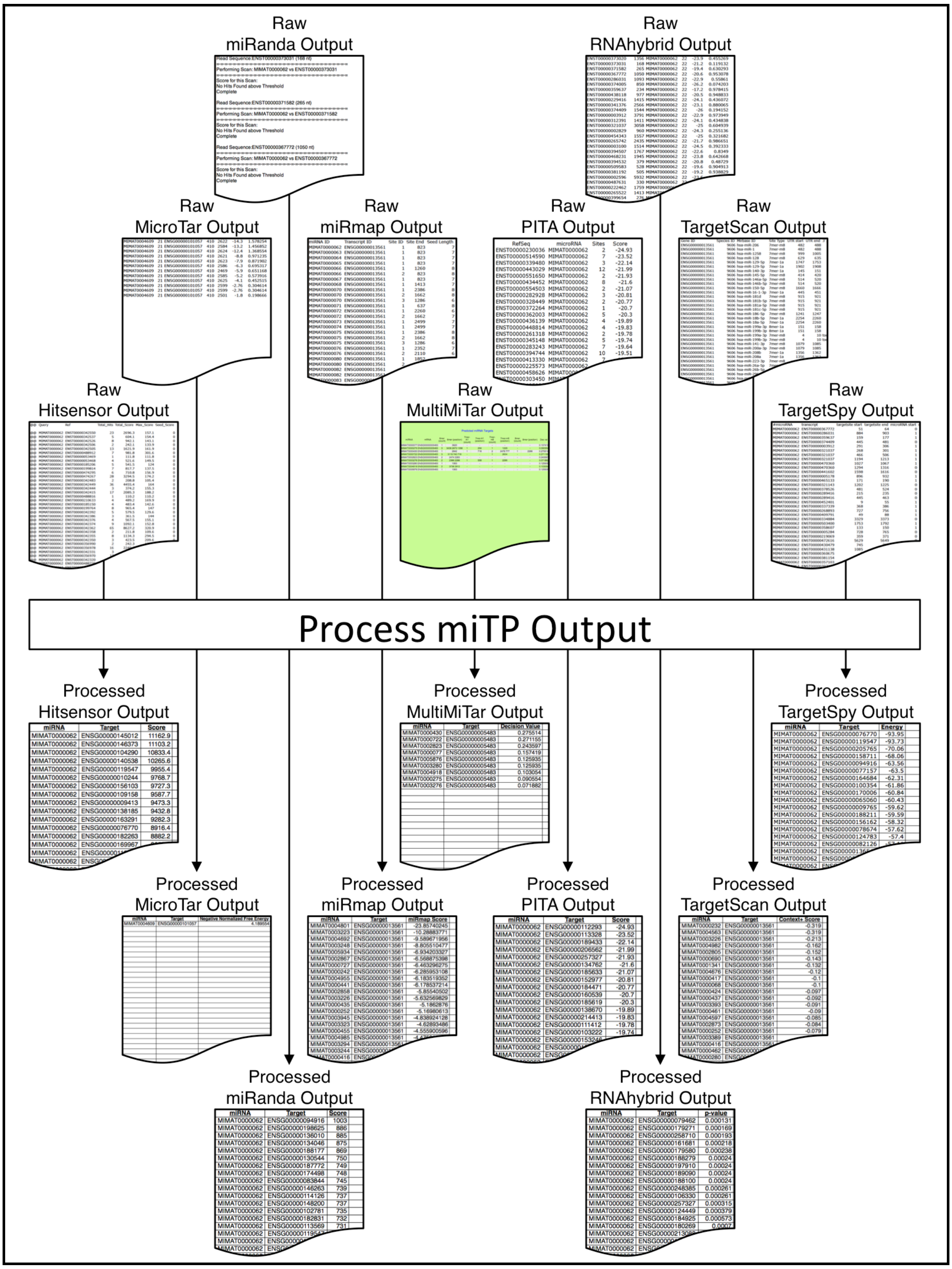


Figure 3.8: miRNA-Target prediction output processing pipeline.

Shown is a simple bioinformatics pipeline depicting the input and output of the processing of raw miRNA-target prediction output from different miTP programs and conversion to a standard layout. Arrows indicate the direction of process flow. For information on symbols, see the bioinformatics pipeline key in Figure 2.2.

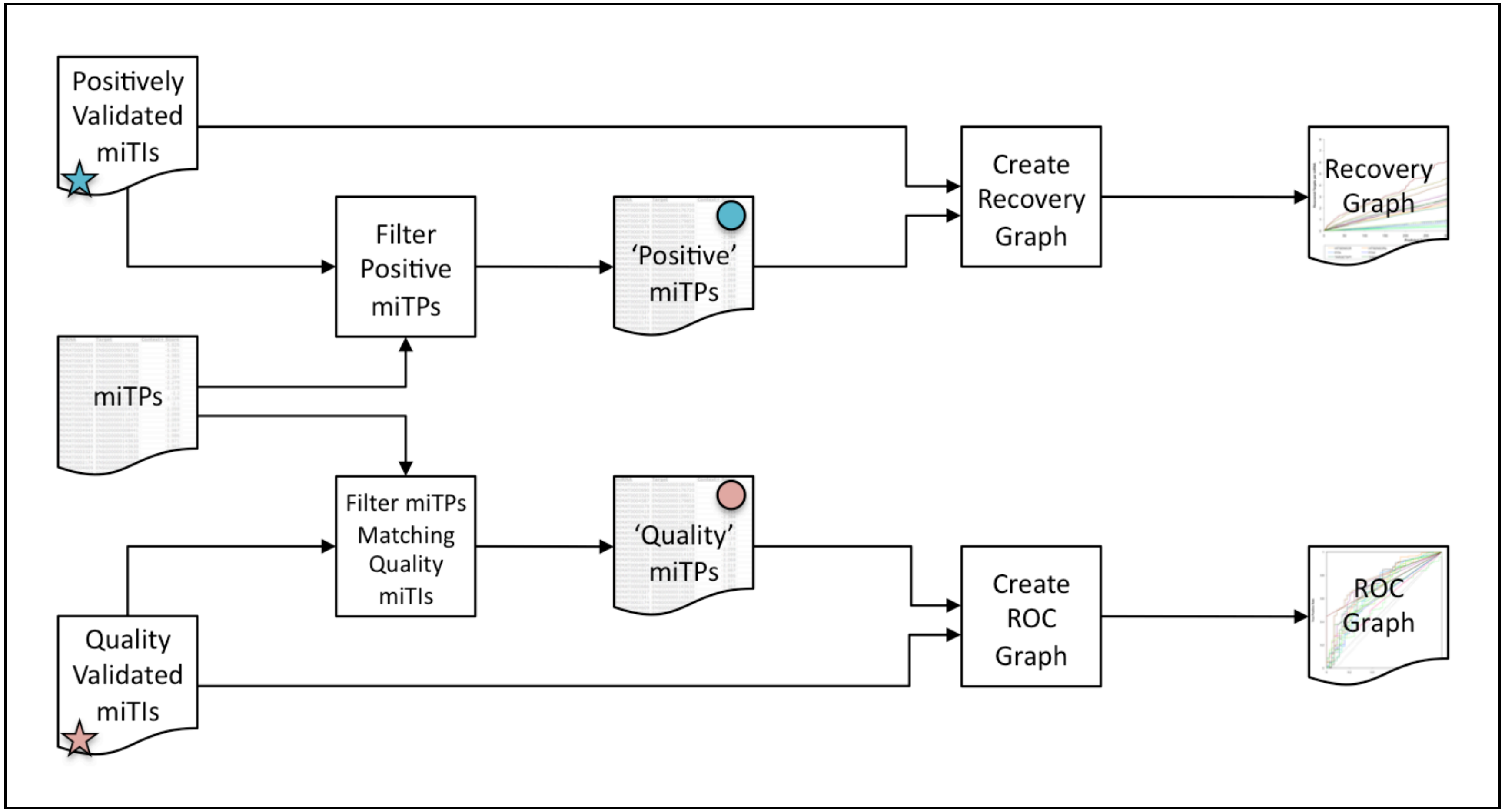


Figure 3.9: miRNA-target prediction benchmark comparison pipeline.

Shown is a bioinformatics pipeline depicting the steps involved in comparing the output of the 9 miTP methods against known miRNA targets and non-targets. Arrows indicate the direction of process flow. For information on symbols, see the pipeline key in Figure 2.2.

I. Score Type Selection

For TargetSpy, RNAhybrid and miRanda, score types were compared with each other using a ROC graph. The web tool StAR (Vergara *et al.* 2008) — available from: http://protein.bio.puc.cl/star/home.php — was used to create a ROC graph for each miTP program. In addition to constructing a ROC graph, the StAR web utility performs statistical analysis, implementing the method described by Delong *et al.* (1988) for comparing the AUC values of different ROC curves using a Mann-Whitney *U* test.

Input files for StAR were prepared using PrepRocStarInput.pl, which reads one or more TSV files containing processed miRNA target prediction output and a TSV file containing directly validated miRNA target interactions (miTIs). This script then creates three output files: a complete data file listing each directly validated miTI, its associated scores according to each miTP method of interest and whether it is positively or negatively validated; a positive target data file containing normalised scores of positively validated targets for each miTP method; and a negative target data file containing normalised scores of negatively validated targets for each miTP method. For each method, scores are normalised to the range -1.0 to 0.0, such that the score of the top-ranking miRNA-target prediction for that method is mapped to -1.0, the minimum possible score for the given method is mapped to 0.0, and the scores of all other miRNA-target predictions by that method are mapped onto an intermediate scale between these two extremes. The positive and negative target data files could then be used as input to the StAR web utility.

For each method, the area under the ROC curve (AUC) of each score type was compared so as to guide selection of a score type for that method, such that if one score type had a significantly higher AUC than another score type, it was given preference. In the absence of a significant difference in AUC, the score type with the better AUC was arbitrarily selected for that method, with the caveat that the selected score type was not significantly better than the alternative.

Two ROC graphs were created for miRanda: one showing the results with a strict seed requirement, the other showing results without this. Both ROC graphs showed two ROC curves: one for the miRanda results ranked by increasing 'Energy', the other for miRanda results ranked by decreasing 'Score'. In both ROC graphs, the area under the ROC curve (AUC) for miRanda's 'Score' ranked output was slightly higher than that for the 'Energy' ranked output, although not significantly so (at a 5% significance level), with a difference in AUC of 0.0115 for miRanda with a strict seed requirement (Mann-Whitney *U* test, P = 0.085), and a difference in AUC of 0.00935 otherwise (Mann-Whitney *U* test, P = 0.162). Because the AUC of miRanda was so similar between score types, 'Score' was arbitrarily chosen as the metric for miRanda.

Two ROC graphs were similarly created for RNAhybrid: one showing the results with a seed binding constraint, the other showing results without this. Both ROC graphs showed two ROC curves: one for the RNAhybrid results ranked by increasing 'Energy', the other for RNAhybrid results ranked by increasing 'p-value' (where p-value ranges from zero to one, and a lower p-value represents a better score). In the ROC graph for RNAhybrid *without* a seed binding constraint, the overall AUC for 'Energy' was 0.672, which was significantly greater than that for 'p-value' miTPs: 0.5639 (Mann-Whitney *U* test, P = 0.026). In the ROC graph for RNAhybrid *with* a seed binding constraint, the overall AUC for 'Energy' was 0.665, which was slightly greater that that achieved by 'p-value' miTPs — 0.663 — but not significantly so (Mann-Whitney *U* test, P = 0.953). However, a visual assessment indicated that the 'p-value' miTPs appeared to have a considerably higher AUC at lower false-positive rates (FPRs) in both ROC graphs (i.e. FPR ≤ 0.15), see Figure 3.10. This seems to indicate that p-value is a better way to distinguish true from false-positives when p-values are low, but that this is a poor classifier at high p-values. Because 'Energy' performed so well overall but 'p-value' appeared to perform better at low false-positive rates, RNAhybrid results with both score types were included in the final ROC graph.

Two ROC graphs were created for TargetSpy: one showing results with a seed requirement, the other showing results without this. Both ROC graphs showed two ROC curves: one for the TargetSpy results ranked by increasing 'Energy', the other for TargetSpy results ranked by decreasing 'Score'. In the ROC graph for TargetSpy *without* a seed requirement, the AUC for 'Energy' was 0.663, while that for 'Score' was 0.657; the difference in AUC between the two score types was not significant (Mann-Whitney *U* test, P = 0.408). In the ROC graph for TargetSpy *with* a seed requirement, the AUC for 'Score' was 0.642, while that of 'Energy' was 0.639; as before, the difference in AUC between the two score types was not significant (Mann-Whitney *U* test, P = 0.84476). Because no significant difference was found between the AUC of the two TargetSpy score types, 'Score' was arbitrarily chosen as the metric for TargetSpy.

Figure 3.10 Legend:

Shown on page 234 overleaf are two ROC graphs created during score type selection for RNAhybrid, in which miRNA-target predictions are shown as ranked by increasing minimum free energy ('Energy') — shown in blue — and by increasing p-value — shown in red. (A) Included is a ROC graph of RNAhyrid miRNA-target predictions *without* a seed binding constraint. The overall AUC of the 'Energy' metric is 0.672, significantly greater than the AUC of 0.5639 achieved by p-value (P = 0.026). (B) Shown is a ROC graph of RNAhybrid miRNA-target predictions *with* a seed binding constraint. The overall AUC of the 'Energy' metric is 0.665, slightly greater than the AUC of 0.663 achieved by p-value, but not significantly so (P = 0.953). Note from graphs A and B that although RNAhybrid miRNA-target predictions ranked by the Energy metric have a higher AUC overall, the AUC of p-value ranked miTPs is greater at low false-positive rates (i.e. FPR ≤ 0.15).

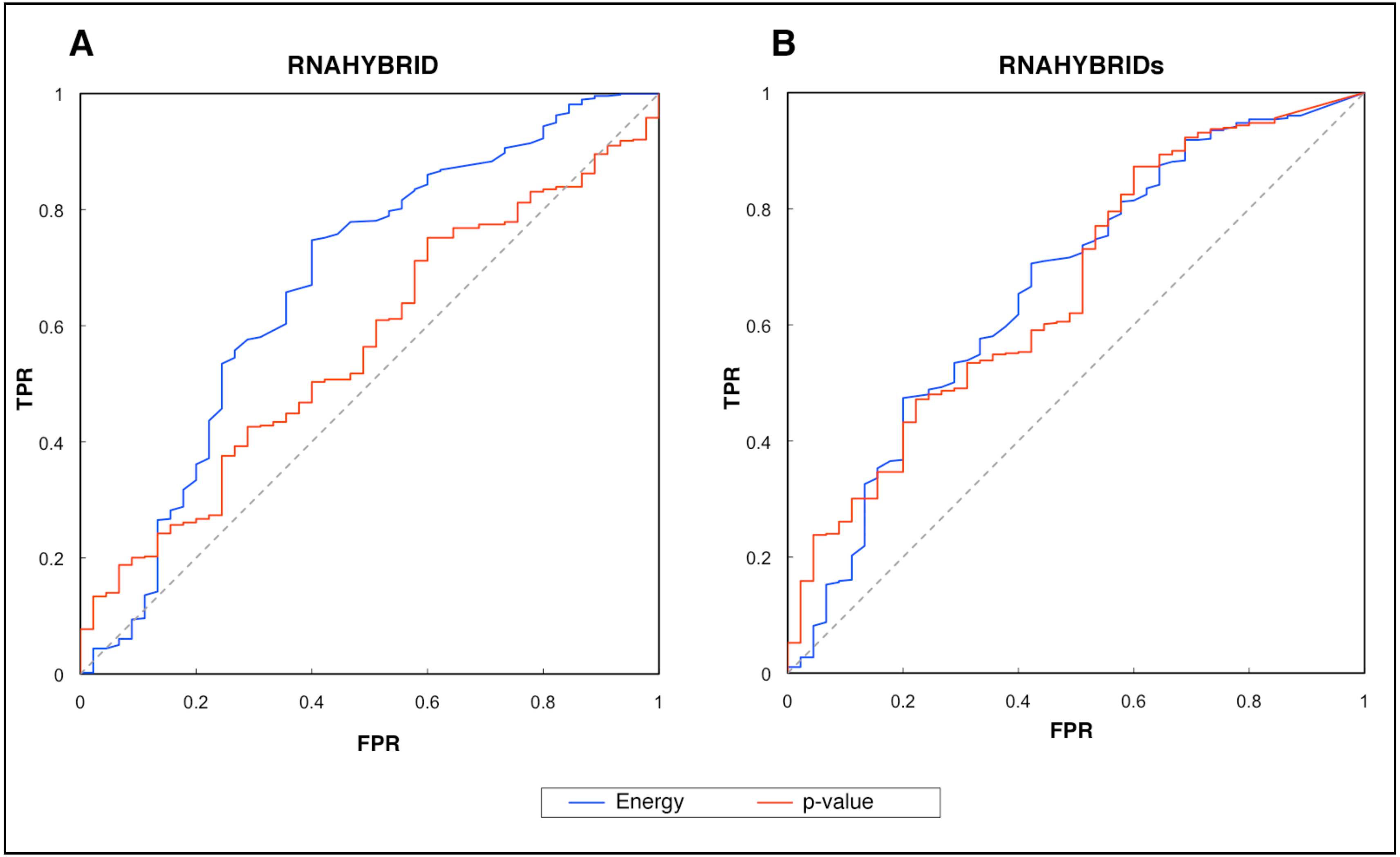


Figure 3.10: ROC graphs for RNAhybrid score type selection.

II. Generating a ROC Graph

Taking processed results from all miRNA-target prediction programs, StAR input was prepared using the Perl script PrepRocStarInput.pl. This included miRanda 'Score' results, RNAhybrid 'Energy' and 'p-value' results, and TargetSpy 'Energy' results. A ROC graph was created using the StAR web tool (Vergara *et al.* 2008), including 17 ROC curves corresponding to 15 sets of miRNA-target predictions for 9 miTP methods. (See Figure 3.11.)

For each miRNA-target prediction method, predictions were compared to the high-quality validated miRNA-target interaction dataset. The ROC curve data files output by StAR were input to the Perl script PlotStarRocCurveChart.pl, which plotted the final ROC graph in Microsoft Excel format using Perl modules Spreadsheet::WriteExcel (McNamara 2012), OLE::Storage_Lite (Takanori and McNamara 2012) and Parse::RecDescent (Braun and Conway 2012).

III. Generating a Recovery Graph

The processed results from all miRNA-target prediction programs were input to the Perl script CreateMitpRecoveryGraph.pl, which compared these to the positively validated miRNA-target interaction dataset. Using this input, this script creates a Microsoft Excel file displaying a miTP recovery graph, with the help of the Perl modules Spreadsheet::WriteExcel (McNamara 2012), OLE::Storage_Lite (Takanori and McNamara 2012) and Parse::RecDescent (Braun and Conway 2012).

For each miRNA-target prediction that has identified a known miRNA-target gene, a miTP recovery graph plots the number of validated miTIs (per miRNA) recovered by the miTP method against the total number of miRNA-target predictions (per miRNA); the recovery curve joining these points conveys the relationship between the number of miRNA targets predicted by each method and the number of known targets identified. As a point of comparison, a random predictor line is estimated using the formula shown in Equation 3.3.

$$L_{rand} = \frac{R_{\max}}{P_{\max}} = \frac{R_T/\mu}{P_T}$$

Equation 3.3: Estimated recovery rate of random microRNA-target predictor. Shown is the equation of the 'random predictor line' in a miTP recovery graph (L_{rand}), which reflects the performance of a miTP method predicting miRNA targets at random. This is equal to the ratio of the maximum number of miRNA targets that can be recovered per miRNA (R_{max}) to the maximum number of miRNA-target predictions that can be made per miRNA (P_{max}). The value of R_{max} is calculated as the ratio of the number of positively validated miRNA targets (R_T) — i.e. the number of 'recoverable' miRNA targets — to the number of unique miRNAs in the set of positively validated miRNA targets (μ). The value of P_{max} is taken as being equivalent to the number of annotated target 3′ UTR sequences being tested (P_T) — i.e. the maximum number of targets that can be predicted for any miRNA. It is not necessary to divide P_T by the number of miRNAs μ, since all annotated 3′ UTR sequences being tested are regarded, at least in principle, as potential targets of each miRNA.

The results of the TargetScan context and conservation scripts were merged in each recovery graph (forming the TSUNION recovery curve, see Figure 3.12 and Figure 3.13) to facilitate a fairer comparison between TargetScan and other methods.

3.4 Results

For the human data, a ROC graph was produced comparing miRNA-target predictions against known miRNA targets and non-targets (true- and false-positives, respectively). For each of the 6 benchmark test species, a recovery graph was produced showing the number of known miRNA targets recovered by each miTP method, graphed against the number of miTP predictions made. These results are described in the following sections.

3.4.1 MicroRNA-Target Prediction ROC Graph

The ROC graph in Figure 3.11 shows a comparison of the ranked output of 9 different miTP programs, while the bar graph displays the **area under the ROC curve** (AUC) for each miTP method. ROC curves can be used to compare method performance where the true-positive rate and false-positive rate can be estimated. The AUC of each method can provide a measure of the performance of that method, such that a higher AUC is associated with a higher true-positive rate and lower false-positive rate. The TargetScan context script has the highest AUC (i.e. 0.753), followed by the TargetScan conservation script (i.e. 0.723), and miRanda with and without a seed requirement (i.e. 0.716 and 0.71, respectively). The TargetScan context script has a significantly higher AUC than 9 of the bottom 10 ranked ROC curves, none of which had an AUC greater than 0.67 (Mann-Whitney *U* test, $P \leq 0.035$), see Table 3.7. The AUC values of the bottom 5 ranked ROC curves — none of which had an AUC higher than 0.639 — are significantly exceeded by those of the TargetScan conservation script (Mann-Whitney *U* test, $P \leq 0.00184$), miRanda *with* a strict seed requirement (Mann-Whitney *U* test, $P \leq 0.0113$), and miRanda *without* a strict seed (Mann-Whitney *U* test, $P \leq 0.0384$). Neither TargetScan nor miRanda has an AUC that is significantly better than those of miRmap or Hitsensor at a 5% significance level. This suggests that TargetScan is the miTP program best able to distinguish between miRNA targets and non-targets, followed closely by miRanda and more distantly by miRmap and Hitsensor. Having said that, all methods except MultiMiTar were found to have an AUC greater than that of a random predictor (where a random predictor is expected to have an AUC of 0.5).

With an area under the ROC curve of 0.461, MultiMiTar was the only miRNA-target prediction program with an AUC below that of a random classifier. This may be due to the low number of predictions output by MultiMiTar — nine involving the miTIs used in the ROC graph — so it may not be a good reflection of MultiMiTar's true performance. Similarly, although the TargetScan conservation script identified 44.7% (i.e. 214 of 479) of known miRNA targets while predicting no false-positives, to attain an AUC of 0.7234, this lack of false-positives can be seen as a disadvantage with respect to assessing its performance, since its ROC curve may not give a full picture of exactly how well it distinguishes miRNA targets from non-targets. The ROC curve data files are accessible in the electronic appendix at: Appendix / home / projects / benchmark / pipelines / mitp-benchmark / roc-curve.html.

3.4.2 MicroRNA-Target Prediction Recovery Graph

Figure 3.12 shows the overall recovery graph for human. For each miRNA-target prediction method, this graph plots the number of recovered known miRNA-target interactions against the number of miRNA-target predictions. The recovery graphs for mouse, rat and zebrafish broadly recapitulate those of human. There were a very small number of validated miTIs for cow and chicken, so the recovery graphs for these organisms was a step function. All six recovery graphs are accessible in the electronic appendix at the following file path: Appendix / home / projects / benchmark / pipelines / mitp-benchmark / validated-target-recovery-graph.html.

The recovery graph shows, for each miTP method, the number of known targets identified for a given number of miRNA-target predictions, or conversely, the number of predictions required to recover a given number of known miRNA targets. In a manner analogous to the ROC curve, a recovery curve with a larger area indicates a better performance.

Figure 3.11 Legend:

The figure on page 240 overleaf shows (A) results of a ROC analysis comparing 17 ROC curves corresponding to 15 sets of miRNA-target predictions by 9 miRNA-target prediction methods; and (B) a ranking of AUC values for the ROC curves. The colour of the bar for each ROC curve in the AUC ranking matches the colour of the corresponding ROC curve in the ROC graph. Given a set of ranked predictions by a miRNA-target prediction method and a set of known miRNA targets and non-targets, the true-positive and false-positive rates can be estimated for different levels of stringency and plotted on a curve; this produces a receiver operating characteristic (ROC) curve (Egan 1975). The area under the ROC curve (AUC) of each miTP method can then provide a measure of the performance of the method, such that a higher AUC is associated with a higher true-positive rate and lower false-positive rate. A prediction method that identifies all true-positives and rejects all true-negatives will have an AUC of 1, while a random classifier will have an AUC of 0.5 (corresponding to the dashed diagonal line in the ROC graph). In order of decreasing AUC values, the ROC curves represent miRNA-target predictions from: the TargetScan context script (TSCONTEXT), the TargetScan conservation script (TSCONSERV), miRanda with strict seed binding (MIRANDAs), miRanda without strict seed (MIRANDA), Hitsensor with increased weighting for seed match miRNA-target sites (HITSENSORs), miRmap (MIRMAP), Hitsensor with default settings (HITSENSOR), PITA with default settings (PITA), RNAhybrid predictions ranked by MFE (RNAHYBRID-E), RNAhybrid predictions constrained by seed binding and ranked by MFE (RNAHYBRIDs-E), RNAhybrid predictions constrained by seed binding and ranked by p-value (RNAHYBRIDs-P), TargetSpy with default settings (TARGETSPY), TargetSpy with seed match requirement (TARGETSPYs), PITA accounting for target site flanking sequence (PITAf), MicroTar (MICROTAR), RNAhybrid predictions ranked by p-value (RNAHYBRID-P) and MultiMiTar (MULTIMITAR).

Table 3.7 Legend:

Shown on page 241 is a table summarising statistical analysis performed on the miTP benchmark ROC graph, taken from the output of the StAR web utility (Vergara *et al.* 2008). (For the benchmark run names, see Table 3.6. For the AUC values, see Table 3.8.) The values in the upper triangle show the difference in AUC between miTP methods (ΔAUC), while the values in the lower triangle give the p-value of the ΔAUC in a Mann-Whitney *U* test. Those p-values less than 5% are shown in green, while p-values greater than 5% are shown in red. Note that because the AUC of MultiMiTar was less than 0.5, results are given for an inverted ROC curve for this miTP program.

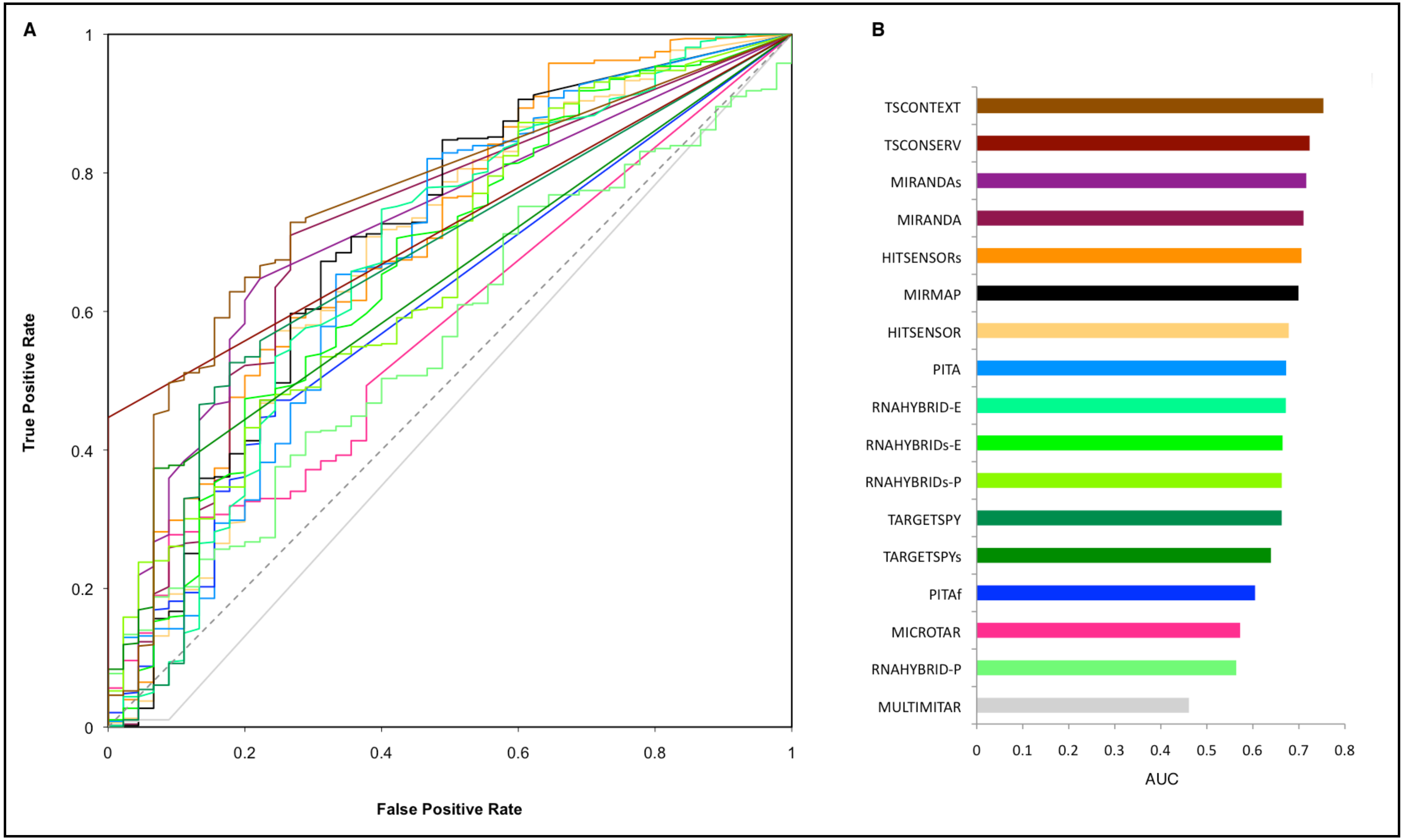


Figure 3.11: Benchmark ROC graph of 9 miTP programs.

Table 3.7: ROC StAR statistical analysis.

	TSCONTEXT	TSCONSERV	MIRANDAs	MIRANDA	HITSENSORs	MIRMAP	HITSENSOR	PITA	RNAHYBRID-E	RNAHYBRIDs-E	RNAHYBRIDs-P	TARGETSPY	TARGETSPYs	PITAf	MICROTAR	RNAHYBRID-P	MULTIMITAR
TSCONTEXT	—	0.0299	0.03716	0.04322	0.04765	0.05414	0.07532	0.08103	0.08161	0.08868	0.09047	0.09058	0.11401	0.14843	0.18103	0.18938	0.21424
TSCONSERV	0.39146	—	0.00726	0.01331	0.01775	0.02424	0.04542	0.05113	0.05170	0.05878	0.06057	0.06068	0.08411	0.11853	0.15113	0.15948	0.18434
MIRANDAs	0.06844	0.84044	—	0.00605	0.01048	0.01698	0.03816	0.04386	0.04444	0.05152	0.05331	0.05342	0.07685	0.11127	0.14387	0.15222	0.17708
MIRANDA	0.09468	0.74871	0.71942	—	0.00443	0.01093	0.03210	0.03781	0.03839	0.04547	0.04725	0.04737	0.07080	0.10522	0.13781	0.14616	0.17103
HITSENSORs	0.30021	0.69441	0.80178	0.91430	—	0.00650	0.02767	0.03338	0.03396	0.04103	0.04282	0.04294	0.06637	0.10079	0.13338	0.14173	0.16660
MIRMAP	0.32842	0.6062	0.74854	0.83452	0.83914	—	0.02118	0.02688	0.02746	0.03454	0.03633	0.03644	0.05987	0.09429	0.12689	0.13524	0.16010
HITSENSOR	0.10897	0.34527	0.38780	0.45889	0.20812	0.57369	—	0.00571	0.00629	0.01336	0.01515	0.01526	0.03869	0.07312	0.10571	0.11406	0.13892
PITA	0.03532	0.29104	0.24657	0.30019	0.35821	0.55683	0.88029	—	0.00058	0.00765	0.00944	0.00956	0.03299	0.06741	0.10000	0.10835	0.13322
RNAHYBRID-E	0.06786	0.29288	0.26673	0.35057	0.47967	0.64090	0.90018	0.98902	—	0.00707	0.00886	0.00898	0.03241	0.06683	0.09942	0.10777	0.13264
RNAHYBRIDs-E	0.01352	0.20332	0.12031	0.17675	0.36879	0.53627	0.77739	0.83210	0.81190	—	0.00179	0.00190	0.02533	0.05975	0.09235	0.10070	0.12556
RNAHYBRIDs-P	0.01240	0.17051	0.18230	0.21855	0.42985	0.53850	0.78754	0.84377	0.86058	0.95261	—	0.00012	0.02354	0.05797	0.09056	0.09891	0.12378
TARGETSPY	0.03195	0.12118	0.15587	0.21209	0.21290	0.39173	0.63480	0.81520	0.84294	0.96634	0.99827	—	0.02343	0.05785	0.09044	0.09879	0.12366
TARGETSPYs	0.00115	0.00170	0.01129	0.03835	0.08457	0.14804	0.33606	0.45370	0.50402	0.56654	0.61075	0.39514	—	0.03442	0.06701	0.07537	0.10023
PITAf	0.00001	0.00184	0.00096	0.00225	0.00966	0.05057	0.06561	0.01926	0.08699	0.11038	0.20237	0.08210	0.34400	—	0.03259	0.04094	0.06581
MICROTAR	0.00001	0.00010	0.00150	0.00581	0.02226	0.04302	0.07955	0.07340	0.05932	0.05538	0.05811	0.10198	0.17319	0.46951	—	0.00835	0.03322
RNAHYBRID-P	0.00006	0.00026	0.00248	0.00502	0.03580	0.05231	0.09897	0.07695	0.02633	0.05743	0.04221	0.10504	0.15981	0.41595	0.86794	—	0.02487
MULTIMITAR	0.00002	0.00000	0.00032	0.00137	0.00145	0.00166	0.01096	0.02414	0.03055	0.03615	0.02716	0.00590	0.00165	0.16611	0.47709	0.64035	—

As can be seen from Figure 3.12, TargetScan again performs well compared to other miRNA-target prediction methods, having a relatively high area under its curve despite low overall sensitivity — see the recovery curves for **TSCONSERV** and **TSCONTEXT** — the TargetScan conservation and context scripts, respectively — to the left of Figure 3.12. (Each miTP method is named here in the same colour as its corresponding recovery curves in Figure 3.12 and Figure 3.13. For information on the benchmark run names, see Table 3.6.) This was followed by **MIRANDAs** and **MIRANDA**, then **TARGETSPYs** and **TARGETSPY**, **MIRMAP**, **HITSENSORs** and **HITSENSOR**, then other methods. Overall, RNAhybrid predictions perform quite poorly as a binary classifier, such that its recovery curves straddle the dashed random prediction line — with the notable exception of **RNAHYBRIDs-P** (i.e. RNAhybrid with a seed binding constraint, ranked by p-value), which has a performance that is better than **PITA**, although not as good as **PITAf**.

Figure 3.13 shows the same recovery graph as Figure 3.12, but focusing on the top 500 ranked predictions for each method. These would be the predictions used most as they would be considered most enriched with *bona fide* miRNA-target interactions. This emphasises the difference in performance between TargetScan and other methods for top ranked miRNA-target predictions. Note the recovery curve of the TargetScan conservation script (i.e. **TSCONSERV**) above all others. This is followed by the **TSUNION** recovery curve — which unifies the results of the two TargetScan scripts — and then by the TargetScan context script (i.e. **TSCONTEXT**). This shows that, at least for the set of known validated miRNA-target interactions used here, TargetScan is better at identifying known miTIs among its top predictions than the other methods being compared.

Figure 3.12 Legend:

The figure on page 244 overleaf shows a graph of known miRNA-target interactions recovered as a function of the number of miRNA-target predictions. Except for MicroTar (which was included only in the ROC analysis), each miRNA-target prediction run in Figure 3.11 is represented in this graph in the same colour. An additional curve is included in this figure: TSUNION — the averaged results of TSCONSERV and TSCONTEXT — created to allow the results of TargetScan as a whole to be compared to those of other methods. Recovery curves for RNAhybrid include: RNAhybrid predictions ranked by MFE (RNAHYBRID-E), RNAhybrid predictions ranked by p-value (RNAHYBRID-P), RNAhybrid predictions constrained by seed binding and ranked by MFE (RNAHYBRIDs-E), and RNAhybrid predictions constrained by seed binding and ranked by p-value (RNAHYBRIDs-P). For all other recovery curve names, see Table 3.6. The dashed grey line reflects the expected performance of a random predictor.

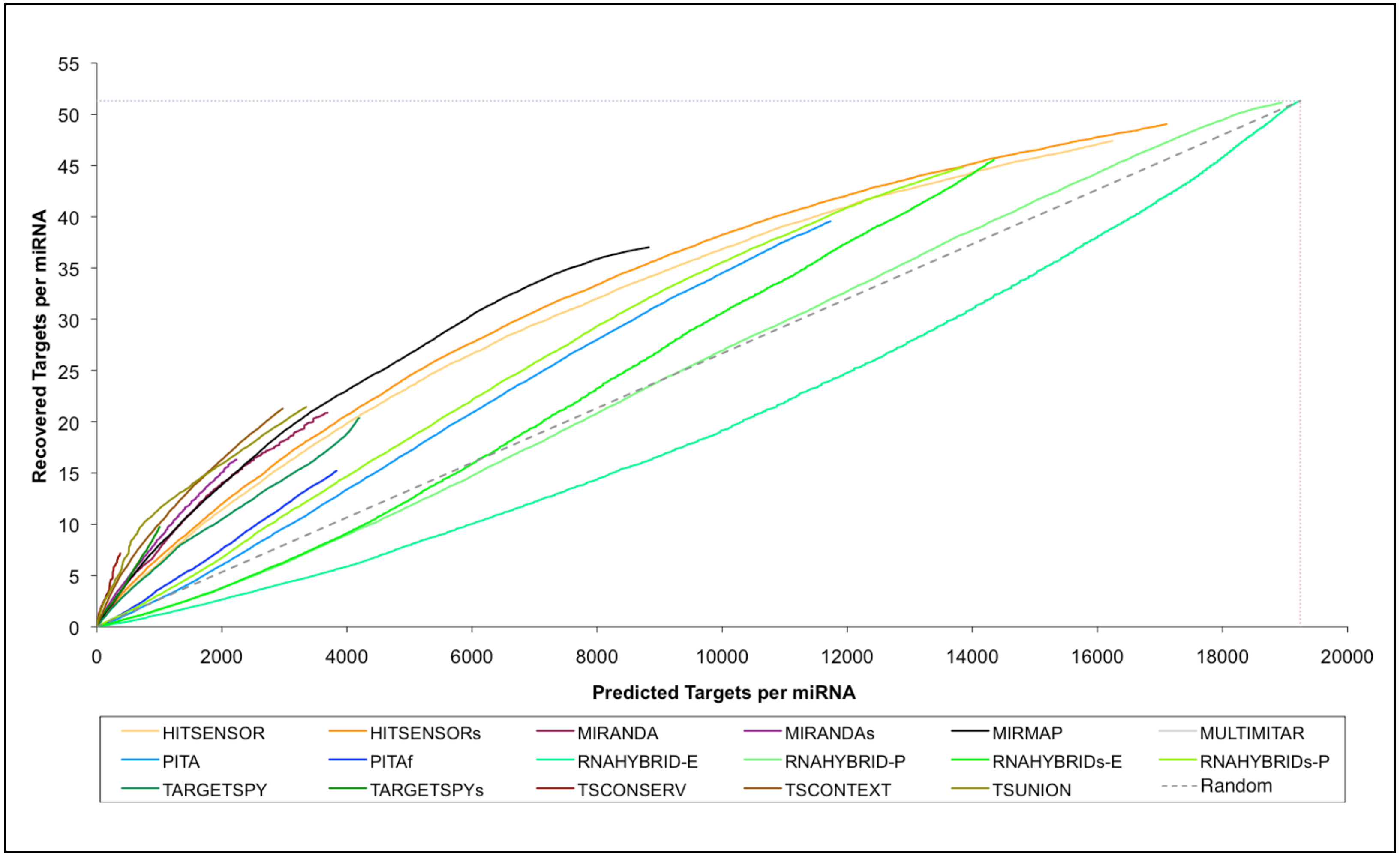


Figure 3.12: Benchmark recovery rates of 8 miTP programs.

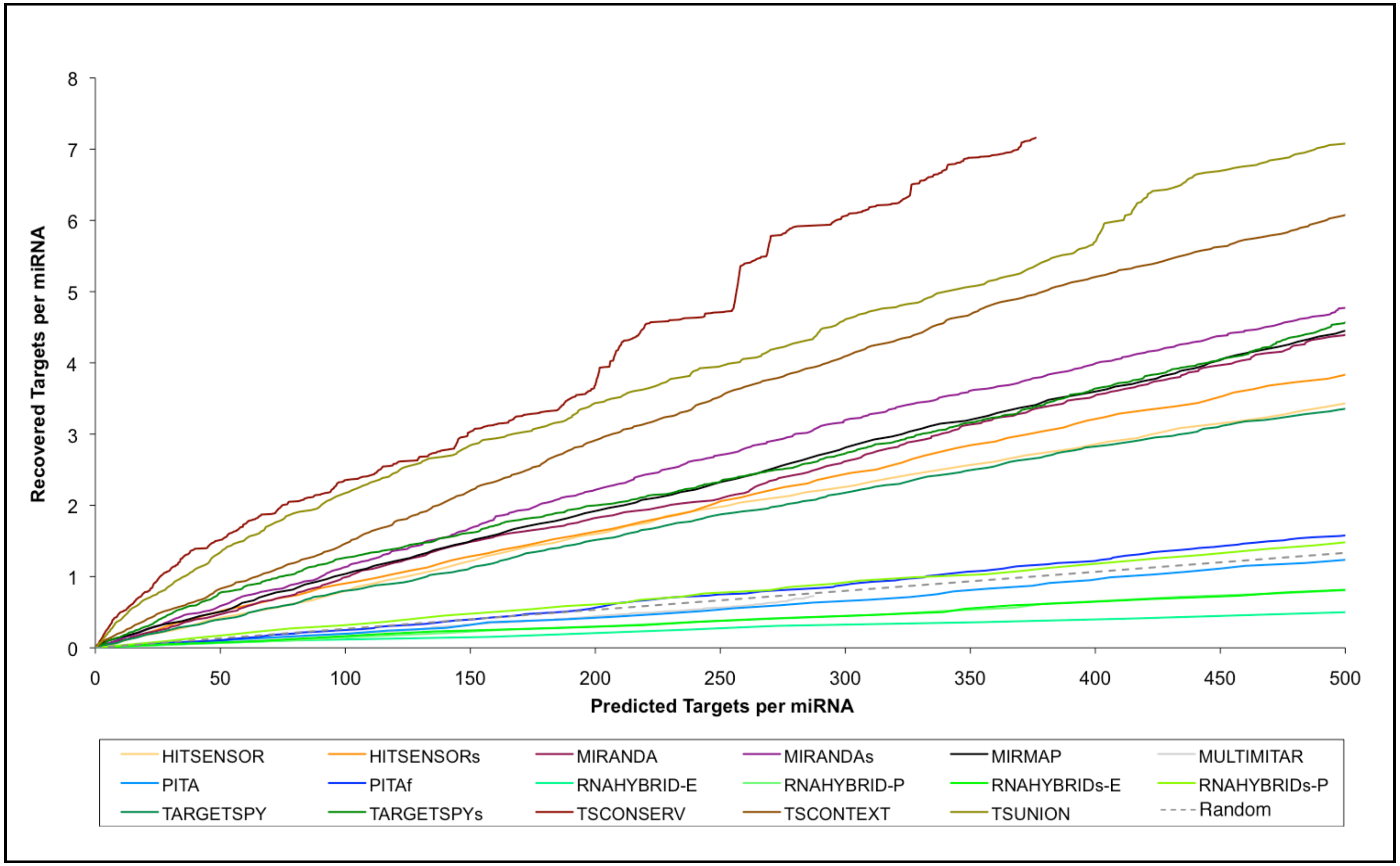


Figure 3.13: Benchmark recovery rates among top-ranking microRNA-target predictions.

Shown is a graph of known miRNA-target interactions recovered as a function of the number of miRNA-target predictions, among the top-ranking predictions of 8 of the methods benchmarked (MicroTar having been excluded). Recovery curves are colour-coded as in Figure 3.12.

Table 3.8: Summary of miTP benchmark results.

	ROC Graph	Recovery Graph	
	AUC	Sensitivity (Top 500)	Sensitivity (Overall)
HITSENSOR	0.6780	6.69%	92.39%
HITSENSORs	0.7056	7.47%	95.61%
MICROTAR	0.5723	N/A	N/A
MIRANDA	0.7101	8.57%	40.70%
MIRANDAs	0.7161	9.30%	31.78%
MIRMAP	0.6991	8.68%	72.15%
MULTIMITAR	0.4610	1.42%	1.42%
PITA	0.6723	2.42%	77.09%
PITAf	0.6048	3.07%	29.65%
RNAHYBRID-E	0.6717	0.96%	99.99%
RNAHYBRID-P	0.5639	1.56%	99.66%
RNAHYBRIDs-E	0.6646	1.58%	88.79%
RNAHYBRIDs-P	0.6628	2.89%	87.34%
TARGETSPY	0.6627	6.54%	39.68%
TARGETSPYs	0.6393	8.89%	19.00%
TSCONSERV	0.7234	13.96%	13.96%
TSCONTEXT	0.7533	11.84%	41.47%
TSUNION	N/A	13.79%	41.79%

Table 3.8 Legend:

Shown is a summary of the key results for the miTP benchmark. The **AUC** column shows the area under the ROC curve attained by each method in the ROC analysis. The **Sensitivity (Top 500)** column shows the percentage of known miRNA targets recovered by each miTP method in the top 500 miRNA-target predictions. The **Sensitivity (Overall)** column shows the percentage of known targets recovered overall by each miRNA-target prediction method. Benchmark runs shown for RNAhybrid include: RNAhybrid predictions ranked by MFE (RNAHYBRID-E), RNAhybrid predictions ranked by p-value (RNAHYBRID-P), RNAhybrid predictions constrained by seed binding and ranked by MFE (RNAHYBRIDs-E), and RNAhybrid predictions constrained by seed binding and ranked by p-value (RNAHYBRIDs-P). As in the recovery graphs of Figure 3.12 and Figure 3.13, the results of TSCONSERV and TSCONTEXT were averaged to produce a TSUNION curve, which allows the results of TargetScan as a whole to be compared to those of other methods. For other benchmark run names, see Table 3.6.

3.5 Discussion

With the caveat that the results of this benchmark study are subject to the data and methods used, the general ranking of miRNA-target prediction methods was consistent across both the ROC and recovery analyses. TargetScan and miRanda performed best overall, attaining a higher AUC than other methods in ROC analysis, and recovering more known miRNA targets than other methods studied, for a given number of predictions. However, both methods were relatively limited in their sensitivity: in terms of known positive miRNA targets recovered, TargetScan had sensitivity of 41.79%, while miRanda had sensitivity of 40.7%. Of the more sensitive methods overall, miRmap and Hitsensor performed best, with overall sensitivity of 72.15% and 95.61%, respectively, and also high sensitivity among top ranking miRNA-target predictions (i.e. 8.68% for miRmap and 7.47% for Hitsensor). However, this sensitivity came at some cost to specificity, such that in ROC analysis, miRmap had an AUC of 0.6991, while Hitsensor (with increased weighting for seed match sites) attained an AUC of 0.7056. Nevertheless, these two methods had sensitivity comparable to miRanda among the top-ranked miRNA-target predictions (see Figure 3.13 and Table 3.8).

TargetSpy and MultiMiTar — the miTP methods based on machine learning — did not perform very well overall. TargetSpy (without seed constraint) attained a reasonably good AUC of 0.6627 in ROC analysis, but attained relatively low sensitivity of 39.68%. MultiMiTar performed considerably worse, having an AUC of 0.461 (worse than random) in ROC analysis and maximum sensitivity of 1.42% in the recovery graph. Both methods produced a relatively small number of miRNA-target predictions: approximately 4000 and 288 per miRNA, respectively, compared to an average of about 9000 miTPs per miRNA among the other benchmarked methods. Whether the low number of predictions is an issue with machine learning *per se* is debatable; perhaps a stringent cutoff is used by these methods because the accuracy of the machine learning algorithm was found to deteriorate beyond a certain point. It could be argued, at least in some instances, that a reduction in sensitivity is acceptable if the miRNA-target predictions that are made have a high degree of accuracy. This might be the case for TargetSpy, which achieved a high AUC considering its relatively low sensitivity and, at the limit of its sensitivity, recovered almost as many known miRNA targets as the TargetScan context script, (note the short green line for **TARGETSPYs** in the bottom left of Figure 3.12). However, the same could not be said for MultiMiTar, which achieved neither a high AUC nor appreciable sensitivity.

The miRNA-target prediction methods based primarily on thermodynamic considerations — i.e. PITA, RNAhybrid and MicroTar — performed relatively poorly at the binary classification of miRNA-target interactions. Among these methods, PITA performed best, with an AUC of up to 0.6723 in ROC analysis and sensitivity levels of 3.07% (with flank) and 2.42% (without flank) among the top 500 miRNA-target predictions. RNAhybrid had a maximum AUC of 0.6717 in the ROC graph, but tracked the random predictor line in the recovery graph, with a maximum sensitivity of 2.89% among the top 500 predictions, when seed constrained miRNA-target predictions were ranked by increasing p-value. MicroTar had the worst performance of the three: due to its prohibitive computational cost (e.g. in extreme cases, it could take more than 12 hours to test a single miRNA-target interaction), it was excluded from the main benchmark study and only tested against the comparatively small number of high-quality validated miTIs used to construct the ROC graph, achieving an AUC of 0.5723. The inclusion of accessibility in the thermodynamic calculations of PITA may account for its improved performance over RNAhybrid, which focuses only on the hybridisation energy of the miRNA-target duplex. MicroTar similarly focuses on the miRNA-target duplex, but models the folding of the target mRNA both before and after hybridisation; this thoroughness is associated with considerable computational cost, but is unfortunately not rewarded with improved identification of miRNA-target interactions.

In the recovery graph in Figure 3.12, the differences in performance of RNAhybrid miRNA-target predictions under different score types and settings can be seen. This recovery graph shows that RNAhybrid predictions ranked by increasing MFE (i.e. RNAHYBRID-E and RNAHYBRIDs-E in Figure 3.12) are a poor binary classifier of miRNA-target interactions, with a recovery performance at or below that of a random classifier. In the 'Top-500' predictions, RNAHYBRID and RNAHYBRIDs, as ranked by MFE, attained sensitivity levels of only 0.96% and 1.58%, respectively.

This compares unfavourably with an average of 6.09% sensitivity among all miTP methods benchmarked. Predictions of RNAHYBRID ranked by p-value (i.e. RNAHYBRID-P in Figure 3.12) had 'Top-500' sensitivity of 1.56%, reflecting a generally poor performance by RNAhybrid. However, one exception was observed: when a miRNA seed binding constraint is imposed (i.e. RNAHYBRIDs-P in Figure 3.12), the overall performance of RNAhybrid improves considerably and with 'Top-500' sensitivity of 2.89%, is comparable with that of PITA.

Whether the observed differences in performance of miTP methods are statistically significant is a non-trivial question. Fortunately, the StAR web utility (Vergara *et al.* 2008) integrates statistical analysis of ROC curves, implementing the method described by Delong *et al.* (1988) for comparing the AUC values of different ROC curves using a Mann-Whitney *U* test. The results of this statistical analysis for the ROC graph in Figure 3.11 are shown in Table 3.7. These indicate that TargetScan has a significantly higher AUC than the majority of methods, including RNAhybrid, PITA, TargetSpy, MicroTar and MultiMiTar ($0.081 \leq \Delta AUC \leq 0.214$, $\alpha = 0.05$). Similarly, miRanda has significantly higher AUC than TargetSpy, MicroTar and MultiMiTar, as well as PITA and RNAhybrid under some settings ($0.071 \leq \Delta AUC \leq 0.171$, $\alpha = 0.05$). Notably, neither TargetScan nor miRanda have a performance that is significantly better than miRmap or Hitsensor at a 5% significance level (see Table 3.7). This may be due in part to the relatively low number of high-quality positively and negatively validated miRNA-target interactions — with 479 positively validated miTIs and 45 negatively validated miTIs, there were only 524 high-quality validations in the dataset. Perhaps with a larger set of such high-quality miRNA-target validations, statistical analysis of ROC graphs might allow clearer distinctions to be drawn between the performance of different miRNA-target prediction methods.

Aside from its ability to identify and characterise miRNA-target interactions, other aspects of miRNA-target prediction software may be of importance. The development model of miRmap may present a new way of tackling the problem of choosing which miTP method to use from a selection of disparate methods, by incorporating the features of different miTP methods within one open source Python package. Whether this model will become widely used remains to be seen, but does depend on continuing active development by the authors of the miRmap Python package (Vejnar and Zdobnov 2012) and by its community of users. Python is (arguably) highly suited to open source development, since its design philosophy emphasises readability and maintainability. Without such active maintenance, the feature set implemented in miRmap will inevitably lag behind those used by other actively developed miRNA-target prediction software. For example, the miRmap package at time of publication by Vejnar and Zdobnov (2012) lacked the target abundance and seed-pairing stability features implemented in TargetScan by Garcia *et al.* (2011). Furthermore, newly discovered miRNA-target site types such as bulge sites (Chi *et al.* 2012) and centred pairing sites (Shin *et al.* 2010) are yet to be included in miRmap or indeed any miRNA-target prediction method.

One aspect of concern is the narrow focus of current miRNA-target prediction methods on identifying miRNA-target sites within the 3′ UTR sequence; it has been suggested that this may have become a self-reinforcing bias (Rigoutsos 2009, Peter 2010). Indeed, the lack of validated miRNA-target sites localised to the coding region of the target gene effectively prevented miRNA-target prediction within the CDS from being considered within the benchmark study in this chapter. Recent developments will at least alleviate concerns of a bias towards target prediction within 3′ UTRs, with the publication of methods that identify miRNA targets within coding regions, such as DIANA-microT-CDS (Reczko *et al.* 2012) and more recently PACCMIT-CDS (Marín *et al.* 2013). This will perhaps allow more miRNA-target sites to be identified within the coding regions of potential target genes, and in turn for such target sites to be validated as functional, since many miTI validation experiments are guided by preliminary miRNA-target predictions.

The use of miRNA-target predictions to identify candidate targets for validation in itself can pose a problem, since this biases the set of validated miRNA-target interactions towards those miTP methods that are most widely used (Sturm *et al.* 2010). To some extent such circularity is unavoidable, although large-scale miRNA-target validation studies such as those of Baek *et al.* (2008) or Chi *et al.* (2009) can ameliorate this problem.

The greatest strength of this benchmark — comparison of miRNA-target prediction methods on a consistent dataset — could also be seen as its most limiting weakness, since it effectively excluded those miTP methods for which standalone software is not currently available. This ruled out a comparison of the software for EIMMo, PicTar and DIANA-microT, all of which performed well in previous benchmarks (Sethupathy *et al.* 2006, Baek *et al.* 2008, Selbach *et al.* 2008, Alexiou *et al.* 2009), and none of which are available to download. It also prevented an assessment of miRanda predictions ranked using mirSVR, which has been reported to improve the ranking of miRNA-target predictions produced by miRanda (Betel *et al.* 2010), but which has yet to be integrated into miRanda proper. A benchmark study of miRNA-target prediction methods remains incomplete without including an assessment of all current miTP methods, and the prospect of such a benchmark awaits a time when the release of standalone software becomes the norm among the population of miRNA-target prediction methods. Nevertheless, incomplete as it is, the benchmark described here has distinguished those methods that achieve the best balance between precision and sensitivity (i.e. TargetScan and miRanda) from those that lack either precision (e.g. RNAhybrid), sensitivity (e.g. TargetSpy), or both.

Chapter 4: Prediction of Placenta-Specific MicroRNA-Target Interactions

4.1 Introduction

Chapter 2 focused on the role of functional shift of placental genes in the evolution of the placenta. However, this inevitably produces an incomplete picture of evolutionary change in relation to the emergence of placenta, since there are numerous ways in which such novelty can arise (Hughes and Friedman 2010). Morphological innovation introduced by genetic novelty is frequently mediated by changes in the regulatory regime controlling expression of genes (Davidson and Erwin 2006).

MicroRNAs are a potentially valuable source of information on the evolution of gene expression regulation, since they more or less continually emerge during evolution, tend to be highly conserved and undergo secondary loss only rarely (Wheeler *et al.* 2009). MiRNAs have been shown to regulate a broad range of biological processes — including insulin secretion (Poy *et al.* 2004), B-cell differentiation (Chen 2004a) and differentiation of adipocytes (Esau *et al.* 2004) — and are believed to play a role in the maintenance of tissue identity (Chang *et al.* 2004, Girard *et al.* 2008), and more broadly in the evolution of vertebrate complexity (Heimberg *et al.* 2008, Wheeler *et al.* 2009, Peterson *et al.* 2009a).

MiRNAs similarly play a key role in placental development. Many miRNAs are expressed specifically or exclusively in the placenta (Liang *et al.* 2007, Landgraf *et al.* 2007), and placental miRNA expression patterns differ over the gestation period (Luo *et al.* 2009), perhaps reflecting a developmental program. MiRNAs have been shown to play a role in the regulation of trophoblast proliferation (Chao *et al.* 2010, Luo *et al.* 2012) and the formation of capillaries at foetal-maternal interface (Sekita *et al.* 2008), while dysregulation of miRNAs has been observed in cases of pre-term delivery (Mayor-Lynn *et al.* 2011) and notably in pre-eclampsia (Poliseno *et al.* 2006, Pineles *et al.* 2007, Hu *et al.* 2009, Mayor-Lynn *et al.* 2011).

For example, the gene retrotransposon-like 1 (RTL1) is critical for the development of placenta in a manner sensitive to its expression levels (Kotzot 2004, Sekita *et al.* 2008). Both over-expression and under-expression of RTL1 are associated with deleterious phenotypes affecting the growth and development of the young in human (Kotzot 2004) and in mouse (Sekita *et al.* 2008). Several miRNAs are transcribed from the antisense strand of the RTL1 locus (Seitz *et al.* 2003); these miRNAs regulate the expression of their host gene, maintaining its expression levels temporally within the optimal range (Davis *et al.* 2005).

Another example of the influence of miRNA in placenta is furnished by the chromosome 19 miRNA cluster (C19MC) first identified by Bentwich *et al.* (2005); this remains the largest miRNA cluster observed in human (Flor and Bullerdiek 2012). The C19MC cluster comprises 46 miRNA precursors, many of which share similarities with the neighbouring cluster of miRNAs containing mir-371, mir-372 and mir-373. This similarity has led to suggestions that the C19MC cluster may have originated in a duplication of one of these three neighbouring miRNAs (Bentwich *et al.* 2005, Zhang *et al.* 2008).

MiRNAs homologous to those in the C19MC have not been identified in any non-primate genome, and it is believed to be a primate-specific cluster (Bentwich *et al.* 2005, Zhang *et al.* 2008). A clue to its rapid expansion in the primate lineage has been offered by the presence of Alu elements (Batzer and Deininger 2002), interspersed among the miRNAs in C19MC; Alu-mediated expansion has been suggested to account for the emergence of such a large number of miRNAs in a comparatively short period of time (Zhang *et al.* 2008, Lehnert *et al.* 2009).

There is some debate as to whether the members of C19MC are transcribed by RNA polymerase II or III (Borchert *et al.* 2006, Bortolin-Cavaille *et al.* 2009), but there is little controversy about where it is expressed. Expression of C19MC miRNAs has been observed in placenta (Bentwich *et al.* 2005, Liang *et al.* 2007) and particularly in embryonic stem cells (Bar *et al.* 2008, Cao *et al.* 2008, Ren *et al.* 2009), but has not been observed in adult tissues (Liang *et al.* 2007, Landgraf *et al.* 2007). The C19MC cluster is genomically imprinted, such that only the paternally inherited gene is expressed in placenta (Tsai *et al.* 2009, Noguer-Dance *et al.* 2010), while both alleles are methylated (and therefore not expressed) in normal adult tissues such as muscle or brain (Noguer-Dance *et al.* 2010). However, despite considerable progress in determining its evolutionary history and expression patterns, relatively little is known about the function of the miRNAs in the C19MC cluster, although for example, 23 members of this cluster have been shown to have a tumour suppressive function, an oncogenic role, or both (Flor and Bullerdiek 2012).

Taken together, these studies demonstrate that miRNAs have played a key role in the evolution and development of placenta through the regulation of their target genes. In this chapter we have performed evolutionary analyses of miRNA-target interactions for a set of 114 placental genes (from Chapter 2) and 141 placental miRNAs assembled specifically for this chapter. For each placental miRNA and gene, miRNA-target prediction was performed in 20 different species, to identify miRNA-target pairs that have a predicted functional interaction that is conserved across species and to assess the prevalence of such interactions and the nature of their emergence, whether they emerge soon after the emergence of the miRNA itself or more slowly over time.

4.2 Materials

A set of placental miRNAs was assembled, including those miRNAs known to be important to the development and/or function of the placenta, and those miRNAs with placenta-specific expression patterns. The placental genes analysed in Chapter 2 were used again in this chapter. The species of interest in this chapter are also identical to those in Chapter 2, as the same requirements of good genome coverage, genome quality and taxon sampling remain. The genomic nucleotide sequence data for each species of interest was obtained from the Ensembl genomic database server (Flicek *et al.* 2012), while the sequences of placental miRNA and target gene 3′ UTRs were downloaded from Ensembl through the BioMart interface (Smedley *et al.* 2009). Placental miRNA sequences were obtained from a local copy of the miRBase database (Griffiths-Jones 2004, Griffiths-Jones 2006, Griffiths-Jones *et al.* 2008, Kozomara and Griffiths-Jones 2011). All scripts used during this process can be accessed from the following location in the electronic appendix: Appendix / home / code / scripts.html.

4.2.1 Placental MicroRNA Set Assembly

The set of placenta-specific miRNAs is composed of miRNAs from seven different data sources. Each placental miRNA has a placenta-specific expression pattern, or its dysregulation is associated with a disease state. Each data source is listed in Table 4.1. Before being merged into one non-redundant data set, each miRNA (whether precursor or mature miRNA) was mapped to a common standard nomenclature — in this case that of miRBase release 18. Where a study specified a precursor, all corresponding mature miRNAs that could not be ruled out were included. Correspondingly, where a study specified a mature miRNA, all corresponding miRNA precursors that could not be ruled out were included. The result was a set of 159 mature miRNAs with 141 corresponding miRNA precursors. Of the 141 precursor miRNAs, 127 were human miRNAs and 14 were mouse miRNAs. Information on each of the 141 precursor miRNAs is given in Table 4.2, and the process itself is summarised in Figure 4.1.

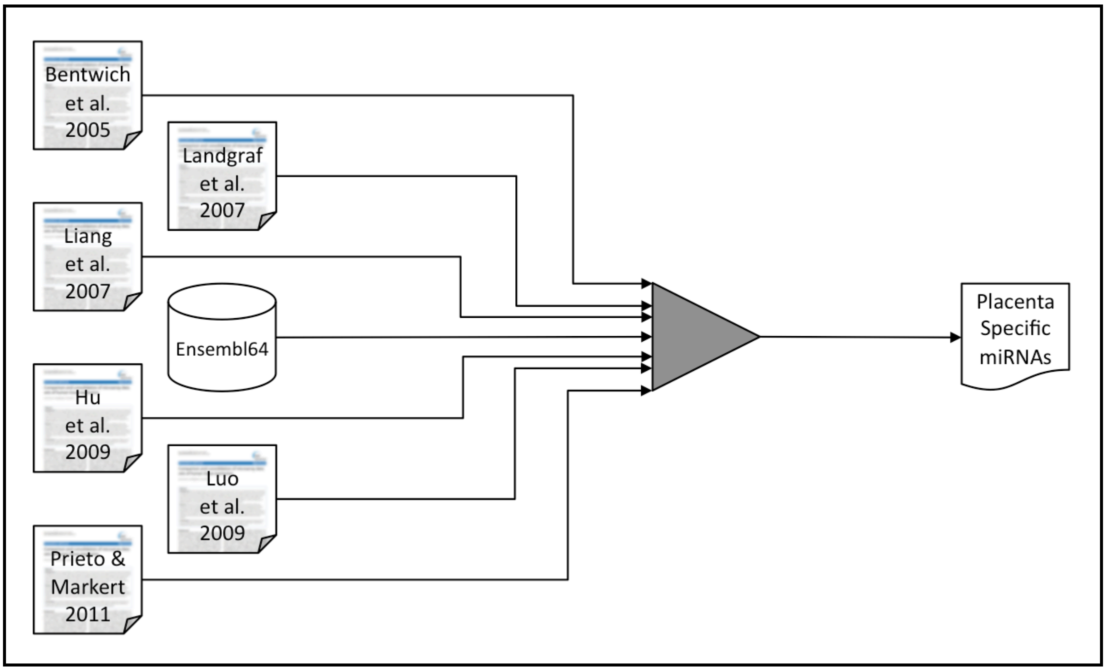


Figure 4.1: Placental microRNA set assembly pipeline.

Shown is a bioinformatics pipeline depicting the steps involved in placental miRNA set assembly. Arrows indicate the direction of process flow. For information on symbols used, see Figure 2.2.

Table 4.1: Placental microRNA dataset sources.

miRNA Dataset	# miRNAs	Details	References
Bentwich	43	Original discovery of chromosome 19 miRNA cluster (C19MC)	Bentwich *et al.* (2005)
Ensembl64 C19MC	46	C19MC miRNAs annotated in Ensembl release 64	Flicek *et al.* (2011)
Landgraf	8	Expression profile showing 100% normalised tissue enrichment in placenta	Landgraf *et al.* (2007)
Liang	51	MicroRNAs over-expressed in placenta; of which 40 are in C19MC	Liang *et al.* (2007)
Hu	7	MicroRNAs over-expressed in placenta of patients with pre-eclampsia	Hu *et al.* (2009)
Luo	72	High expression in placenta: in first trimester, at term or throughout gestation	Luo *et al.* (2009)
Prieto	14	MicroRNAs with functional effects in normal and disease states of placenta	Prieto and Markert (2011)

Table 4.2: Placenta-specific microRNAs.

miRNA ID	miRBase Accession	Ensembl Gene ID	Mature miRNAs
hsa-let-7a-1	MI0000060	ENSG00000199165	hsa-let-7a-5p
hsa-let-7a-2	MI0000061	ENSG00000198975	hsa-let-7a-5p
hsa-let-7a-3	MI0000062	ENSG00000198986	hsa-let-7a-5p
hsa-let-7b	MI0000063	ENSG00000207875	hsa-let-7b-5p
hsa-let-7d	MI0000065	ENSG00000199133	hsa-let-7d-3p
hsa-mir-1-2	MI0000437	ENSG00000207694	hsa-miR-1
hsa-mir-1-1	MI0000651	ENSG00000199017	hsa-miR-1
hsa-mir-15b	MI0000438	ENSG00000207779	hsa-miR-15b-5p, hsa-miR-15b-3p
hsa-mir-16-1	MI0000070	ENSG00000208006	hsa-miR-16-5p, hsa-miR-16-1-3p
hsa-mir-16-2	MI0000115	ENSG00000198987	hsa-miR-16-5p, hsa-miR-16-2-3p
hsa-mir-21	MI0000077	ENSG00000199004	hsa-miR-21-5p
hsa-mir-23a	MI0000079	ENSG00000207980	hsa-miR-23a-3p
hsa-mir-24-1	MI0000080	ENSG00000207617	hsa-miR-24-3p
hsa-mir-24-2	MI0000081	ENSG00000209707	hsa-miR-24-3p
hsa-mir-25	MI0000082	ENSG00000207547	hsa-miR-25-3p
hsa-mir-26a-1	MI0000083	ENSG00000199075	hsa-miR-26a-5p
hsa-mir-26b	MI0000084	ENSG00000199121	hsa-miR-26b-5p, hsa-miR-26b-3p
hsa-mir-26a-2	MI0000750	ENSG00000207789	hsa-miR-26a-5p
hsa-mir-27a	MI0000085	ENSG00000207808	hsa-miR-27a-3p
hsa-mir-27b	MI0000440	ENSG00000207864	hsa-miR-27b-3p
hsa-mir-29b-1	MI0000105	ENSG00000207748	hsa-miR-29b-3p, hsa-miR-29b-1-5p
hsa-mir-29b-2	MI0000107	ENSG00000207790	hsa-miR-29b-3p, hsa-miR-29b-2-5p
hsa-mir-30c-2	MI0000254	ENSG00000199094	hsa-miR-30c-5p
hsa-mir-30d	MI0000255	ENSG00000199153	hsa-miR-30d-5p
hsa-mir-30b	MI0000441	ENSG00000207582	hsa-miR-30b-5p
hsa-mir-30c-1	MI0000736	ENSG00000207962	hsa-miR-30c-5p
hsa-mir-30e	MI0000749	ENSG00000198974	hsa-miR-30e-5p
hsa-mir-31	MI0000089	ENSG00000199177	hsa-miR-31-5p
hsa-mir-34a	MI0000268	ENSG00000207865	hsa-miR-34a-5p
hsa-mir-34b	MI0000742	ENSG00000207811	hsa-miR-34b-5p, hsa-miR-34b-3p
hsa-mir-34c	MI0000743	ENSG00000207562	hsa-miR-34c-5p

Table 4.2: Placenta-specific microRNAs. (continued)

miRBase 18 ID	miRBase Accession	Ensembl Gene ID	Mature miRNAs
hsa-mir-92b	MI0003560	ENSG00000208011	hsa-miR-92b-3p
hsa-mir-93	MI0000095	ENSG00000207757	hsa-miR-93-5p, hsa-miR-93-3p
hsa-mir-99a	MI0000101	ENSG00000207638	hsa-miR-99a-5p
hsa-mir-99b	MI0000746	ENSG00000207550	hsa-miR-99b-5p, hsa-miR-99b-3p
hsa-mir-100	MI0000102	ENSG00000207994	hsa-miR-100-5p
hsa-mir-106b	MI0000734	ENSG00000208036	hsa-miR-106b-5p
hsa-mir-122	MI0000442	ENSG00000207778	hsa-miR-122-5p
hsa-mir-125b-1	MI0000446	ENSG00000207971	hsa-miR-125b-5p
hsa-mir-125b-2	MI0000470	ENSG00000207863	hsa-miR-125b-5p, hsa-miR-125b-2-3p
hsa-mir-140	MI0000456	ENSG00000208017	hsa-miR-140-5p
hsa-mir-141	MI0000457	ENSG00000207708	hsa-miR-141-3p
hsa-mir-143	MI0000459	ENSG00000208035	hsa-miR-143-3p
hsa-mir-151a	MI0000809	ENSG00000207792	hsa-miR-151a-3p
hsa-mir-154	MI0000480	ENSG00000207978	hsa-miR-154-5p, hsa-miR-154-3p
hsa-mir-181a-2	MI0000269	ENSG00000207595	hsa-miR-181a-5p, hsa-miR-181a-2-3p
hsa-mir-181a-1	MI0000289	ENSG00000207759	hsa-miR-181a-5p, hsa-miR-181a-3p
hsa-mir-184	MI0000481	ENSG00000207695	hsa-miR-184
hsa-mir-191	MI0000465	ENSG00000207605	hsa-miR-191-5p
hsa-mir-196b	MI0001150	ENSG00000207584	hsa-miR-196b-5p
hsa-mir-199a-1	MI0000242	ENSG00000207752	hsa-miR-199a-3p
hsa-mir-199a-2	MI0000281	ENSG00000208024	hsa-miR-199a-3p
hsa-mir-199b	MI0000282	ENSG00000207581	hsa-miR-199b-5p, hsa-miR-199b-3p
hsa-mir-200c	MI0000650	ENSG00000207713	hsa-miR-200c-3p
hsa-mir-205	MI0000285	ENSG00000207623	hsa-miR-205-5p, hsa-miR-205-3p
hsa-mir-210	MI0000286	ENSG00000199038	hsa-miR-210
hsa-mir-214	MI0000290	ENSG00000207949	hsa-miR-214-3p
hsa-mir-221	MI0000298	ENSG00000207870	hsa-miR-221-3p
hsa-mir-222	MI0000299	ENSG00000207725	hsa-miR-222-5p, hsa-miR-222-3p
hsa-mir-224	MI0000301	ENSG00000207621	hsa-miR-224-5p, hsa-miR-224-3p
hsa-mir-324	MI0000813	ENSG00000199053	hsa-miR-324-3p

Table 4.2: Placenta-specific microRNAs. (continued)

miRNA ID	miRBase Accession	Ensembl Gene ID	Mature miRNAs
hsa-mir-331	MI0000812	ENSG00000199172	hsa-miR-331-3p
hsa-mir-335	MI0000816	ENSG00000199043	hsa-miR-335-5p, hsa-miR-335-3p
hsa-mir-342	MI0000805	ENSG00000199082	hsa-miR-342-3p, hsa-miR-342-5p
hsa-mir-361	MI0000760	ENSG00000199051	hsa-miR-361-5p
hsa-mir-371a	MI0000779	ENSG00000199031	hsa-miR-371a-5p, hsa-miR-371a-3p
hsa-mir-372	MI0000780	ENSG00000199095	hsa-miR-372
hsa-mir-373	MI0000781	ENSG00000199143	hsa-miR-373-5p, hsa-miR-373-3p
hsa-mir-377	MI0000785	ENSG00000199015	hsa-miR-377-5p, hsa-miR-377-3p
hsa-mir-381	MI0000789	ENSG00000199020	hsa-miR-381
hsa-mir-424	MI0001446	ENSG00000199097	hsa-miR-424-5p, hsa-miR-424-3p
hsa-mir-449a	MI0001648	ENSG00000198983	hsa-miR-449a
hsa-mir-450a-1	MI0001652	ENSG00000199132	hsa-miR-450a-5p
hsa-mir-450a-2	MI0003187	ENSG00000207755	hsa-miR-450a-5p, hsa-miR-450a-3p
hsa-mir-451a	MI0001729	ENSG00000207794	hsa-miR-451a
hsa-mir-483	MI0002467	ENSG00000207805	hsa-miR-483-3p
hsa-mir-491	MI0003126	ENSG00000207609	hsa-miR-491-5p, hsa-miR-491-3p
hsa-mir-497	MI0003138	ENSG00000207791	hsa-miR-497-5p
hsa-mir-498	MI0003142	ENSG00000207869	hsa-miR-498
hsa-mir-503	MI0003188	ENSG00000208005	hsa-miR-503
hsa-mir-512-1	MI0003140	ENSG00000207645	hsa-miR-512-5p, hsa-miR-512-3p
hsa-mir-512-2	MI0003141	ENSG00000207644	hsa-miR-512-5p, hsa-miR-512-3p
hsa-mir-515-1	MI0003144	ENSG00000207616	hsa-miR-515-5p, hsa-miR-515-3p
hsa-mir-515-2	MI0003147	ENSG00000207615	hsa-miR-515-5p, hsa-miR-515-3p
hsa-mir-516b-2	MI0003167	ENSG00000207925	hsa-miR-516b-5p, hsa-miR-516b-3p
hsa-mir-516b-1	MI0003172	ENSG00000207946	hsa-miR-516b-5p, hsa-miR-516b-3p
hsa-mir-516a-1	MI0003180	ENSG00000207767	hsa-miR-516a-5p, hsa-miR-516a-3p

Table 4.2: Placenta-specific microRNAs. (continued)

miRNA ID	miRBase Accession	Ensembl Gene ID	Mature miRNAs
hsa-mir-516a-2	MI0003181	ENSG00000207620	hsa-miR-516a-5p, hsa-miR-516a-3p
hsa-mir-517a	MI0003161	ENSG00000207734	hsa-miR-517-5p, hsa-miR-517a-3p
hsa-mir-517b	MI0003165	ENSG00000207837	hsa-miR-517-5p, hsa-miR-517b-3p
hsa-mir-517c	MI0003174	ENSG00000207838	hsa-miR-517-5p, hsa-miR-517c-3p
hsa-mir-518f	MI0003154	ENSG00000207706	hsa-miR-518f-5p, hsa-miR-518f-3p
hsa-mir-518b	MI0003156	ENSG00000207862	hsa-miR-518b
hsa-mir-518c	MI0003159	ENSG00000207553	hsa-miR-518c-5p, hsa-miR-518c-3p
hsa-mir-518e	MI0003169	ENSG00000207987	hsa-miR-518e-5p, hsa-miR-518e-3p
hsa-mir-518a-1	MI0003170	ENSG00000207803	hsa-miR-518a-5p, hsa-miR-518a-3p
hsa-mir-518d	MI0003171	ENSG00000207747	hsa-miR-518d-5p, hsa-miR-518d-3p
hsa-mir-518a-2	MI0003173	ENSG00000207699	hsa-miR-518a-5p, hsa-miR-518a-3p
hsa-mir-519e	MI0003145	ENSG00000207810	hsa-miR-519e-5p, hsa-miR-519e-3p
hsa-mir-519c	MI0003148	ENSG00000207788	hsa-miR-519c-5p, hsa-miR-519c-3p
hsa-mir-519b	MI0003151	ENSG00000207825	hsa-miR-519b-5p, hsa-miR-519b-3p
hsa-mir-519d	MI0003162	ENSG00000207981	hsa-miR-519d
hsa-mir-519a-1	MI0003178	ENSG00000207992	hsa-miR-519a-5p, hsa-miR-519a-3p
hsa-mir-519a-2	MI0003182	ENSG00000207723	hsa-miR-519a-3p
hsa-mir-520e	MI0003143	ENSG00000207599	hsa-miR-520e
hsa-mir-520f	MI0003146	ENSG00000207555	hsa-miR-520f
hsa-mir-520a	MI0003149	ENSG00000207594	hsa-miR-520a-5p, hsa-miR-520a-3p
hsa-mir-520b	MI0003155	ENSG00000207722	hsa-miR-520b
hsa-mir-520c	MI0003158	ENSG00000207738	hsa-miR-520c-5p, hsa-miR-520c-3p
hsa-mir-520d	MI0003164	ENSG00000207735	hsa-miR-520d-5p, hsa-miR-520d-3p
hsa-mir-520g	MI0003166	ENSG00000207799	hsa-miR-520g
hsa-mir-520h	MI0003175	ENSG00000207861	hsa-miR-520h
hsa-mir-521-1	MI0003176	ENSG00000207634	hsa-miR-521

Table 4.2: Placenta-specific microRNAs. (continued)

miRNA ID	miRBase Accession	Ensembl Gene ID	Mature miRNAs
hsa-mir-521-2	MI0003163	ENSG00000207549	hsa-miR-521
hsa-mir-522	MI0003177	ENSG00000207806	hsa-miR-522-5p, hsa-miR-522-3p
hsa-mir-523	MI0003153	ENSG00000208016	hsa-miR-523-5p, hsa-miR-523-3p
hsa-mir-524	MI0003160	ENSG00000207977	hsa-miR-524-5p, hsa-miR-524-3p
hsa-mir-525	MI0003152	ENSG00000207711	hsa-miR-525-5p, hsa-miR-525-3p
hsa-mir-526b	MI0003150	ENSG00000207580	hsa-miR-526b-5p, hsa-miR-526b-3p
hsa-mir-526a-1	MI0003157	ENSG00000207629	hsa-miR-526a
hsa-mir-526a-2	MI0003168	ENSG00000211532	hsa-miR-526a
hsa-mir-527	MI0003179	ENSG00000207979	hsa-miR-527
hsa-mir-532	MI0003205	ENSG00000207758	hsa-miR-532-3p
hsa-mir-584	MI0003591	ENSG00000207714	hsa-miR-584-5p, hsa-miR-584-3p
hsa-mir-1283-1	MI0003832	ENSG00000221421	hsa-miR-1283
hsa-mir-1283-2	MI0006430	ENSG00000221548	hsa-miR-1283
hsa-mir-1323	MI0003786	ENSG00000221017	hsa-miR-1323
mmu-mir-136	MI0000162	ENSMUSG00000070129	mmu-miR-136-5p
mmu-mir-144	MI0000168	ENSMUSG00000065401	mmu-miR-144-3p
mmu-mir-467a-1	MI0002402	ENSMUSG00000076162	mmu-miR-467a-5p
mmu-mir-467d	MI0005513	ENSMUSG00000077021	mmu-miR-467d-5p
mmu-mir-467e	MI0006128	ENSMUSG00000076948	mmu-miR-467e-5p
mmu-mir-467a-2	MI0014053	ENSMUSG00000076161	mmu-miR-467a-5p
mmu-mir-467a-3	MI0014056	ENSMUSG00000076160	mmu-miR-467a-5p
mmu-mir-467a-4	MI0014059	ENSMUSG00000076159	mmu-miR-467a-5p
mmu-mir-467a-5	MI0014062	ENSMUSG00000076158	mmu-miR-467a-5p
mmu-mir-467a-6	MI0014065	ENSMUSG00000076157	mmu-miR-467a-5p
mmu-mir-467a-7	MI0014069		mmu-miR-467a-5p
mmu-mir-467a-8	MI0014072	ENSMUSG00000076155	mmu-miR-467a-5p
mmu-mir-467a-9	MI0014074		mmu-miR-467a-5p
mmu-mir-467a-10	MI0014076	ENSMUSG00000077033	mmu-miR-467a-5p

Table 4.2 Legend:

Shown on pages 259 to 263 inclusive are tables of information for all 141 placenta-specific miRNAs studied in this thesis. For each miRNA ID, the corresponding miRBase accession, Ensembl Gene ID and mature miRNAs are listed.

4.2.2 Placental Gene Set Assembly

The set of placental genes used here is the same as that used in Chapter 2. Because miRNA-target prediction analysis did not impose the same constraints, in terms of number of species, as the selective pressure analysis conducted in Chapter 2, the following genes — excluded in that chapter — were reintroduced here: pregnancy specific beta-1-glycoprotein 7 (PSG7), X antigen family members 2 and 3 (XAGE2 and XAGE3), and insulin-like 4 (INSL4). For each gene family member, the same representative transcript used in Chapter 2 was used again here.

4.2.3 Obtaining Genome Sequence Data

The complete genome assemblies of the 22 species of interest were manually downloaded from the Ensembl server (Flicek *et al.* 2012). All genome assemblies comprised nucleotide FASTA files: in 17 species these were arranged into chromosomes, while in the remaining 5 genome assemblies were non-chromosomal (see Table 2.1). For the human genome assembly GRCh37.p5, the genome assembly files included 114 genomic patch files, representing 40 fix patches, 65 novel patches and 9 haplotype sequence patches.

4.2.4 Obtaining Placental MicroRNA Sequence Data

The precursor sequence of each placental miRNA was obtained using the Perl script GetMirbaseSequences.pl on a local copy of miRBase. The resulting FASTA sequence file contained the precursor sequence of 141 miRNAs — 127 in human and 14 in mouse.

4.2.5 Obtaining Placental Gene Family 3′ UTR Sequence Data

Target gene 3′ UTR sequences were obtained as follows. The set of representative Transcript Ensembl IDs for each placental gene and its homologs (as obtained in Chapter 2) were used as a filter for a series of Ensembl BioMart 'Sequences' queries — one per species — with the attributes 'Ensembl Gene ID' and '3′ UTR'. The output from each such query was a file containing a FASTA entry for each representative transcript. If the representative transcript had an annotated 3′ UTR, the FASTA entry contained its 3′ UTR sequence. Otherwise, it contained one of the non-sequence strings 'Sequence unavailable' or 'No UTR is annotated for this transcript'. These non-sequence entries were not removed, as they were to be used later on to distinguish between cases where a species contained no homologs from those where a species contained one or more homologs but none had an annotated 3′ UTR. The resulting 3′ UTR sequences and non-sequence entries for each species were then concatenated into a single FASTA file.

This complete sequence file was then filtered to create a sequence file that only contained annotated 3′ UTR sequences, using the script FilterSequencesFASTA.pl. Non-sequences were filtered out, as were all genes for which the annotated 3′ UTR was shorter than 37 nucleotides. The minimum 3′ UTR length was set to 37, as this is the sum of the average length of a mature miRNA (i.e. 22 bases) and the distance downstream of the open reading frame (i.e. 15 bases), in which miRNA-target interactions have been found to have reduced effectiveness (Grimson *et al.* 2007). As used in this thesis, a **non-sequence** is placeholder text that is used by databases such as Ensembl to indicate that no sequence is present (e.g. 'Sequence unavailable').

The result was two FASTA sequence files: one containing a FASTA entry — either sequence or non-sequence — for all 2495 representative transcripts, the other containing FASTA sequence entries for the 1050 representative transcripts with an annotated 3′ UTR sequence of sufficient length to perform miRNA-target prediction. Figure 4.2 shows the number of placental genes in each species and the proportion of these genes with an annotated 3′ UTR sequence. Note the complete absence of annotated 3′ UTR sequences for these genes in elephant and lizard. Because of this, elephant and lizard were effectively excluded from the gene family miRNA-target prediction analysis in this chapter.

Figure 4.2 Legend:

The figure on page 267 overleaf depicts the heterogeneous levels of 3′ UTR sequence annotation for the 2495 gene family members of 114 placental genes across 22 species. Each column indicates the total number of placental genes of interest in the given genome; the filled section indicates the number of genes for which there is an annotated 3′ UTR nucleotide sequence in Ensembl that is at least 37 nucleotides in length (the minimum 3′ UTR length required for miRNA-target prediction) (Flicek *et al.* 2012), while the section outlined by a dashed line indicates the number of genes lacking an annotated 3′ UTR in Ensembl. Columns marked by an asterisk (*) denote that the species in question had almost no usable 3′ UTR sequences for the genes of interest, while columns marked by a double asterisk (**) denote that those species completely lacked annotated 3′ UTR sequence for the genes of interest.

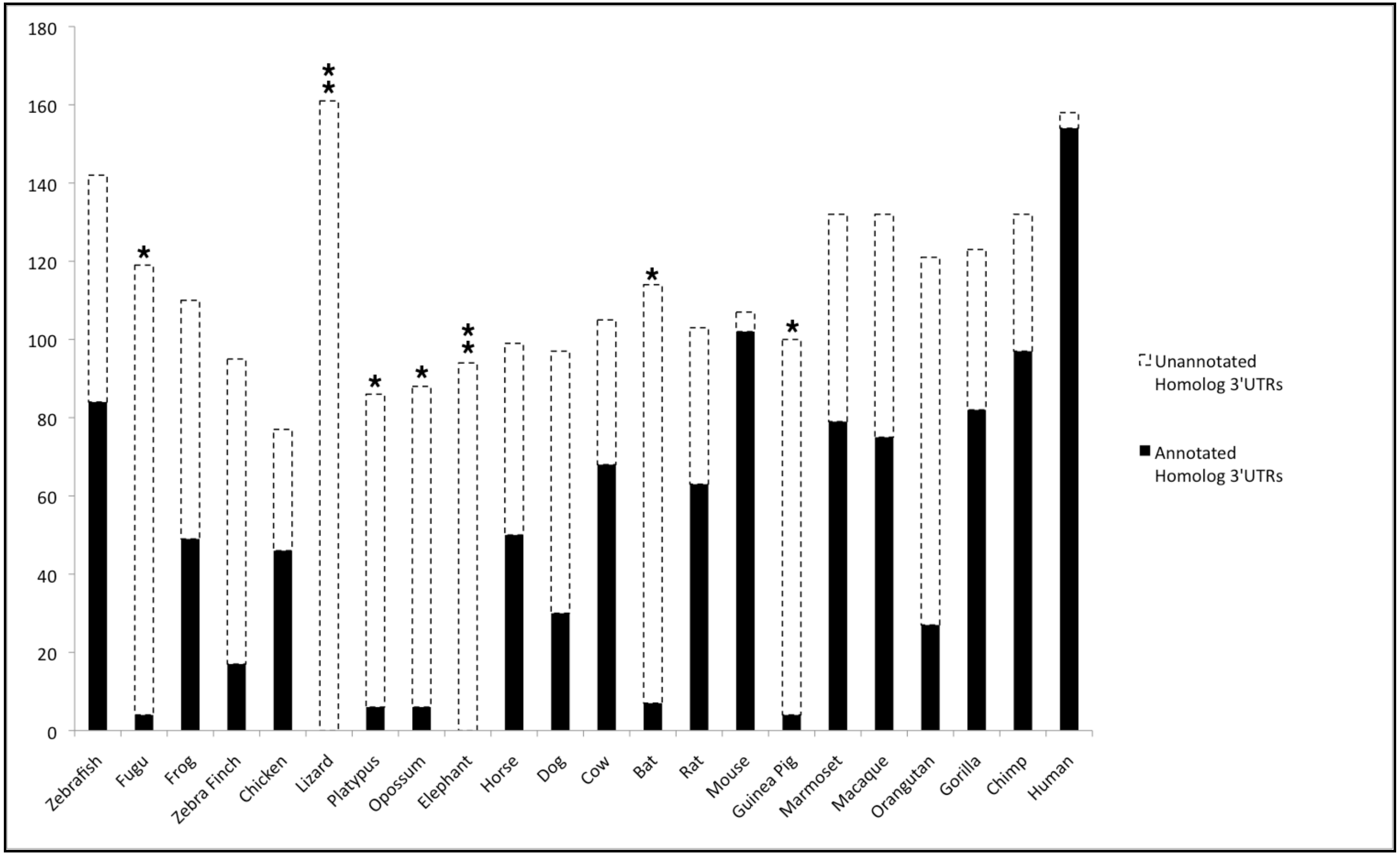


Figure 4.2: Levels of 3′ UTR annotation for 114 placental genes and their homologs in 22 species.

4.3 Methods

In order to identify whether a given miRNA-target interaction is conserved across species, it is first necessary to identify the respective homologs of the cognate miRNA and target. The gene family miRNA-target prediction therefore involves two steps: (i) identifying miRNA and target gene homologs in each species of interest, and (ii) predicting miRNA-target interactions between each miRNA and its target gene across all species in which they are present. The placental gene homologs were previously obtained from the Ensembl database (Flicek *et al.* 2012) in Chapter 2, so it only remained to identify the homologs of the placental miRNAs.

4.3.1 Identifying MicroRNA Homologs

This section describes the procedures followed to identify the homologs of 141 placental miRNAs in 22 species. It is important for a miRNA search to be as comprehensive as possible (Shomron *et al.* 2009). To ensure that miRNA gene families were as complete as possible, homology information was (I) extracted from miRBase, (II) downloaded from Ensembl, (III) obtained through a miRNA homolog search in all 22 species of interest, and (IV) integrated from these three sources into one dataset. The process undertaken is outlined below and summarised in Figure 4.3. All scripts used during this process can be accessed from the following location in the electronic appendix: Appendix / home / code / scripts.html.

I. miRBase Gene Family Data Download

With the set of miRNAs in human and mouse, homologs were obtained for the 18 species represented in miRBase using CreateMirbaseGeneFamilyTable.pl with the 141 miRNA precursor miRBase accessions. (See Table 4.3.) This script queries a local copy of miRBase using a set of user-defined filters and creates a gene family table for the miRNA precursors and/or miRNA gene families that pass those filters, in which each row contains the members of a particular miRNA gene family across different species, and each column contains the miRNAs of a particular species across different miRNA gene families.

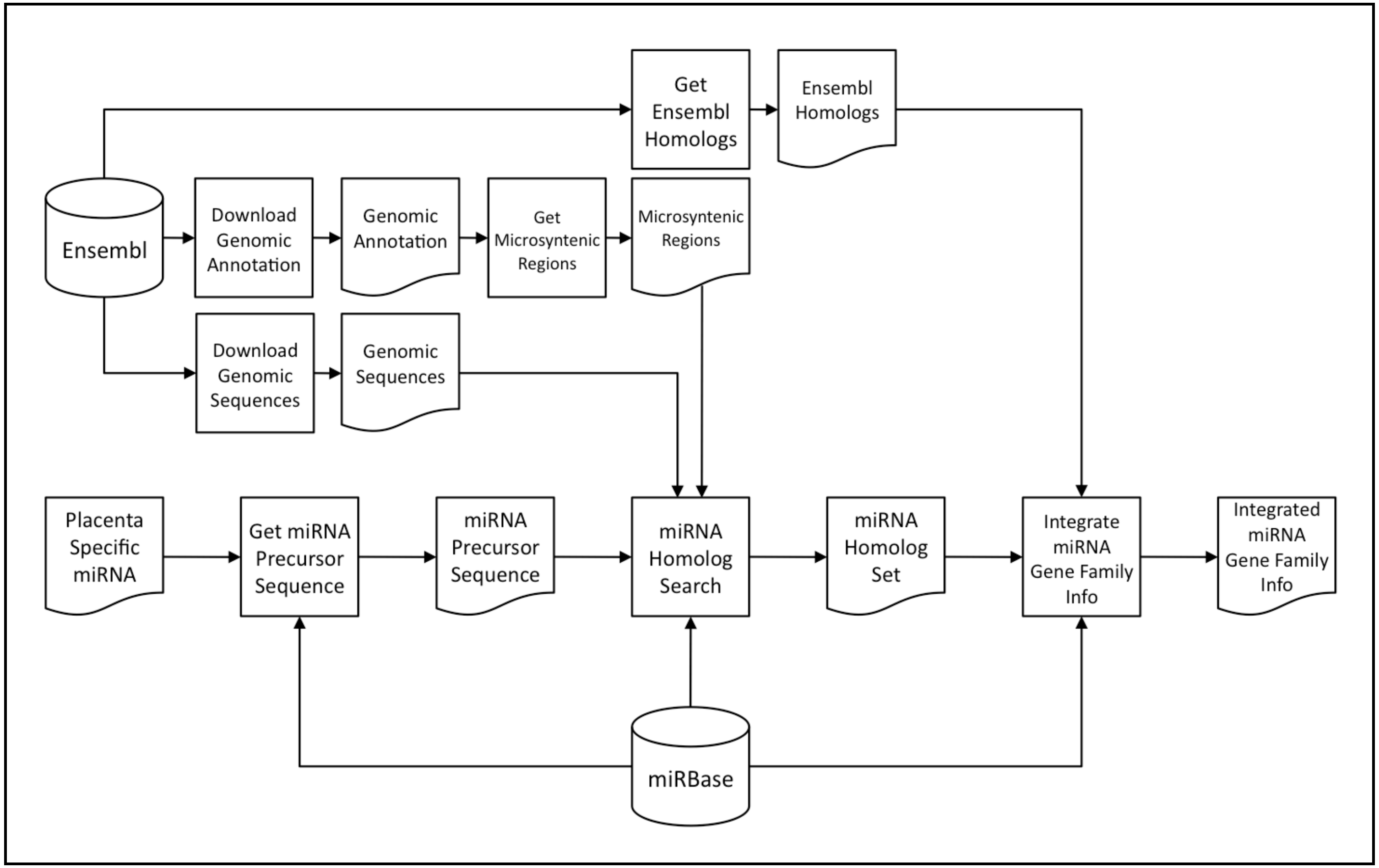


Figure 4.3: Placental microRNA homolog identification pipeline.

Shown is a bioinformatics pipeline depicting the steps involved in placental miRNA homolog identification. Arrows indicate the direction of process flow. For information on symbols used, see the pipeline key in Figure 2.2.

Table 4.3: miRBase data for the species studied in this thesis.

Species	Scientific Name	miRBase Entries	miRBase Genome Assembly	Ensembl Genome Assembly
Bat	*Myotis lucifugus*	0	N/A	myoLuc2
Chicken	*Gallus gallus*	499	WASHUC2	WASHUC2
Chimp	*Pan troglodytes*	600	PanTro2.1	CHIMP2.1.4
Cow	*Bos taurus*	662	BTAU4.0	UMD3.1
Dog	*Canis familiaris*	323	CanFam2.0	CanFam2.0
Elephant	*Loxodonta africana*	0	N/A	loxAfr3
Frog	*Xenopus tropicalis*	188	JGI4.1	JGI_4.2
Fugu	*Takifugu rubripes*	264	N/A	FUGU4.0
Gorilla	*Gorilla gorilla*	85	N/A	gorGor3.1
Guinea Pig	*Cavia porcellus*	0	N/A	cavPor3
Horse	*Equus caballus*	341	N/A	EquCab2
Human	*Homo sapiens*	1527	GRCh37	GRCh37.p5
Lizard	*Anolis carolinensis*	282	N/A	AnoCar2.0
Macaque	*Macaca mulatta*	479	MMUL1.0	MMUL_1.0
Marmoset	*Callithrix jacchus*	0	N/A	C_jacchus3.2.1
Mouse	*Mus musculus*	741	NCBIM37	NCBIM37
Opossum	*Monodelphis domestica*	156	MONDOM5	monDom5
Orangutan	*Pongo abelii*	581	N/A	PPYG2
Platypus	*Ornithorhynchus anatinus*	337	N/A	OANA5
Rat	*Rattus norvegicus*	408	RGSC3.4	RGSC3.4
Zebra Finch	*Taeniopygia guttata*	224	N/A	taeGut3.2.4
Zebrafish	*Danio rerio*	344	Zv9	Zv9

II. Ensembl Gene Family Data Download

With the set of miRNAs in human and mouse, homologs were obtained for all 22 species using a set of Ensembl BioMart 'Homologs' queries to obtain a paralog table and ortholog table as in Chapter 2. The Perl script named CreateEnsemblGeneFamilyTable.pl was then used to create a gene family table from the orthology and paralogy data.

III. miRNAminer(Emulator) Homolog Search

miRNAminer (Artzi *et al.* 2008) is software that takes, as input, a precursor and mature miRNA sequence and searches for homologs in a target genome sequence database. After an initial BLAST search (Camacho *et al.* 2009, Altschul *et al.* 1997, Altschul *et al.* 1990) performed with the query miRNA precursor sequence against the target genome, the miRNAminer algorithm filters BLAST hits by E-value per chromosome (set to 0.05 by default); remaining matches are then extended by genomic context of the BLAST hit by examining 50 bases upstream and downstream of the start and end, respectively, of the BLAST hit.

All possible extensions of each match within user-specified threshold lengths — by default, these are set to a minimum of 70 and maximum of 180 bases — are then filtered according to the following criteria: (i) RNA minimum free energy (MFE), (ii) minimum precursor intramolecular base-pairing, (iii) requirement for hairpin loop, (iv) minimum precursor alignment identity, (v) minimum mature sequence alignment identity, (vi) maximum mature sequence mismatches, (vii) seed conservation, and (viii) maximum overlap of mature sequence and hairpin loop. Putative homologs that satisfy these filters are returned by miRNAminer.

In addition to BLAST, miRNAminer makes use of JAligner (Moustafa 2005) — an open-source Java implementation of the Smith-Waterman algorithm (Smith and Waterman 1981) — for alignment of the query miRNA to each target, as well as RNAfold from the ViennaRNA package (Gruber *et al.* 2008, Hofacker 2003, Hofacker *et al.* 1994) to estimate RNA folding energy (Artzi *et al.* 2008).

Some aspects of the standard implementation of the miRNAminer algorithm rendered it less applicable to this miRNA homolog search. It only accepts input of one mature miRNA per query, but many of the placenta-specific miRNAs had two annotated mature miRNAs. The standard implementation uses the Smith-Waterman algorithm for local alignment of the query miRNA with its putative homolog sequences, but it was felt that this would not distinguish well between candidate homolog sequences with the same local alignment score to a query but with different overall lengths. Additionally, miRNAminer filters E-values by chromosome. Because of the proportional relationship of E-value to database size (see Equation 4.1), the difference in chromosome lengths would result in differences of BLAST search sensitivity across a target genome. For example, one would expect to see more BLAST hits by chance in chromosome 1 than chromosome 19, simply due to its much greater length (i.e. 249Mb in the former and 59Mb in the latter, in Ensembl release 65).

$$E = l_q \cdot l_D \cdot 2^{-S'}$$

Equation 4.1: Calculation of BLAST E-value.

The E-value of a BLAST hit (E) is shown in terms of the length of the query sequence (l_q), the length of the database sequence (l_D), and the normalised score of the BLAST hit (S'). See Altschul *et al.* (1997).

For these reasons, a modified form of the miRNAminer algorithm was implemented by this author in Perl for use with the placenta-specific miRNA dataset. The resulting Perl script, miRNAminerEmulator.pl, emulates the functionality of miRNAminer but incorporates the following changes to the algorithm:

- Two mature miRNAs can be included in a query miRNA precursor.

- E-values are estimated for the entire target genome.

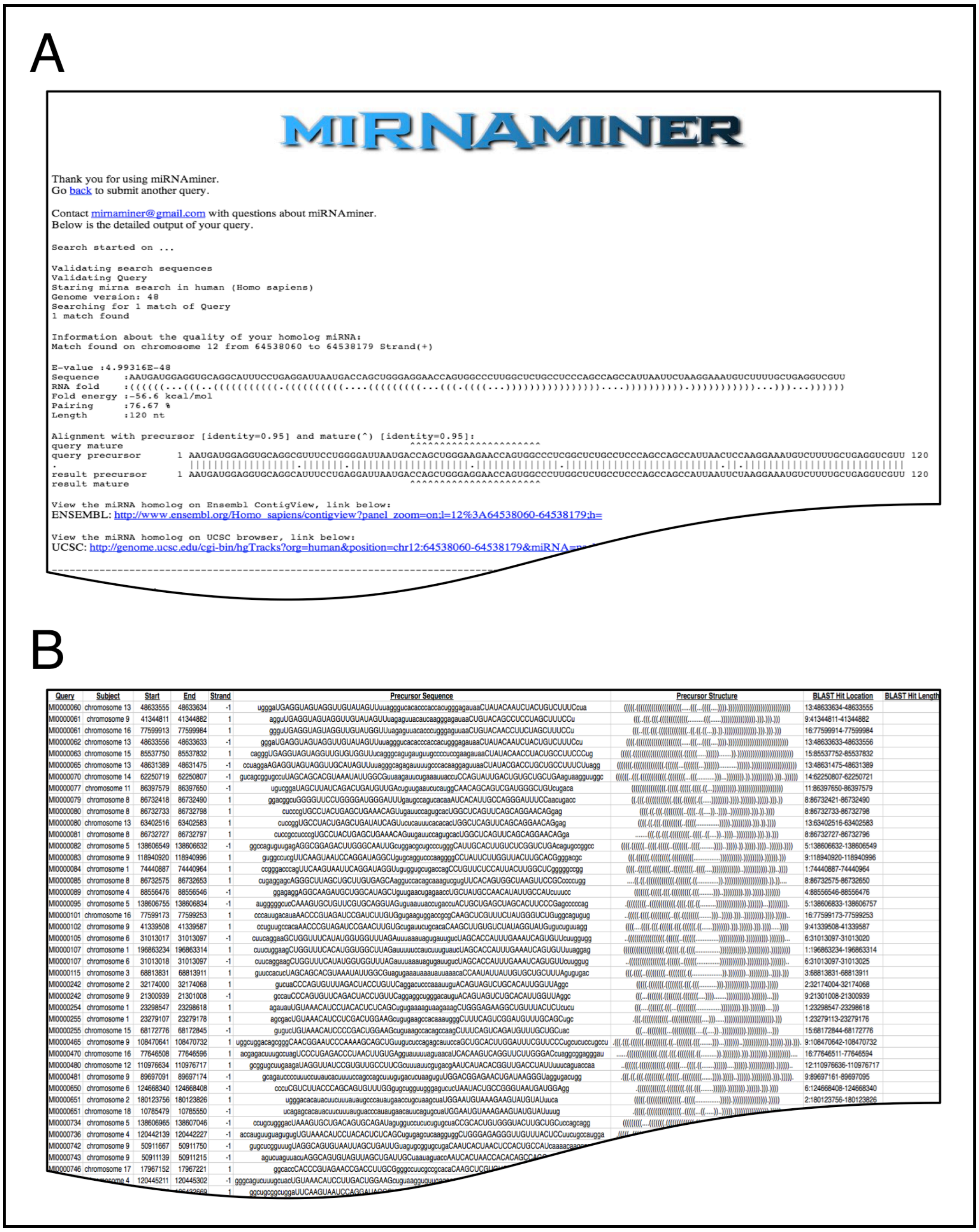
- Each query is aligned with putative homologs using the Needleman-Wunsch global alignment algorithm (Needleman and Wunsch 1970), as implemented in EMBOSS Needle (Rice *et al.* 2000).

- A **flank length** can be specified to set the length of the BLAST hit context sequence, such that context sequences are only evaluated if they are of a length within one flank length of the query sequence.

The introduction of a flank length parameter ensures that candidate homologs are only evaluated if they are of similar length to the query, which markedly reduces the number of context sequences that must be evaluated. While greater lengths of genomic context sequence are possible in principle, those most similar in length to the query are favoured in any case by both alignment and structural filters, so that in practice it is not worthwhile to assess context sequences of very different length to the query pre-miRNA.

Sample output for both the standard implementation and the emulator described here are shown in Figure 4.4. The Perl script miRNAminerEmulator.pl uses BLAST+ (Camacho *et al.* 2009) and ViennaRNA RNAfold (Gruber *et al.* 2008, Hofacker 2003, Hofacker *et al.* 1994) in the same way as the standard miRNAminer. However, instead of using JAligner for alignment of each query miRNA to its putative homolog match, miRNAminerEmulator.pl uses EMBOSS Needle (Rice *et al.* 2000) — an implementation of the global alignment Needleman-Wunsch algorithm (Needleman and Wunsch 1970). Needleman-Wunsch was used in miRNAminerEmulator.pl because this alignment algorithm favours sequences with high similarity across the entire alignment over those with high similarity only in local regions — a desirable property in this case, since well-defined homologous regions were sought.

Each miRNA of interest was input to miRNAminerEmulator.pl, which searched each genome in turn for homologs of that miRNA, which were then output to a table showing information about sequence, position and other attributes for each homolog. The settings used were varied depending on the evolutionary distance between query and target species. Some example settings are shown in Table 4.4.



Thank you for using miRNAminer.
Go back to submit another query.

Contact mirnaminer@gmail.com with questions about miRNAminer.
Below is the detailed output of your query.

```
Search started on ...

Validating search sequences
Validating Query
Staring mirna search in human (Homo sapiens)
Genome version: 48
Searching for 1 match of Query
1 match found

Information about the quality of your homolog miRNA:
Match found on chromosome 12 from 64538060 to 64538179 Strand(+)

E-value :4.99316E-48
Sequence    :AAUGAUGGAGGUGCAGGCAUUUCCUGAGGAUUAAUGACCAGCUGGGAGGAACCAGUGGCCCUUGGCUCUGCCUCCCAGCCAGCCAUUAAUUCUAAGGAAAUGUCUUUUGCUGAGGUCGUU
RNA fold    :((((((...(((..((((((((((((.(((((((((((....((((((((((...(((.((((...))))))))))))))))))))....))))))))))).))))))))))...)))...))))))
Fold energy :-56.6 kcal/mol
Pairing     :76.67 %
Length      :120 nt

Alignment with precursor [identity=0.95] and mature(^) [identity=0.95]:
query mature                                                ^^^^^^^^^^^^^^^^^^^^^^
query precursor      1 AAUGAUGGAGGUGCAGGCGUUUCCUGGGGAUUAAUGACCAGCUGGGAAGAACCAGUGGCCCUCGGCUCUGCCUCCCAGCCAGCCAUUAACUCCAAGGAAAUGUCUUUUGCUGAGGUCGUU 120
.                      ||||||||||||||||||.|||||||.|||||||||||||||||||.||||||||||||||.||||||||||||||||||||||||||.||.||||||||||||||||||||||||||||
result precursor     1 AAUGAUGGAGGUGCAGGCAUUUCCUGAGGAUUAAUGACCAGCUGGGAGGAACCAGUGGCCCUUGGCUCUGCCUCCCAGCCAGCCAUUAAUUCUAAGGAAAUGUCUUUUGCUGAGGUCGUU 120
result mature                                               ^^^^^^^^^^^^^^^^^^^^^^

View the miRNA homolog on Ensembl ContigView, link below:
```

ENSEMBL: http://www.ensembl.org/Homo_sapiens/contigview?panel_zoom=on;l=12%3A64538060-64538179;h=

```
View the miRNA homolog on UCSC browser, link below:
```

UCSC: http://genome.ucsc.edu/cgi-bin/hgTracks?org=human&position=chr12:64538060-64538179&miRNA=

Query	Subject	Start	End	Strand	Precursor Sequence	Precursor Structure	BLAST Hit Location	BLAST Hit Length
MI0000060	chromosome 13	48633555	48633634	-1	ugggaUGAGGUAGUAGGUUGUAUAGUUuuaggucacacccaccacugggagauaaCUAUACAAUCUACUGUCUUUCcua	(((((.((((((((((((((((((((.....(((...((((....)))).))))))))))))))))))))))))))	13:48633634-48633555	
MI0000061	chromosome 9	41344811	41344882	1	agguUGAGGUAGUAGGUUGUAUAGUUuagaguuacaucaagggagauaaCUGUACAGCCUCCUAGCUUUCCu	(((..(((.(((.((((((((((((.........(((......))))))))))))))).))).))).)))	9:41344811-41344882	
MI0000061	chromosome 16	77599913	77599984	1	ggguUGAGGUAGUAGGUUGUAUGGUUuagaguuacacccugggaguuaaCUGUACAACCUUCUAGCUUUCCu	(((..(((.(((.((((((((((((..((.((.((...)).)).)).))))))))))))).))).))).)))	16:77599914-77599984	
MI0000062	chromosome 13	48633556	48633633	-1	gggaUGAGGUAGUAGGUUGUAUAGUUuuaggucacacccaccacugggagauaaCUAUACAAUCUACUGUCUUUCcu	((((.((((((((((((((((((((.....(((...((((....)))).)))))))))))))))))))))))))	13:48633633-48633556	
MI0000063	chromosome 15	85537750	85537832	1	cagggUGAGGUAGUAGGUUGUGUGGUUucagggcagugauguugccccuccgaagauaaCUAUACAACCUACUGCCUUCCCug	(((((.((((((((((((((((((((((.(((((.....)))))........)).))))))))))))))))))))))))	15:85537752-85537832	
MI0000065	chromosome 13	48631389	48631475	-1	ccuaggaAGAGGUAGUAGGUUGCAUAGUUuuagggcagagauuuugcccacaaggaguuaaCUAUACGACCUGCUGCCUUUCUuagg	((((((.((((((((((((.(((((...((((((...)))))).........))))).))))))))))))))))))	13:48631475-48631389	
MI0000070	chromosome 14	62250719	62250807	-1	gucagcggugccuUAGCAGCACGUAAAUAUUGGCGuuaagauucugaaauuaccuCCAGUAUUGACUGUGCUGCUGAaguaagguuggc	(((((..(((.((((((((((.(((((((...(((..........)))...))))))).)).)))))))))).)))..)))))	14:62250807-62250721	
MI0000077	chromosome 11	86397579	86397650	-1	ugucggaUAGCUUAUCAGACUGAUGUUGAcuguugaaucucauggCAACAGCAGUCGAUGGGCUGUcugaca	(((((((((((((((.(((((.(((((.((((.((...)))))))))))).))))))))))))))))))	11:86397650-86397579	
MI0000079	chromosome 8	86732418	86732490	1	ggacggcuGGGGUUCCUGGGGAUGGGAUUUgaugccagucacaaAUCACAUUGCCAGGGAUUUCCaacugacc	((.(((.((((((((((((.((((.((((((.((.....))))))))).)))).))))))).))))).))).))	8:86732421-86732490	
MI0000080	chromosome 8	86732733	86732798	1	cucccgUGCCUACUGAGCUGAAACAGUugauuccagugcacUGGCUCAGUUCAGCAGGAACAGgag	((((.((.(((.((((((((..((((..((....))..))))..)))))))).))).)).))))	8:86732733-86732798	
MI0000080	chromosome 13	63402516	63402583	1	cuccggUGCCUACUGAGCUGAUAUCAGUucucauuucacacacUGGCUCAGUUCAGCAGGAACAGgag	((((.((.(((.((((((((..(((((.............))))).)))))))).))).)).))))	13:63402516-63402583	
MI0000081	chromosome 8	86732727	86732797	1	cuccgccucccgUGCCUACUGAGCUGAAACAGuugauuccagugcacUGGCUCAGUUCAGCAGGAACAGga	.......(((.((.(((.((((((((..((((..((....))..))))..)))))))).))).)).)))	8:86732727-86732796	
MI0000082	chromosome 5	138606549	138606632	-1	ggccaguguugagAGGCGGAGACUUGGGCAAUUGcuggacgcugcccugggCAUUGCACUUGUCUCGGUCUGAcagugccggcc	((((.((((((..((((.(((((..(((((((..((((.........))))..))))).)).))))).)))))..)))))).))))	5:138606632-138606549	
MI0000083	chromosome 9	118940920	118940996	1	guggccucgUUCAAGUAAUCCAGGAUAGGCUgugcaggucccaaggggCCUAUUCUUGGUUACUUGCACGgggacgc	(((.((((((.((((((((.(((((((((...............))))))))).)))))))).)))))).)))	9:118940920-118940996	
MI0000084	chromosome 1	74440887	74440964	1	ccgggacccagUUCAAGUAAUUCAGGAUAGGUuguggugcugaccagCCUGUUCUCCAUUACUUGGCUCgggggccgg	((((..((((((.((((((((..(((((((((..((((....))))))))))))..))))))))))).))))..))))	1:74440887-74440964	
MI0000085	chromosome 8	86732575	86732653	1	cugaggagcAGGGCUUAGCUGCUUGUGAGCAagguccacagcaaagucgugUUCACAGUGGCUAAGUUCCGCccccugg	....((.((.((((((((((((.(((((((.((.............)).))))))))))))))))))).)).))....	8:86732575-86732650	
MI0000089	chromosome 4	88556476	88556546	-1	ggagaggAGGCAAGAUGCUGGCAUAGCUguugaacugagaaccUGCUAUGCCAACAUAUUGCCAUcuuucc	(((((((.(((((.(((.((((((((.(((.........)))..)))))))).))).))))).)))))))	4:88556546-88556476	
MI0000095	chromosome 5	138606755	138606834	-1	augggggcucCAAAGUGCUGUUCGUGCAGGUAGuguaauuaccugaccuACUGCUGAGCUAGCACUUCCCGagcccccag	.((((((((..((((((((((((.((((.(((.((..........)))))))))))))).))))))...))))))))).	5:138606833-138606757	
MI0000101	chromosome 16	77599173	77599253	1	cccauugacauaAACCCGUAGAUCCGAUCUUGUGgugaaguggaccgcgCAAGCUCGUUUCUAUGGGUCUGuggcagugug	..(((((.((((.((((((((..(((.(((((((........)))..))))).)))..))))))))).)))).))))))..	16:77599173-77599253	
MI0000102	chromosome 9	41339508	41339587	1	ccuguugccacaAACCCGUAGAUCCGAACUUGUGcugauucugcacaCAAGCUUGUGUCUAUAGGUAUGugucuguuagg	((((.....((((.(((.(((((.(((.((((((.((.......)))))))).))).)))))).))).))))......))))	9:41339508-41339587	
MI0000105	chromosome 6	31013017	31013097	-1	cuucaggaaGCUGGUUUCAUAUGGUGGUUUAGAuuuaaauagugauugucUAGCACCAUUUGAAAUCAGUGUUcuuggugg	..(((((((((((((((((.((((((..((((((..............)))))))))))).)))))))))).)))))))...	6:31013097-31013020	
MI0000107	chromosome 1	196863234	196863314	1	cuucuggaagCUGGUUUCACAUGGUGGCUUAGauuuuuccaucuuuguaucUAGCACCAUUUGAAAUCAGUGUUuuaggag	((((((((((((((((((.((((((.((.((((...............)))))))))))).))))))))).)))))))))	1:196863234-196863314	
MI0000107	chromosome 6	31013018	31013097	-1	cuucaggaagCUGGUUUCAUAUGGUGGUUUAGauuuaaauagugauugucUAGCACCAUUUGAAAUCAGUGUUcuuggug	..((((((((((((((((.((((((..((((((..............)))))))))))).))))))))).))))))))..	6:31013097-31013025	
MI0000115	chromosome 3	68813831	68813911	1	guuccacucUAGCAGCACGUAAAUAUUGGCGuagugaaauaaauauuaaacaCCAAUAUUAUUGUGCUGCUUUAgugugac	(((.((((..((((((((..((((((((.((....................)).))))))))..))))))))..)))).)))	3:68813831-68813911	
MI0000242	chromosome 2	32174000	32174068	1	gucuaCCCAGUGUUUAGACUACCUGUUCaggacucccaaauuguACAGUAGUCUGCACAUUGGUUAggc	(((((.(((((((.((((((((.(((.(((...........))).)))))))))))).))))))).)))))	2:32174004-32174068	
MI0000242	chromosome 9	21300939	21301008	-1	gccauCCCAGUGUUCAGACUACCUGUUCaggaggcugggacauguACAGUAGUCUGCACAUUGGUUAggc	(((...((((((((.((((((((.((((((....))))........)))))))))).))))))))...)))	9:21301008-21300939	
MI0000254	chromosome 1	23298547	23298618	1	agauauUGUAAACAUCCUACACUCUCAGCugugaaaaguaagaaagCUGGGAGAAGGCUGUUUACUCUcucu	(((....((((((.(((...(((((((((..............))))))))).))).))))))....)))	1:23298547-23298618	
MI0000255	chromosome 1	23279107	23279178	1	agcgacUGUAAACAUCCUCGACUGGAAGcugugaagccacaaaugggCUUUCAGUCGGAUGUUUGCAGCugc	.(((.((((((((((..((((((((((((((....))).....)))))))))))))))))))))).)))	1:23279113-23279176	
MI0000255	chromosome 15	68172776	68172845	-1	gugucUGUAAACAUCCCCGACUGGAAGcuguaagccacagccaagCUUUCAGUCAGAUGUUUGCUGCuac	(((...((((((((...((((((((((.....((.....))..)))))))))).)))))))...)))	15:68172844-68172776	
MI0000465	chromosome 9	108470641	108470732	1	uggcuggacagcgggCAACGGAAUCCCAAAAGCAGCUGuugucuccagagcauuccaGCUGCACUUGGAUUUCGUUCCCugcucuccugccu	.(((.(((.((((((.(((((((((.((..(((((((.(((........)))...)))))))..))))))))))).)))))).))).))).	9:108470642-108470732	
MI0000470	chromosome 16	77646508	77646596	1	acgagacuuugccuagUCCCUGAGACCCUAACUUGUGAgguauuuuaguaacaUCACAAGUCAGGUUCUUGGGACcuaggcggagggau	......((((((((((((.((((.(((.((((((((.((..........)).)))))))).))).)))))))).)))))))))))....	16:77646511-77646594	
MI0000480	chromosome 12	110976634	110976717	1	gcggugcuugaagaUAGGUUAUCCGUGUUGCCUUCGcuuuauucgugacgAAUCAUACACGGUUGACCUAUUuuucaguaccaa	..((((((.((((((((((((.(((((((....((((.((.........)))))).....))))).)))))))))))))))).))))))..	12:110976636-110976717	
MI0000481	chromosome 9	89697091	89697174	-1	gcagaucccccuuuccuuaucacuuuuccagccagcuuugugacucuaaguguUGGACGGAGAACUGAUAAGGGUaggugacugg	.(((.((.(((.((((((((.((((((..((((((((.......)))).)))))..)))))).)))))))).))).)).))))).	9:89697161-89697095	
MI0000650	chromosome 6	124668340	124668408	-1	cccuCGUCUUACCCAGCAGUGUUUGGgugcugguugggagucucUAAUACUGCCGGGUAAUGAUGGAgg	.((((((((((((.((((((((.(((.(((.........))))).)))))))).)))))).))).))))	6:124668408-124668340	
MI0000651	chromosome 2	180123756	180123826	1	ugggacacauacuucuuuauaugcccauaugaaccugcuaagcuaUGGAAUGUAAAGAAGUAUGUAUuuca	(((((.((((((((((((((..(((((..............))))).)))))))))))))))).)))))	2:180123756-180123826	
MI0000651	chromosome 18	10785479	10785550	-1	ucagagcacauacuucuuuauguacccauaugaacauucagugcuaUGGAAUGUAAAGAAGUAUGUAUuuug	.(((((.((((((((((((((..(((((...((.....))..))))).))))))))))))))		
MI0000734	chromosome 5	138606965	138607046	-1	ccugcugggacUAAAGUGCUGACAGUGCAGAUagugguccucucugugcuaCCGCACUGUGGGUACUUGCUGCuccagcagg	((((((((....((((((.((((((((		
MI0000736	chromosome 4	120442139	120442227	-1	accauguuguagugugUGUAAACAUCCUACACUCUCAGCugugagcucaagguggCUGGGAGAGGGUUGUUUACUCCuucugccaugga	.(((((..(((		
MI0000742	chromosome 9	50911667	50911750	-1	gugcucgguuugUAGGCAGUGUAAUUAGCUGAUUGuagugcggugcugaCAAUCACUAACUCCACUGCCAUcaaaacaag			
MI0000743	chromosome 9	50911139	50911215	-1	agucuaguuacuAGGCAGUGUAGUUAGCUGAUUGCuaauaguaccAAUCACUAACCACACAGCCAG			
MI0000746	chromosome 17	17967152	17967221	1	ggcaccCACCCGUAGAACCGACCUUGCGgggccuucgccgcacaCAAGCUCGUG			
	mosome 4	120445211	120445302	-1	gggcagucuuugcuacUGUAAACAUCCUUGACUGGAAGcuguaagguguuga			
			[illegible]	1	ggcugcggcuggaUUCAAGUAAUCCAGGAUAGG			

Figure 4.4: miRNAminer/Emulator output.

Shown are samples of output from (A) miRNAminer (Artzi *et al.* 2008) in HTML format, and (B) miRNAminerEmulator.pl in TSV format. In the miRNAminerEmulator output, the **Query** column gives the precursor miRNA query identifier; the **Subject** column gives the name of the subject sequence in the target genome (e.g. chromosome 9); the **Start**, **End** and **Strand** columns give the putative homolog's chromosomal start position, chromosomal end position, and chromosomal strand, respectively; the **Precursor Sequence** and **Precursor Structure** columns give the sequence and predicted secondary structure, respectively, of the putative homolog; and subsequent columns give the values of the various miRNAminer parameters for the putative homolog.

Table 4.4: miRNAminerEmulator settings used.

Genome	flank	maxDG	minBP	maxPG	minPI	maxOL	minMI	minL	maxMM	maxE	minSI
Bat	20	-23.60	0.53	9	0.750	7	0.92	55	1	0.002	6
Chicken	20	-23.59	0.53	21	0.635	7	0.79	27	5	0.071	5
Chimp	20	-23.59	0.53	26	0.620	7	0.80	23	5	0.052	5
Cow	20	-23.59	0.53	9	0.750	7	0.92	55	1	0.002	6
Dog	20	-23.59	0.53	9	0.750	7	0.92	55	1	0.002	6
Elephant	20	-23.60	0.53	9	0.750	7	0.92	55	1	0.002	6
Frog	20	-23.59	0.53	33	0.595	7	0.80	23	5	0.104	5
Fugu	20	-23.59	0.53	40	0.570	7	0.76	22	7	0.170	6
Gorilla	20	-23.59	0.53	26	0.620	7	0.80	23	5	0.052	5
Horse	20	-23.59	0.53	9	0.750	7	0.92	55	1	0.002	6
Lizard	20	-23.59	0.53	21	0.635	7	0.79	27	5	0.071	5
Macaque	20	-23.59	0.53	26	0.620	7	0.80	23	5	0.052	5
Marmoset	20	-23.59	0.53	26	0.620	7	0.80	23	5	0.052	5
Mouse	20	-23.59	0.53	11	0.770	7	0.88	52	3	0.120	6
Opossum	20	-23.59	0.53	23	0.600	7	0.79	23	5	0.242	7
Orangutan	20	-23.59	0.53	26	0.620	7	0.80	23	5	0.052	5
Platypus	20	-23.59	0.53	35	0.560	7	0.69	22	8	0.071	2
Rat	20	-23.59	0.53	11	0.770	7	0.88	52	3	0.120	6
Zebrafish	20	-23.59	0.53	40	0.570	7	0.76	22	7	0.170	6

Table 4.4 Legend:

Shown are example miRNAminerEmulator settings used in this chapter. Parameters include the **flank**, to set genome context sequence length, maximum thermodynamic ΔG of precursor (**maxDG**), minimum proportion of intramolecular base pairs (**minBP**), maximum number of gaps in precursor alignment (**maxPG**), minimum proportion of identical sites in precursor alignment (**minPI**), maximum overlap between mature region and hairpin loop (**maxOL**), minimum proportion of identical sites in mature sequence alignment (**minMI**), minimum BLAST hit length (**minL**), maximum number of mismatches in mature sequence alignment (**maxMM**), maximum BLAST E-value (**maxE**), and minimum number of identical sites in the miRNA seed alignment (**minSI**).

In parallel with the whole-genome miRNA homolog search, homologous microsyntenic regions were searched. In order to improve the sensitivity of the miRNA homolog search while maintaining high specificity, genes lying immediately upstream and downstream of each miRNA of interest were used to establish the boundaries of that miRNA's microsyntenic region. Homologous microsyntenic regions were obtained for each miRNA in each species using the Perl script GetSyntenicRegions.pl. The narrower search space presented by these microsyntenic regions — typically less than 1 million bases, compared to more than 3 billion nucleotides in, for example, the human genome — allowed for a small marginal gain in sensitivity that improved recovery of homologous miRNAs in these regions (e.g. 2% more miRNA homologs in macaque). miRNAminerEmulator.pl was run as before (see Table 4.4) but with an additional argument specifying the microsyntenic region(s) to be searched.

In the human genome assembly, regions of the reference sequence that coincide with a genome fix patch were also searched for putative homologs. Putative homologs found on a fix patch were reported from the reference sequence only if the corresponding sequence on the fix patch was unchanged. If the fix patch appeared to have affected the DNA sequence of a homolog (i.e. if the number of putative homologs in the fix patch differed from the reference sequence, or if the set of sequences in the fix patch did not match those in the reference sequence), the match was taken from the fix patch and the corresponding match in the patched region of the reference sequence was ignored.

IV. Integrating Homology Data

Having obtained homology data from miRBase, Ensembl and from a miRNAminerEmulator.pl search, this homology information was integrated using the Perl script GetIntegratedMirnaGeneFamilyInfo.pl, which took the homology information from these three sources and combined them into one integrated miRNA gene family table and associated miRNA sequence file.

Where a homolog was present in miRBase, the miRBase accession was given in the output gene family table. Where a homolog was absent from miRBase but present in Ensembl, the Ensembl Gene ID was used. Further, if a homolog was identified by miRNAminerEmulator.pl and present in Ensembl, but absent from miRBase, mature regions of the homolog as inferred by miRNAminerEmulator.pl were mapped onto the Ensembl miRNA precursor sequence. In cases where a miRNA homolog was absent from both miRBase and Ensembl, a provisional miRNA ID was used, based on the miRNA ID of the relevant query miRNA (e.g. with query hsa-miR-1-1, a homolog in elephant might be laf-miR-1-P1). This allowed the provenance of each homolog to be determined from its identifier in the integrated miRNA gene family table.

4.3.2 Identifying MicroRNA Targets

This section describes the procedures followed to identify the target genes of 141 placental miRNAs and their homologs in 22 species. MiRNA-target prediction for the gene families of each placenta-specific miRNA and placental gene required a number of steps: (I) obtaining estimates of target abundance for each species of interest, (II) preparing TargetScan input files, (III) performing miRNA-target prediction with TargetScan, and (IV) processing TargetScan output files. The processes involved are outlined below and summarised in Figure 4.5. All scripts used during this process can be accessed from the following location in the electronic appendix: Appendix / home / code / scripts.html.

I. Estimating Target Abundance

As described in Section 3.3.1, TargetScan TA/SPS files had previously been obtained for human, chicken, cow, mouse, rat and zebrafish. Representative 3′ UTRs had also been obtained during the miTP benchmark for a further 9 species. For 7 of these species, target abundance (TA) estimates were obtained using the Perl script GetTargetAbundance.pl, taking that species' representative 3′ UTRs as input. However, fugu and guinea pig had 727 and 232 representative 3′ UTRs, respectively. Given such a small pool of representative 3′ UTRs, it was unlikely that accurate TA estimates could be obtained for these species. Because of this, fugu and guinea pig TA estimates were taken from those of zebrafish and mouse, respectively.

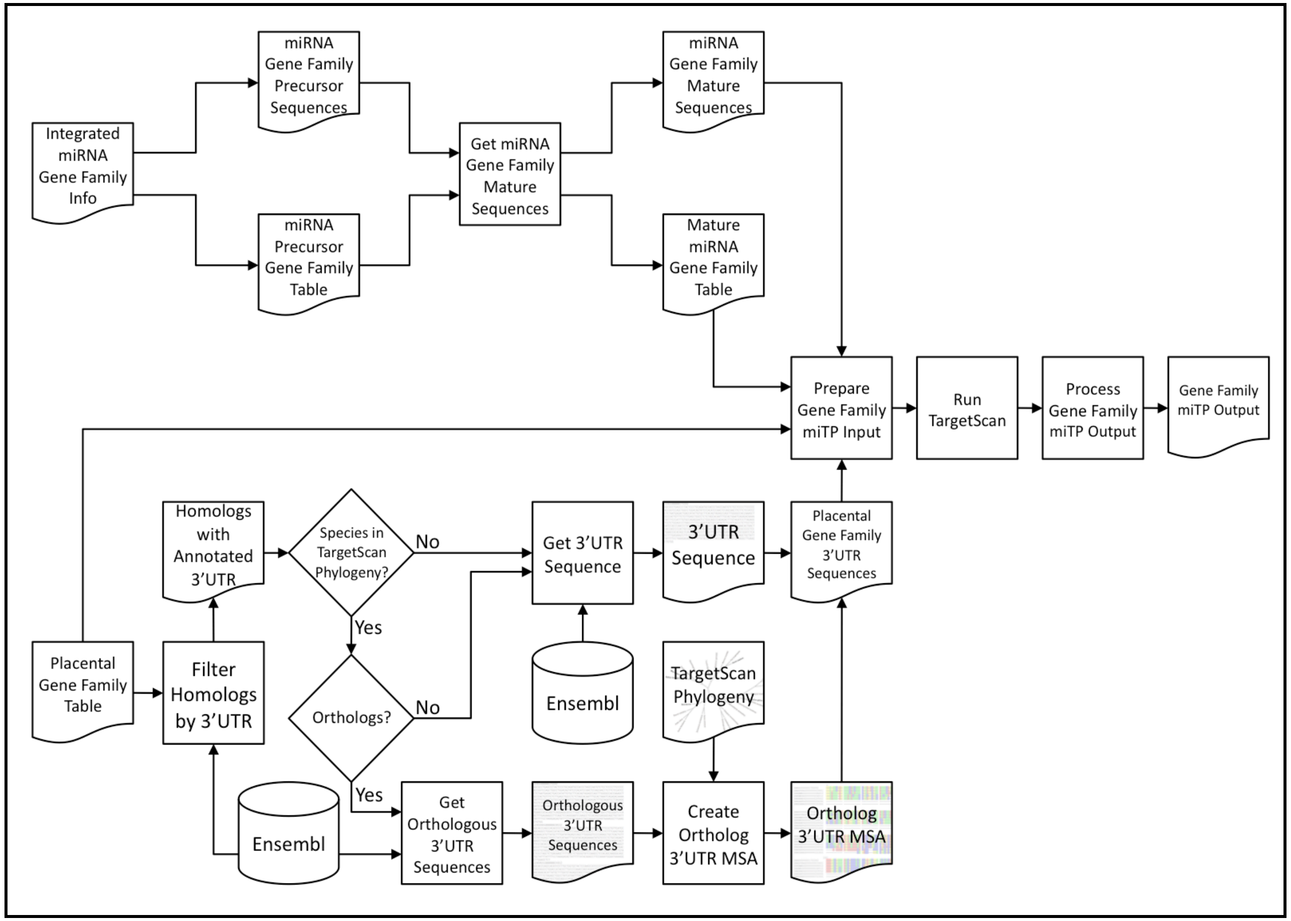

Integrated miRNA Gene Family Info
miRNA Gene Family Precursor Sequences
miRNA Precursor Gene Family Table
Get miRNA Gene Family Mature Sequences
miRNA Gene Family Mature Sequences
Mature miRNA Gene Family Table
Prepare Gene Family miTP Input
Run TargetScan
Process Gene Family miTP Output
Gene Family miTP Output
Homologs with Annotated 3'UTR
Species in TargetScan Phylogeny?
No
Yes
Get 3'UTR Sequence
3'UTR Sequence
Placental Gene Family 3'UTR Sequences
Placental Gene Family Table
Filter Homologs by 3'UTR
Orthologs?
No
Yes
Ensembl
TargetScan Phylogeny
Ensembl
Get Orthologous 3'UTR Sequences
Orthologous 3'UTR Sequences
Create Ortholog 3'UTR MSA
Ortholog 3'UTR MSA

Figure 4.5: Placental microRNA target prediction pipeline.

Shown is a bioinformatics pipeline depicting the steps involved in placental miRNA target prediction. Arrows indicate the direction of process flow. For information on symbols used, see the pipeline key in Figure 2.2.

For the 5 species not in the TargetScan phylogeny, representative 3′ UTRs were obtained for the first time. These sets of representative 3′ UTRs were then used to generate TA estimates for each of the 5 non-TargetScan species. The 18 sets of TA estimates were then used to create 20 species-specific TA/SPS files for all species except elephant and lizard (see Figure 4.2).

II. Preparing TargetScan Input

With the integrated miRNA gene family table and associated precursor sequence file, the Perl script GetMatureMirnaGeneFamilyInfo.pl was used to obtain a mature miRNA gene family table and its corresponding mature miRNA sequence file. For every miRNA precursor in the original miRNA gene family table, the mature miRNA gene family table contained any mature miRNAs corresponding to that miRNA precursor. For precursor miRNAs derived from miRBase, the corresponding mature miRNA accession and sequence were obtained directly from a local copy of miRBase. For Ensembl Gene IDs and miRNAminer hits, a provisional mature miRNA ID was obtained by appending '-5p' or '-3p' to the mature miRNA, depending on its position within the precursor sequence. For Ensembl miRNAs that lack a demarcated mature region, no corresponding mature miRNA could be output.

Five of the species under study — bat, gorilla, orangutan, marmoset and zebra finch — are absent from the phylogeny included with TargetScan, so it was not possible to run the TargetScan conservation script on any miRNA-target pairs in these species. For these species, only the target gene 3′ UTR sequence was required, as an alignment of orthologs is not required by the TargetScan context script. These were obtained by filtering the set of genes in each such species to retain only those with a corresponding annotated 3′ UTR sequence.

For all other species, ortholog alignments were obtained for all target genes with one or more annotated orthologous 3′ UTR sequences. This was done using the ortholog table created in Chapter 2. This ortholog table showed human genes and their orthologs in all other species, but for each other species, an ortholog table would be required from the perspective of that species. The human oriented ortholog table was reoriented to the perspective of each respective species using the Perl script ReorientEnsemblOrthologTable.pl.

Using each species-specific ortholog table along with the sequence file containing annotated 3′ UTRs, a representative ortholog table was obtained using the Perl script GetRepresentativeOrthologs.pl. This script removed those orthologs without an annotated sequence, and chose a representative ortholog where multiple orthologs were present, by taking the ortholog that aligned best to the gene of interest. For alignment, this script uses EMBOSS Needle (Rice *et al.* 2000), an implementation of the global alignment Needleman-Wunsch algorithm (Needleman and Wunsch 1970). Selection of a representative ortholog was necessary because TargetScan only accepts one target gene ortholog per species.

An ortholog sequence file was obtained for each gene of interest, containing the sequences of the gene itself and each of its representative orthologs. There were 16 genes without an ortholog sequence (e.g. pregnancy-specific beta-1-glycoprotein 9, or PSG9). For those genes with one or more orthologs (e.g. lectin, galactoside-binding, soluble, 14, or LGALS14), a guide tree was created for each gene of interest using CreateGuideTree.pl with the species tree discussed in Chapter 2, based on that of Benton and Donoghue (2007). The ortholog sequences and guide tree were input to PRANK v.100802, which output an MSA of orthologs for each gene of interest. In the TargetScan phylogeny species, 838 genes were readied for miTP with TargetScan: 822 ortholog alignments, 16 gene sequences. No annotated 3′ UTRs were available for lizard or elephant.

The Perl script PrepGeneFamilyMITP.pl was used to prepare TargetScan input files across the 20 species with both miRNA and target gene homolog sequences, including miRNA seed and sequence files and target gene alignment files in the TSV format expected by TargetScan, arranged in an ordered directory structure by species. Lizard and elephant were excluded due to lack of target gene data, while the five non-TargetScan species were excluded from TargetScan miRNA-target prediction using the conservation script, and the gene family of tachykinin 3 (TAC3) was excluded because this gene does not itself have an available 3′ UTR sequence.

III. MicroRNA-Target Prediction with TargetScan

TargetScan was run without error on the ICHEC Stokes HPC cluster using one ICHEC taskfarm for each of the four TargetScan scripts: the standard TargetScan script, the branch length bins script, the TargetScan context script and the TargetScan conservation script. The standard TargetScan script was run in parallel with the branch length bins script, followed by a parallel run of the TargetScan context and conservation scripts. See Figure 4.6 for a visual summary of this process. The complete set of taskfarm runs used approximately 250 core hours to test ~220,000 miRNA-target interactions with the four TargetScan scripts — taking less than 4 seconds to test each miTI, on average.

IV. Processing TargetScan Output

The Perl script ProcGeneFamilyMITP.pl was used to process the gene family miTP output. For both the human and mouse miRNA gene families, mature and precursor miRNA gene family tables and associated miRNA info files were included in post-processing. Both the annotated and original target gene tables were also included. This made it easier to distinguish cases where there was no miRNA or gene homolog from those in which there are miRNA and gene homologs but not necessarily available sequences. Although they were not included in the TargetScan analysis, the species elephant and lizard were included in post-processing for completeness.

The final processed output comprised 277 TSV files showing gene family miTP results: one for each of 141 key miRNAs, 114 key genes and 22 species, as well as 2 TSV files showing gene family miTP results specifically for key miRNAs and genes in human and mouse, respectively.

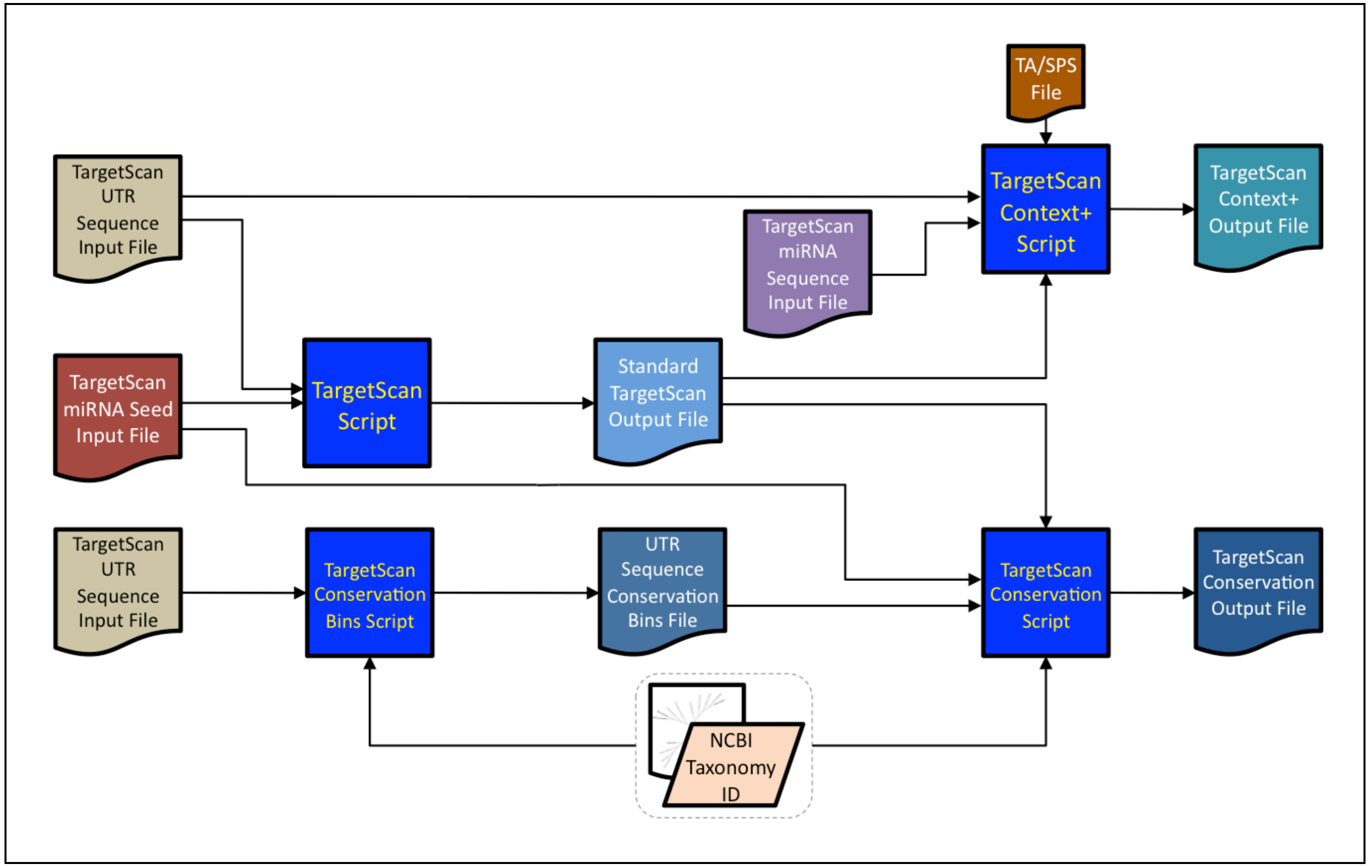

TA/SPS File
TargetScan UTR Sequence Input File
TargetScan Context+ Script
TargetScan Context+ Output File
TargetScan miRNA Sequence Input File
TargetScan miRNA Seed Input File
TargetScan Script
Standard TargetScan Output File
TargetScan UTR Sequence Input File
TargetScan Conservation Bins Script
UTR Sequence Conservation Bins File
TargetScan Conservation Script
TargetScan Conservation Output File
NCBI Taxonomy ID

Figure 4.6: TargetScan script pipeline.

Shown is a bioinformatics pipeline depicting the steps involved in running the TargetScan Perl scripts. Arrows indicate the direction of process flow. The dashed grey rounded rectangle encompasses the NCBI Taxonomy ID of the reference species and the TargetScan phylogeny, both of which are used as input by the TargetScan branch length bins and conservation scripts. For information on other symbols used, see the pipeline key in Figure 2.2.

4.4 Results

Sections 4.4.1 and 4.4.2 outline the results of the miRNA homolog search and gene family miRNA-target prediction, respectively.

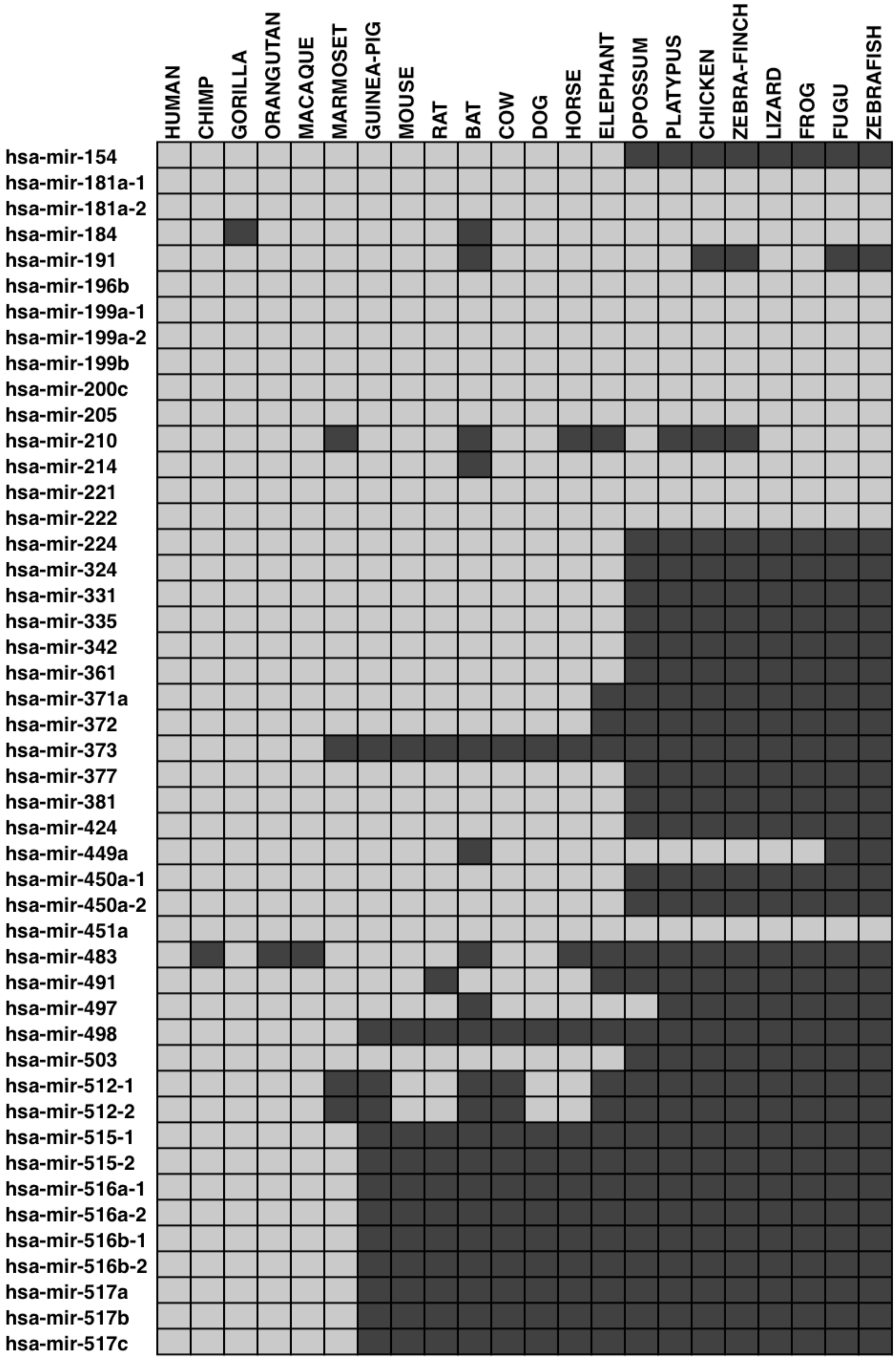
4.4.1 MicroRNA Homologs

MiRNA homologs were identified in the 22 species of interest. Figure 4.7 and Figure 4.8 show gene family presence/absence diagrams for placental miRNAs in human and mouse, respectively. In both figures, each row corresponds to a placental miRNA, each column corresponds to a species, and the colour of each cell indicates whether homologs of the given miRNA are present in the given species: light grey indicates that a homolog is present, while dark grey indicates that a homolog is absent. These figures integrate homology data from three sources: the miRBase miRNA database (Griffiths-Jones 2004, Griffiths-Jones 2006, Griffiths-Jones *et al.* 2008, Kozomara and Griffiths-Jones 2011), the Ensembl genomic database (Flicek *et al.* 2012) and the results of a homolog search against the Ensembl genomic sequences by the script miRNAminerEmulator.pl, which emulates the functionality of miRNAminer (Artzi *et al.* 2008). Note that the integrated miRNA gene family dataset is accessible in the electronic appendix at the following file path: Appendix / home / projects / placenta / pipelines / microrna-homologs / integrated-microrna-gene-family-info.html.

In these presence/absence diagrams, the clade in which each miRNA can be seen, and the lineage of origin of each miRNA may therefore be inferred from the clades in which it is present. It can be seen, for example, that the members of the C19MC miRNA cluster (i.e. hsa-mir-512-1 to hsa-mir-527) are broadly primate-specific.

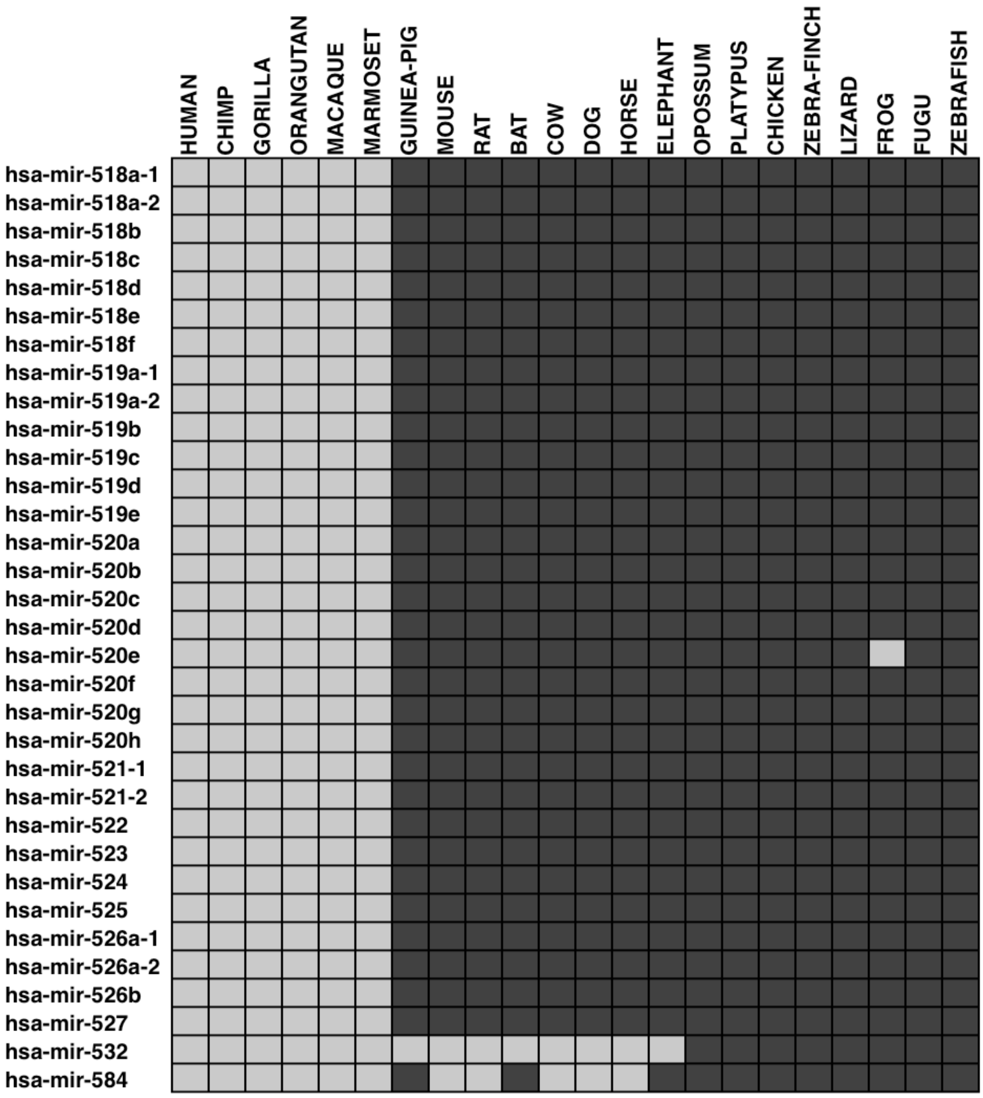
	HUMAN	CHIMP	GORILLA	ORANGUTAN	MACAQUE	MARMOSET	GUINEA-PIG	MOUSE	RAT	BAT	COW	DOG	HORSE	ELEPHANT	OPOSSUM	PLATYPUS	CHICKEN	ZEBRA-FINCH	LIZARD	FROG	FUGU	ZEBRAFISH
hsa-let-7a-1	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□
hsa-let-7a-2	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□
hsa-let-7a-3	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□
hsa-let-7b	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□
hsa-let-7d	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□
hsa-mir-1-1	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□
hsa-mir-1-2	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□
hsa-mir-15b	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□
hsa-mir-16-1	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□
hsa-mir-16-2	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□
hsa-mir-21	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	■	□	□
hsa-mir-23a	□	□	□	□	□	□	□	□	□	■	□	□	□	□	□	□	□	□	□	□	□	□
hsa-mir-24-1	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□
hsa-mir-24-2	□	□	□	□	□	□	□	□	□	■	□	□	□	□	□	□	□	□	□	□	□	□
hsa-mir-25	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□
hsa-mir-26a-1	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□
hsa-mir-26a-2	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□
hsa-mir-26b	□	□	□	□	□	□	□	□	□	■	□	□	□	□	□	□	□	□	□	□	□	□
hsa-mir-27a	□	□	□	□	□	□	□	□	□	■	□	□	□	□	□	□	□	□	□	□	□	□
hsa-mir-27b	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□
hsa-mir-29b-1	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□
hsa-mir-29b-2	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□
hsa-mir-30b	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□
hsa-mir-30c-1	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□
hsa-mir-30c-2	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□
hsa-mir-30d	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□
hsa-mir-30e	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□
hsa-mir-31	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□
hsa-mir-34a	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□
hsa-mir-34b	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□
hsa-mir-34c	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□
hsa-mir-92b	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□
hsa-mir-93	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□
hsa-mir-99a	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□
hsa-mir-99b	□	□	□	□	□	□	□	□	□	■	□	□	□	□	□	□	□	□	□	□	□	□
hsa-mir-100	□	□	□	□	□	□	■	□	□	■	□	□	□	□	□	□	□	□	□	□	□	□
hsa-mir-106b	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□
hsa-mir-122	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□
hsa-mir-125b-1	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□
hsa-mir-125b-2	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□
hsa-mir-1283-1	□	□	□	□	□	□	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
hsa-mir-1283-2	□	□	□	□	□	□	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
hsa-mir-1323	□	□	□	□	□	□	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
hsa-mir-140	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□
hsa-mir-141	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□
hsa-mir-143	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	■	■	□	□	□	□
hsa-mir-151a	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	■	■	■	■	■	■

Figure 4.7: Presence/absence of placental microRNAs in 22 species.



	HUMAN	CHIMP	GORILLA	ORANGUTAN	MACAQUE	MARMOSET	GUINEA-PIG	MOUSE	RAT	BAT	COW	DOG	HORSE	ELEPHANT	OPOSSUM	PLATYPUS	CHICKEN	ZEBRA-FINCH	LIZARD	FROG	FUGU	ZEBRAFISH
hsa-mir-154	□	□	□	□	□	□	□	□	□	□	□	□	□	□	■	■	■	■	■	■	■	■
hsa-mir-181a-1	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□
hsa-mir-181a-2	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□
hsa-mir-184	□	□	■	□	□	□	□	□	□	■	□	□	□	□	□	□	□	□	□	□	□	□
hsa-mir-191	□	□	□	□	□	□	□	□	□	■	□	□	□	□	□	□	■	■	□	□	■	■
hsa-mir-196b	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□
hsa-mir-199a-1	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□
hsa-mir-199a-2	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□
hsa-mir-199b	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□
hsa-mir-200c	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□
hsa-mir-205	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□
hsa-mir-210	□	□	□	□	□	■	□	□	□	■	□	□	■	■	□	■	■	■	□	□	□	□
hsa-mir-214	□	□	□	□	□	□	□	□	□	■	□	□	□	□	□	□	□	□	□	□	□	□
hsa-mir-221	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□
hsa-mir-222	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□
hsa-mir-224	□	□	□	□	□	□	□	□	□	□	□	□	□	□	■	■	■	■	■	■	■	■
hsa-mir-324	□	□	□	□	□	□	□	□	□	□	□	□	□	□	■	■	■	■	■	■	■	■
hsa-mir-331	□	□	□	□	□	□	□	□	□	□	□	□	□	□	■	■	■	■	■	■	■	■
hsa-mir-335	□	□	□	□	□	□	□	□	□	□	□	□	□	□	■	■	■	■	■	■	■	■
hsa-mir-342	□	□	□	□	□	□	□	□	□	□	□	□	□	□	■	■	■	■	■	■	■	■
hsa-mir-361	□	□	□	□	□	□	□	□	□	□	□	□	□	□	■	■	■	■	■	■	■	■
hsa-mir-371a	□	□	□	□	□	□	□	□	□	□	□	□	□	■	■	■	■	■	■	■	■	■
hsa-mir-372	□	□	□	□	□	□	□	□	□	□	□	□	□	■	■	■	■	■	■	■	■	■
hsa-mir-373	□	□	□	□	□	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
hsa-mir-377	□	□	□	□	□	□	□	□	□	□	□	□	□	□	■	■	■	■	■	■	■	■
hsa-mir-381	□	□	□	□	□	□	□	□	□	□	□	□	□	□	■	■	■	■	■	■	■	■
hsa-mir-424	□	□	□	□	□	□	□	□	□	□	□	□	□	□	■	■	■	■	■	■	■	■
hsa-mir-449a	□	□	□	□	□	□	□	□	□	■	□	□	□	□	□	□	□	□	□	□	■	■
hsa-mir-450a-1	□	□	□	□	□	□	□	□	□	□	□	□	□	□	■	■	■	■	■	■	■	■
hsa-mir-450a-2	□	□	□	□	□	□	□	□	□	□	□	□	□	□	■	■	■	■	■	■	■	■
hsa-mir-451a	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□
hsa-mir-483	□	■	□	■	■	□	□	□	□	■	□	□	■	■	■	■	■	■	■	■	■	■
hsa-mir-491	□	□	□	□	□	□	□	□	■	□	□	□	□	■	■	■	■	■	■	■	■	■
hsa-mir-497	□	□	□	□	□	□	□	□	□	■	□	□	□	□	□	■	■	■	■	■	■	■
hsa-mir-498	□	□	□	□	□	□	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
hsa-mir-503	□	□	□	□	□	□	□	□	□	□	□	□	□	□	■	■	■	■	■	■	■	■
hsa-mir-512-1	□	□	□	□	□	■	■	□	□	■	■	□	□	■	■	■	■	■	■	■	■	■
hsa-mir-512-2	□	□	□	□	□	■	■	□	□	■	■	□	□	■	■	■	■	■	■	■	■	■
hsa-mir-515-1	□	□	□	□	□	□	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
hsa-mir-515-2	□	□	□	□	□	□	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
hsa-mir-516a-1	□	□	□	□	□	□	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
hsa-mir-516a-2	□	□	□	□	□	□	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
hsa-mir-516b-1	□	□	□	□	□	□	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
hsa-mir-516b-2	□	□	□	□	□	□	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
hsa-mir-517a	□	□	□	□	□	□	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
hsa-mir-517b	□	□	□	□	□	□	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
hsa-mir-517c	□	□	□	□	□	□	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■

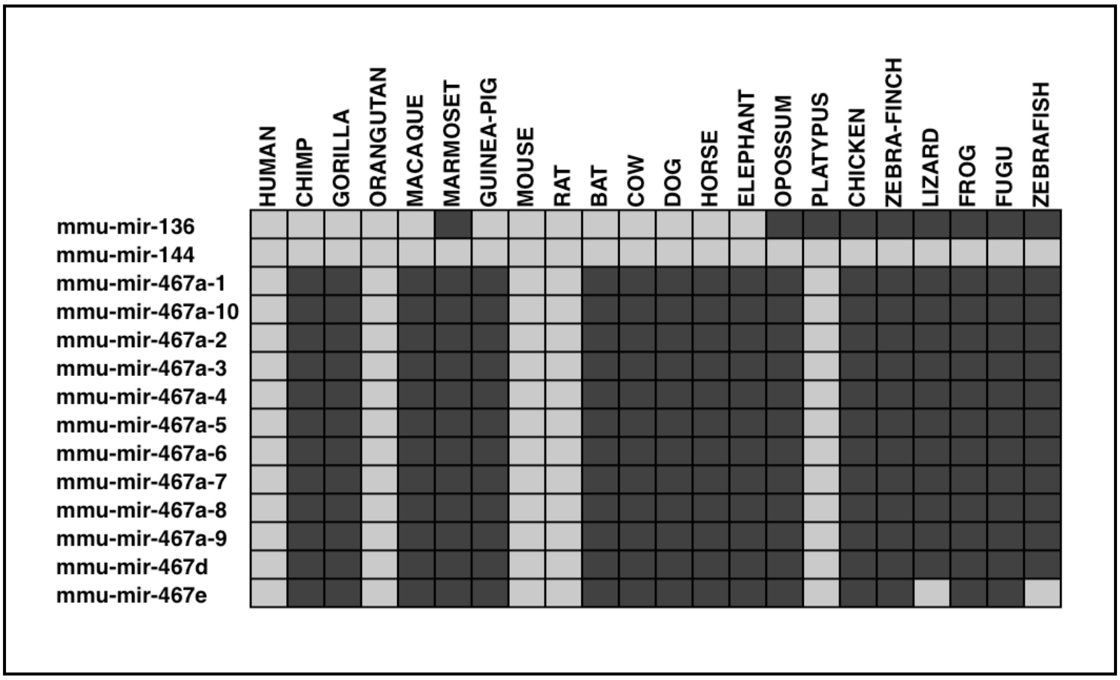
Figure 4.7: Presence/absence of placental microRNAs in 22 species. (continued)



	HUMAN	CHIMP	GORILLA	ORANGUTAN	MACAQUE	MARMOSET	GUINEA-PIG	MOUSE	RAT	BAT	COW	DOG	HORSE	ELEPHANT	OPOSSUM	PLATYPUS	CHICKEN	ZEBRA-FINCH	LIZARD	FROG	FUGU	ZEBRAFISH
hsa-mir-518a-1	□	□	□	□	□	□	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
hsa-mir-518a-2	□	□	□	□	□	□	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
hsa-mir-518b	□	□	□	□	□	□	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
hsa-mir-518c	□	□	□	□	□	□	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
hsa-mir-518d	□	□	□	□	□	□	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
hsa-mir-518e	□	□	□	□	□	□	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
hsa-mir-518f	□	□	□	□	□	□	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
hsa-mir-519a-1	□	□	□	□	□	□	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
hsa-mir-519a-2	□	□	□	□	□	□	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
hsa-mir-519b	□	□	□	□	□	□	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
hsa-mir-519c	□	□	□	□	□	□	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
hsa-mir-519d	□	□	□	□	□	□	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
hsa-mir-519e	□	□	□	□	□	□	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
hsa-mir-520a	□	□	□	□	□	□	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
hsa-mir-520b	□	□	□	□	□	□	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
hsa-mir-520c	□	□	□	□	□	□	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
hsa-mir-520d	□	□	□	□	□	□	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
hsa-mir-520e	□	□	□	□	□	□	■	■	■	■	■	■	■	■	■	■	■	■	■	□	■	■
hsa-mir-520f	□	□	□	□	□	□	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
hsa-mir-520g	□	□	□	□	□	□	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
hsa-mir-520h	□	□	□	□	□	□	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
hsa-mir-521-1	□	□	□	□	□	□	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
hsa-mir-521-2	□	□	□	□	□	□	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
hsa-mir-522	□	□	□	□	□	□	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
hsa-mir-523	□	□	□	□	□	□	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
hsa-mir-524	□	□	□	□	□	□	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
hsa-mir-525	□	□	□	□	□	□	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
hsa-mir-526a-1	□	□	□	□	□	□	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
hsa-mir-526a-2	□	□	□	□	□	□	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
hsa-mir-526b	□	□	□	□	□	□	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
hsa-mir-527	□	□	□	□	□	□	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
hsa-mir-532	□	□	□	□	□	□	□	□	□	□	□	□	□	□	■	■	■	■	■	■	■	■
hsa-mir-584	□	□	□	□	□	□	■	□	□	■	□	□	□	■	■	■	■	■	■	■	■	■

Figure 4.7: Presence/absence of placental microRNAs in 22 species. (continued)

Shown on pages 284 to 286 inclusive are patterns of presence and absence in 22 vertebrate species of homologs for 127 human placental miRNAs. Each row shows the homologs of a specific miRNA, while each column shows miRNA homologs in a particular species. Presence is denoted by a light grey cell, while absence is denoted by a dark grey cell. Homolog information was obtained from three sources: miRBase (Griffiths-Jones 2004, Griffiths-Jones 2006, Griffiths-Jones *et al.* 2008, Kozomara and Griffiths-Jones 2011), Ensembl (Flicek *et al.* 2012), and an implementation of the miRNAminer algorithm (Artzi *et al.* 2008).



	HUMAN	CHIMP	GORILLA	ORANGUTAN	MACAQUE	MARMOSET	GUINEA-PIG	MOUSE	RAT	BAT	COW	DOG	HORSE	ELEPHANT	OPOSSUM	PLATYPUS	CHICKEN	ZEBRA-FINCH	LIZARD	FROG	FUGU	ZEBRAFISH
mmu-mir-136	□	□	□	□	□	■	□	□	□	□	□	□	□	□	■	■	■	■	■	■	■	■
mmu-mir-144	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□
mmu-mir-467a-1	□	■	■	□	■	■	■	□	□	■	■	■	■	■	■	□	■	■	■	■	■	■
mmu-mir-467a-10	□	■	■	□	■	■	■	□	□	■	■	■	■	■	■	□	■	■	■	■	■	■
mmu-mir-467a-2	□	■	■	□	■	■	■	□	□	■	■	■	■	■	■	□	■	■	■	■	■	■
mmu-mir-467a-3	□	■	■	□	■	■	■	□	□	■	■	■	■	■	■	□	■	■	■	■	■	■
mmu-mir-467a-4	□	■	■	□	■	■	■	□	□	■	■	■	■	■	■	□	■	■	■	■	■	■
mmu-mir-467a-5	□	■	■	□	■	■	■	□	□	■	■	■	■	■	■	□	■	■	■	■	■	■
mmu-mir-467a-6	□	■	■	□	■	■	■	□	□	■	■	■	■	■	■	□	■	■	■	■	■	■
mmu-mir-467a-7	□	■	■	□	■	■	■	□	□	■	■	■	■	■	■	□	■	■	■	■	■	■
mmu-mir-467a-8	□	■	■	□	■	■	■	□	□	■	■	■	■	■	■	□	■	■	■	■	■	■
mmu-mir-467a-9	□	■	■	□	■	■	■	□	□	■	■	■	■	■	■	□	■	■	■	■	■	■
mmu-mir-467d	□	■	■	□	■	■	■	□	□	■	■	■	■	■	■	□	■	■	■	■	■	■
mmu-mir-467e	□	■	■	□	■	■	■	□	□	■	■	■	■	■	■	□	■	■	□	■	■	□

Figure 4.8: Presence/absence of mouse microRNAs in 22 species.

The figure above shows patterns of presence and absence in 22 vertebrate species of homologs for 14 human placental miRNAs. Each row shows the homologs of a specific miRNA, while each column shows miRNA homologs in a particular species. Presence is denoted by a light grey cell, while absence is denoted by a dark grey cell. Homolog information was obtained from three sources: miRBase (Griffiths-Jones 2004, Griffiths-Jones 2006, Griffiths-Jones *et al.* 2008, Kozomara and Griffiths-Jones 2011), Ensembl (Flicek *et al.* 2012), and an implementation of the miRNAminer algorithm (Artzi *et al.* 2008).

4.4.2 MicroRNA Targets

Gene family miRNA-target predictions were performed for 127 and 14 placental miRNAs in human and mouse, respectively, against 114 human placental genes. Of a possible 14,478 miRNA-target pairs in human, 2,856 were predicted by TargetScan to have a functional interaction. Given 1,596 possible miRNA-target pairs in mouse, 214 were predicted to have a functional interaction by TargetScan. Full gene family miRNA-target prediction results are accessible in the electronic appendix at the following file path: Appendix / home / projects / placenta / pipelines / microrna-targets / microrna-target-predictions.html.

The clades in which miRNA-target interactions are predicted to act are shown for human in Figure 4.9 and Figure 4.10, and mouse in Figure 4.11 and Figure 4.12. For example, the 'Primates' column of Figure 4.9 shows 834 predicted miRNA-target interactions that are conserved across primates; the vast majority of these (i.e. 663) are from miRNA-target pairs that began to co-exist within the genome of the ancestral primate, while only 171 primate-specific miRNA-target interactions are predicted between miRNAs and target genes that co-existed in the genome prior to the emergence of primates. MiRNAs and targets that have co-existed since the ancestral primate were also predicted to interact in the ape clade (153) and even specifically in human (73). This large proportion of primate-specific miRNA-target interactions is not necessarily unexpected, since the miRNAs in the C19MC cluster are primate-specific and it would be expected that the miRNAs in this cluster would have acquired targets among those genes responsible for placental development in the primate clade.

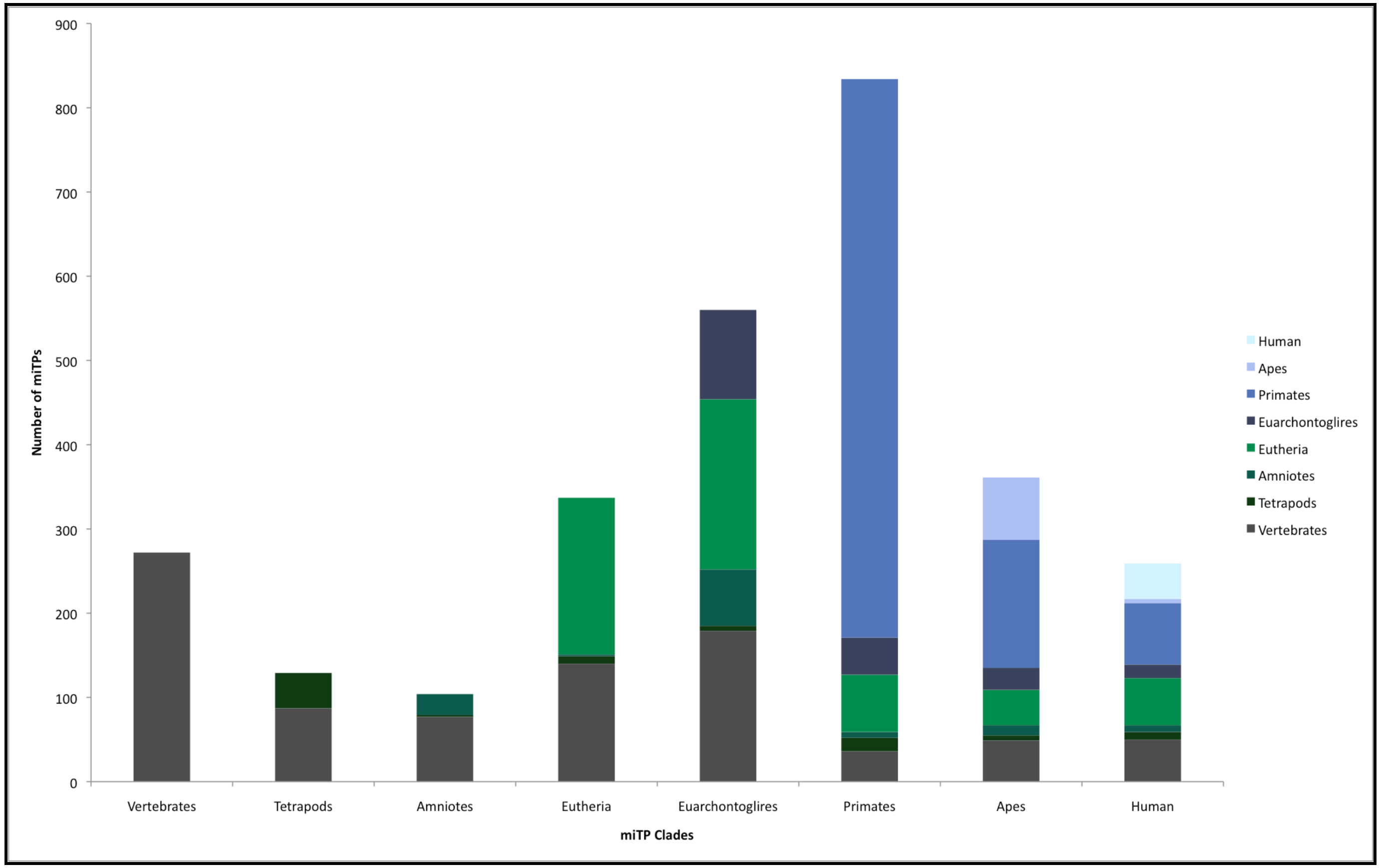


Figure 4.9: Origin of human placental microRNAs and their target sites.

Each stacked column above represents the number of miRNA-target predictions for the given clade, and each differently coloured column within the stacked column represents a clade of origin (i.e. the clade in which both the miRNA and target gene were present). The clade colour code used here is identical to that used in Figure 1.6.

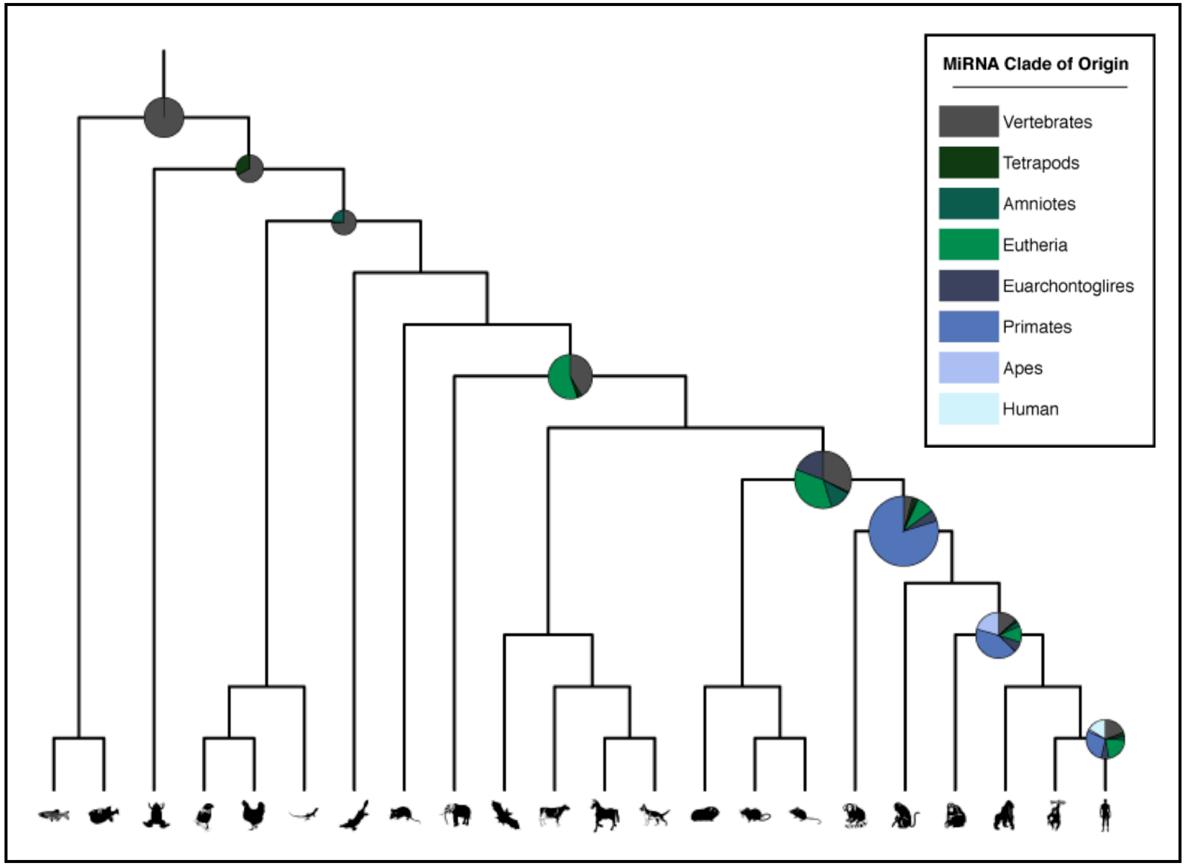


Figure 4.10: Phylogeny depicting origin of human microRNAs and their target sites.

Shown is a phylogeny of all 22 species, overlaid with pie charts, such that the area of each pie chart represents the number of miRNA-target predictions for the given clade, and each differently coloured segment within the pie chart represents a clade of origin. The clade colour code used here is identical to that used in Figure 1.6.

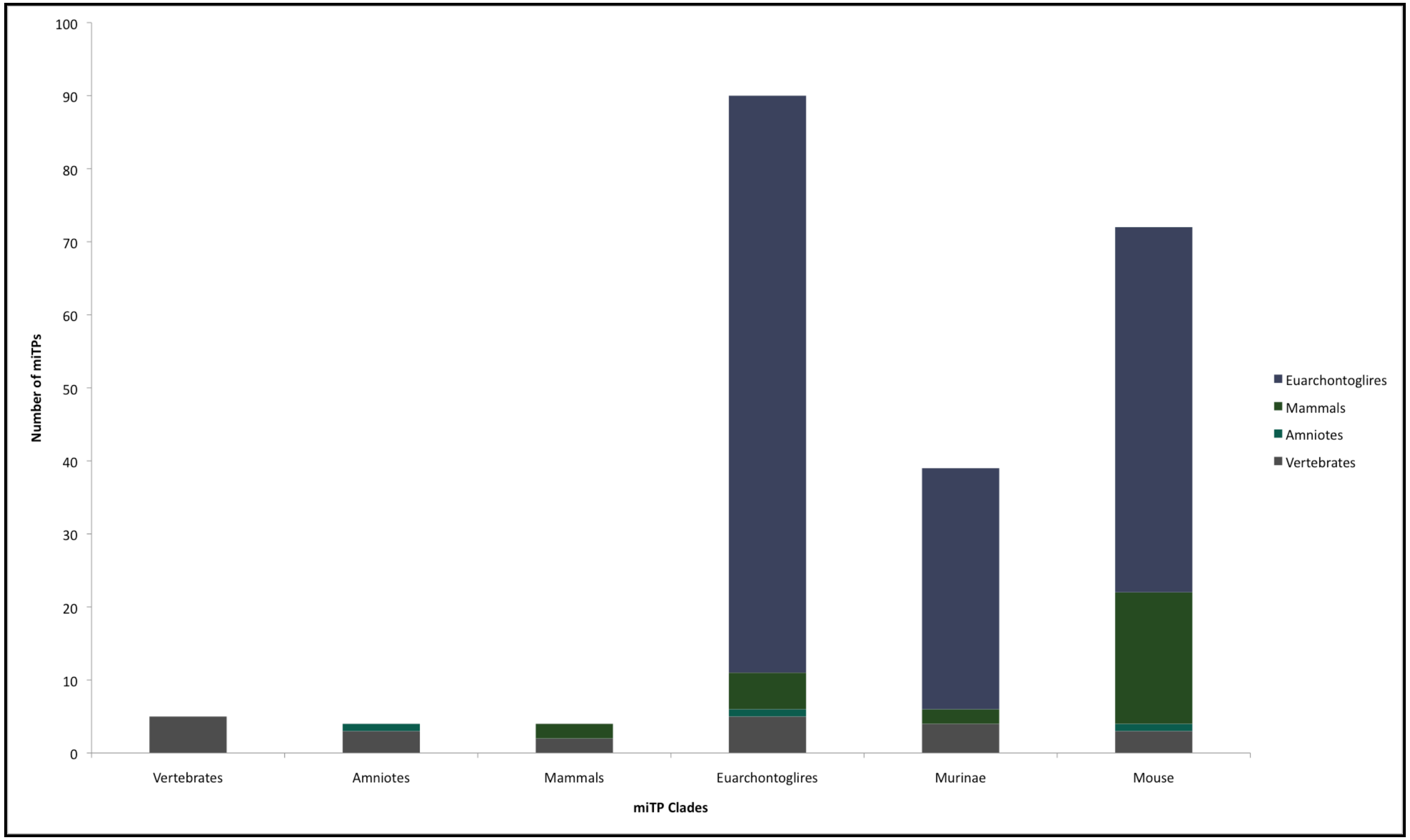


Figure 4.11: Origin of mouse placental microRNAs and their target sites.

Each stacked column above represents the number of miRNA-target predictions for the given clade, and each differently coloured column within the stacked column represents a clade of origin (i.e. the clade in which both the miRNA and target gene were present). The clade colour code used here is identical to that used in Figure 1.6.

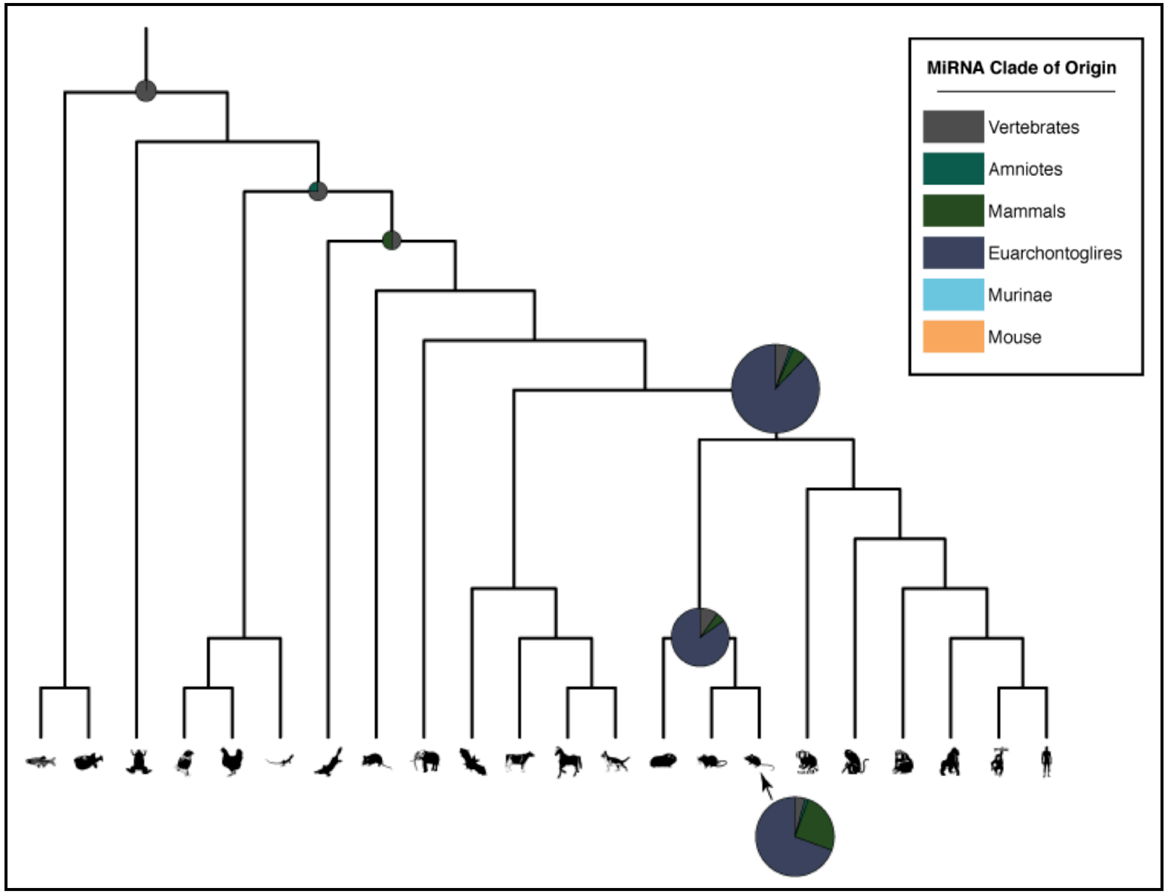


Figure 4.12: Phylogeny depicting origin of mouse microRNAs and their target sites.

Shown is a phylogeny of all 22 species, overlaid with pie charts, such that the area of each pie chart represents the number of miRNA-target predictions for the given clade, and each differently coloured segment within the pie chart represents a clade of origin (i.e. the clade in which both the miRNA and target gene were present). The clade colour code used here is identical to that used in Figure 1.6.

Figure 4.13 shows a heatmap diagram of the miRNA-target predictions made for the C19MC cluster against the set of 114 placental genes (the 24 genes for which no miRNA-target prediction was made are not included in Figure 4.13). Each column shows miRNA-target predictions for a specific placental gene, while each row shows predicted miRNA-target interactions for a specific miRNA. Each cell is divided into two triangles: the red triangle in the upper-left of each cell gives the miRNA-target prediction result for the TargetScan context script, while the blue triangle to the lower-right of each cell gives the miRNA-target prediction result for the TargetScan conservation script. Brighter values indicate a stronger miRNA-target prediction — stronger miRNA-target repression in the case of the TargetScan context script, stronger miRNA-target conservation in the case of the TargetScan conservation script.

Figure 4.13 shows that no preferentially conserved miRNA-target interactions were predicted by the TargetScan conservation script for any of the miRNAs in the C19MC cluster. This is not unexpected, since the cluster itself is not conserved across an appreciable branch length on the TargetScan vertebrate phylogeny and the TargetScan conservation script tends to make a relatively small number of high-confidence miRNA-target predictions. A trend that is noticeable among the miRNA-target predictions made by the TargetScan context script is the numerous patterns of multiple miRNAs being predicted to interact with the same gene, or of multiple genes predicted to interact with the same miRNA. These can be seen as small columns or rows (respectively) of miRNA-target predictions within the heatmap diagram. These might be indicative of coordinated regulation of a gene by multiple miRNAs — as observed by, for example, Wu *et al.* (2010) — or of the regulation by a single miRNA of the expression of multiple genes in a coordinated manner.

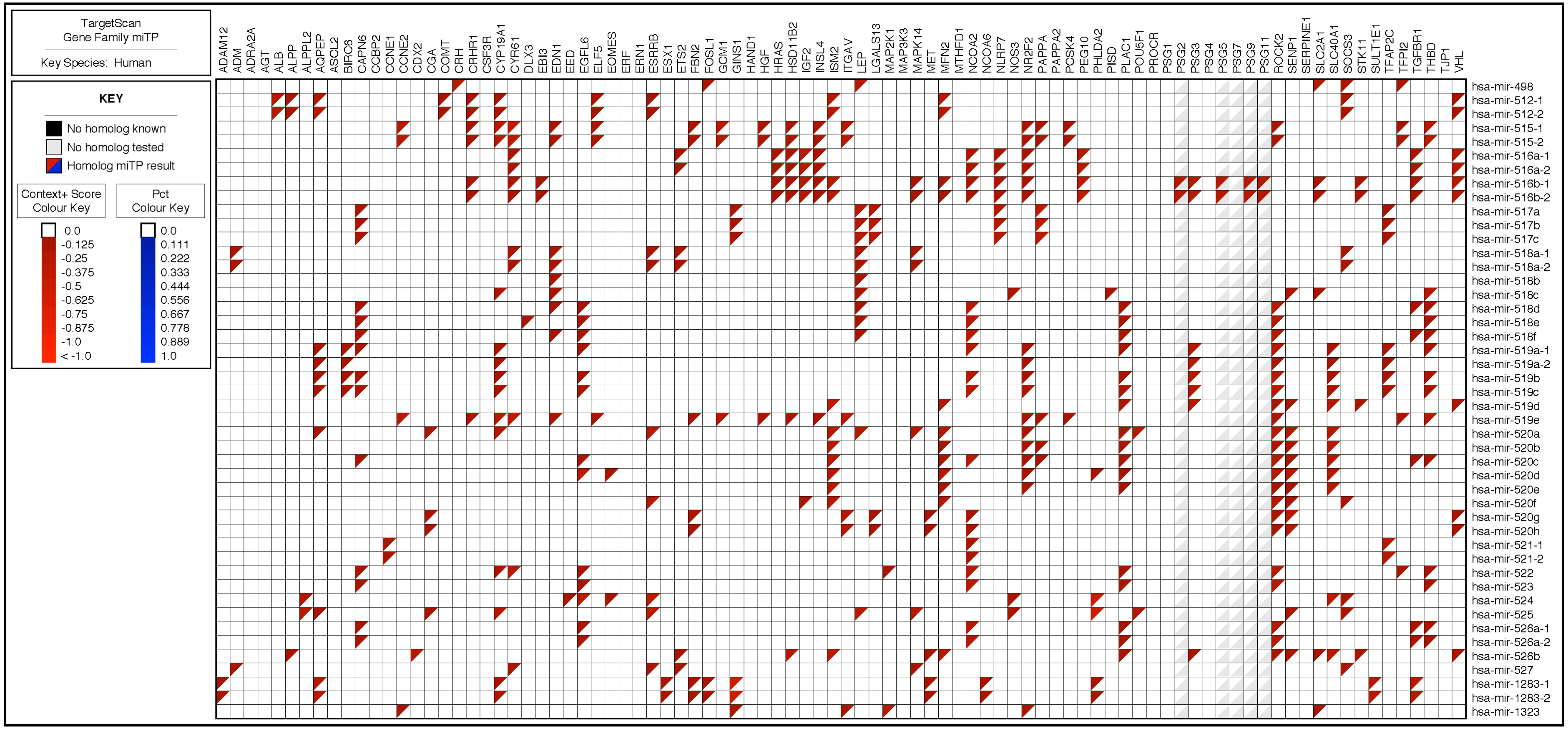


Figure 4.13: MicroRNA-target predictions for C19MC microRNAs in human.

The above figure shows predicted miRNA-target interactions for the chromosome 19 microRNA cluster (C19MC) among the set of placental genes. Each column shows results for a particular gene, while each row shows predictions for a given miRNA. Each red triangle represents a miRNA-target prediction made by the TargetScan Context+ script (Garcia *et al.* 2011, Grimson *et al.* 2007).

4.5 Discussion

The patterns of homology identified for each of the miRNAs in this chapter are in broad agreement with the hypothesis that once a miRNA emerges, substitutions within functionally important regions are rare and secondary loss of the miRNA is rarer still (Sperling and Peterson 2009). Similarly, patterns of miRNA-target prediction, involving homologous miRNAs and target genes, are also observed across clades, so that for many monophyletic groups of miRNAs, interactions between the miRNA and its target genes evolved relatively rapidly after the evolution of the miRNA itself. Furthermore, miRNA targeting patterns form a complex regulatory network, with coordinated regulation of genes by multiple miRNAs and of parallel control by one miRNA of many genes.

The patterns of miRNA homology can be seen in Figure 4.7. For example, hsa-mir-517a is conserved in primate species but found in no other genome, while the let-7 miRNA family is conserved throughout all vertebrate species studied. Some exceptions to this trend are noticeable in Figure 4.7. For example, many miRNAs otherwise conserved throughout Eutheria appear to be absent from bat (e.g. hsa-mir-23a, hsa-mir-26b). This may indicate lineage-specific loss of these miRNAs; alternatively it may reflect the level of sequence coverage or quality in the genome in question, or possibly a false-negative prediction by the homolog search method used. The apparent homology of a frog sequence to hsa-mir-520e may similarly indicate a distant homology, or it may possibly be more reflective of a false-positive prediction by the homolog search method. Although both implementations take account of BLAST E-values, it should be noted that those E-values apply to the sequence alignment only and not to the miRNA homology prediction as a whole. For example, neither miRNAminer or the emulator described in this chapter estimate E-values for miRNA homology that take account of the expected frequency of hairpin structure or other miRNA features in addition to that of a high-scoring sequence alignment (Artzi *et al.* 2008). Such a comprehensive statistic would be highly desirable, and would facilitate more accurate estimation of the extent of false-positive and false-negative homology predictions.

The patterns of emergence of miRNA-target interactions can be seen in Figure 4.9, Figure 4.10, Figure 4.11, and Figure 4.12, while the patterns of predicted human miRNA-target interactions involving C19MC miRNAs are shown in Figure 4.13. False-positive prediction rates among miRNA-target prediction methods remain quite high, with estimates ranging from 28% (Lewis *et al.* 2005) to 67% (Baek *et al.* 2008). Even with the most accurate method, a set of miRNA-target predictions might be better seen as being enriched for miRNA-target interactions rather than constituting a set of miTIs *per se*. Nevertheless, even if only a fraction of the predicted miRNA-target interactions presented in Figure 4.13 are functional *in vivo*, it is difficult to imagine that the picture presented — of coordinated regulation of genes by multiple miRNAs and of parallel control by one miRNA of many genes — would not remain. The validation of an interaction between, say, a placenta-critical gene such as leptin (LEP) or a placenta-specific gene such as isthmin 2 (ISM2) with multiple miRNAs in the C19MC cluster would constitute compelling evidence that the expression of these genes is finely controlled by the miRNAs of the C19MC.

Some of these common miRNA-target predictions reflect a shared miRNA seed sequence across multiple miRNAs, although this is not always the case. For example, the mir-517 miRNAs predicted to target LEP share a seed with each other, as do the mir-518 miRNAs — except for hsa-mir-518e, which appears to have undergone a seed shift in the 3′ direction — but the seed sequences of these two sets of miRNAs differ from each other. Furthermore, miRNAs hsa-mir-518d, hsa-mir-518e and hsa-mir-518f are all predicted by TargetScan to target LEP with both arms of the miRNA hairpin (i.e. both mature miRNAs). That the expression of this gene would be tightly regulated is not unexpected, given its key role in pregnancy in Eutheria and especially among primates; leptin is known to play a key role in implantation, trophoblast development and placental function throughout pregnancy in a time-dependent manner (Henson and Castracane 2006, Maymó *et al.* 2011). Its expression at optimal levels and at the right time would therefore be critical for healthy placental development. If validated, the regulation of LEP by the mir-517 miRNAs, and the nature of its response to each of the individual miRNAs, would augment current knowledge of how this key protein is regulated in placenta.

With those miRNAs that share a seed sequence, it has been proposed that it would be possible to distinguish which member of a miRNA family is most likely to represent a functional regulator of a given target by taking account of the level of Watson-Crick complementarity of the 3′ region of the miRNA to the target gene sequence (Brennecke *et al.* 2005).

In Figure 4.9 and Figure 4.10, the overall high proportion of primate-specific miRNAs with miRNA-target interactions conserved throughout primates is likely affected by the presence of the 46 miRNAs from the primate-specific C19MC cluster. Nevertheless, this indicates that many of these primate-specific miRNAs acquired miRNA-target interactions in a relatively short space of time after their emergence, which supports the argument by Chen and Rajewsky (2007) and Shomron *et al.* (2009) that miRNA-target sites can emerge quite rapidly in evolutionary terms. Their tissue-specific origin is also congruent with the hypothesis of miRNA and miRNA-target emergence posited by Chen and Rajewsky (2007) — see Section 1.9.6 — although broader expression patterns have not yet been observed. Whether the C19MC cluster will be found to act in tissues other than the placenta remains to be seen, but in any case, there is little doubt that this cluster of small miRNAs has made a big impact on the placenta (Flor and Bullerdiek 2012).

In summary, this chapter has shown that many — though not all — placental miRNAs are conserved throughout Eutheria but confined to that clade, and that for many of these miRNAs — though not all — interactions with target genes evolved relatively rapidly after the emergence of the miRNA itself. This pattern is recapitulated by the miRNAs in the C19MC — conserved throughout primates but also confined to that lineage — for which many miRNA-target interactions arose almost as soon as the miRNAs themselves.

Chapter 5: Discussion

In a critique of the widespread use of selective pressure analysis, with codon based methods using maximum likelihood, to identify instances of positive selection and associated functional shift, Hughes and Friedman (2010) argue that an excess of non-synonymous substitutions over synonymous substitutions does not necessarily imply positive selection, that in any case such an excess will be masked by more recent synonymous substitutions over time, and that positive selection by sequence substitution is just one of many possible mechanisms of functional shift. The authors list several alternative mechanisms by which novelty can arise in the evolution of a protein-coding gene: substitution at a single amino acid site, sequence insertion or deletion, gene fusion, loss of splice signals with consequent loss of exons, and regulatory changes that alter the expression of the gene (Hughes and Friedman 2010). Hughes and Friedman (2010) go on to argue that such alternative mechanisms, although less readily analysed by bioinformatic methods, "almost certainly account for the overwhelming majority of cases of positive selection on duplicate genes" (Hughes and Friedman 2010, p.88).

This issue is akin to the well-known metaphor of the drunk under the streetlight, searching for his lost keys only in the small area lit by the streetlight, because that is the easiest place to search. While it is arguably the case that statistical methods for identifying positive selection by comparison of models of codon substitution have a longer history and are better developed than methods to identify positive selection mediated by other mechanisms such as gene fusion, exon shuffling and regulatory innovation, there has been progress in development of such methods. Bioinformatics approaches and resources are available for identifying gene fusion events (Kim *et al.* 2010) and cases of exon shuffling (Cancherini *et al.* 2010). While indels have not been used to analyse the selective pressure acting on a gene, the explicit treatment of alignment gaps as evolutionary events has been used to improve multiple sequence alignment (Löytynoja and Goldman 2008), and there is certainly potential for alignment gaps to inform phylogenetic analyses (Dessimoz and Gil 2010). In terms of regulatory innovation, challenges remain in computational prediction of transcription factor binding sites (Wasserman and Sandelin 2004, Tompa *et al.* 2005), but considerable progress has been made during the last decade in the development of miRNA-target prediction methods (Reyes-Herrera and Ficarra 2012). Thus, the set of methods for identifying and characterising evolutionary novelty progress at different rates in different areas, so the methods available will inevitably be incomplete until sufficient progress is made in all areas. However, this should not be used as an argument against using those methods that are currently amenable to bioinformatics analysis. In the metaphor of the drunk under the streetlight, as progress in bioinformatics sheds new light on an ever expanding space, we can explore the newly elucidated areas while acknowledging that our search is far from exhaustive.

In this thesis, we have sought to make use of complementary approaches to estimate the extent of functional shift in placental genes in the Eutherian lineage: in Chapter 2, a selective pressure analysis based on codon substitution models, and in Chapter 4, an analysis of miRNA-target interactions within the gene families of each placental gene. This aims to address two of the mechanisms of evolutionary novelty listed by Hughes and Friedman (2010). The former looked at the changes made directly to placental gene sequences, while the latter examined putative changes to the regulation of placental gene expression.

Care was taken in both major strands of the project to ensure that the methodology used would minimise errors and false-positive predictions. In Chapter 2, the placental gene homologs used were those identified by the Ensembl Compara pipeline (Vilella *et al.* 2009), currently the only genome-scale homology resource that uses tree reconciliation (Altenhoff and Dessimoz 2009); the sequences of each gene family were aligned using an alignment method that has performed well as a preparatory step for selective pressure analysis (Fletcher and Yang 2010, Markova-Raina and Petrov 2011); and selective pressure analysis itself was performed using software that is both accurate and powerful (Anisimova *et al.* 2001, Yang *et al.* 2009, Zhai *et al.* 2012) and whose predictions of positive selection have been validated *in vitro* (Levasseur *et al.* 2006, Huang *et al.* 2012, Loughran *et al.* 2012). Although false-positive rates were not explicitly estimated during selective pressure analysis, in a comparable analysis by Fletcher and Yang (2010) false-positive estimates ranged from 2% to 29%.

In Chapter 4, placental miRNA homologs were obtained from three sources — Ensembl, miRBase and a homolog search using miRNAminerEmulator.pl (an implementation of the miRNAminer algorithm); and the miRNA-target prediction method used was chosen in a performance benchmark of currently available miRNA-target prediction methods, which found TargetScan to be the best-performing of the methods compared — this was the subject of Chapter 3.

Despite the steps taken to minimise errors, positive selection was inferred due to an alignment error in at least one instance (see Section 2.4.2), and there are likely to be some errors in miRNA homology prediction (see Section 4.4.1) and miRNA-target prediction, for which false-positive rates are known to be high (Bartel 2009, Peter 2010) — for example, the ratio of predicted targets to estimated false-positives was 7:2 in a study by Lewis *et al.* (2005), and even the most accurate methods were found to have false discovery rates of approximately 67% by Baek *et al.* (2008). Nevertheless, such errors are unlikely to be so prevalent as to substantially affect the broad conclusions derived from these analyses.

The results of Chapter 2 are in broad agreement with those of Hou *et al.* (2009), with higher estimates of positive selection at the Eutherian stem in this thesis — 44.1% as opposed to the 27.9% observed by Hou *et al.* (2009) — perhaps reflecting the presence of multi-gene families (gene families with one-to-many homologs). When the results in Chapter 2 for the subset of gene families with one-to-one orthologs are compared with the results of Hou *et al.* (2009), the trend is less marked: estimated rates of positive selection in Chapter 2 drop to 36.1%, higher than those of Hou *et al.* (2009), but not significantly so ($\alpha = 0.05$). Whether the different results in this thesis are due to the choice of alignment method is debatable, but since the alignment method used was one of the key methodological differences between this study and that of Hou *et al.* (2009), it is likely that the alignment method used had — and can have — a considerable impact on the detection of positive selection.

In Chapter 4, a combination of miRNA homology search and miRNA-target prediction identified placental miRNA homologs and their targets. The results of these analyses showed that many miRNAs evolved targets relatively soon after their own emergence. This was especially apparent in the case of the C19MC miRNAs, a cluster of miRNAs identified by (Bentwich *et al.* 2005) that are primate-specific (Bentwich *et al.* 2005) (Zhang *et al.* 2008). As the results of Chapter 4 indicate, these miRNAs also have many miRNA-target interactions that are predicted to occur throughout primates; subject to *in vitro* validation, this would confirm that these miRNAs acquired targets relatively swiftly following the emergence and expansion of the C19MC cluster.

The miRNA-target predictions for miRNAs specific to placental mammals broadly recapitulated the pattern of the C19MC cluster, with an initial burst of new miRNA-target interactions shortly after their emergence, followed by a more or less continuous acquisition of targets throughout subsequent evolution (see Figure 4.9 and Figure 4.10).

The omission from the miRNA-target prediction benchmark in Chapter 3 of PicTar, EIMMo and DIANA-microT is unfortunate, especially considering the good performance of these three methods in previous benchmark studies (Sethupathy *et al.* 2006, Baek *et al.* 2008, Selbach *et al.* 2008, Alexiou *et al.* 2009). Nevertheless, many — though not all — of the features exploited by these methods are also used by at least one of the methods compared in Chapter 3.

Active development continues of both new and existing miRNA-target prediction methods, and as new mechanisms and features of miRNA-target interactions are uncovered experimentally, it is reasonable to expect that these will be employed in future iterations of miRNA-target prediction software. Recently identified miRNA-target site types such as bulge sites (Chi *et al.* 2012) and centred pairing sites (Shin *et al.* 2010) may be incorporated in future miRNA-target prediction methods.

For identification of miRNA-target sites within coding regions, contextual features that could potentially be discriminative of functional target sites might include repeat sequences, since at least one miRNA has been observed to target repeat-rich coding regions enriched for miRNA target sites (Schnall-Levin *et al.* 2011), or the presence of rare codons for the given organism that induce ribosome stalling and allow the miRNA to bind to its target site, as achieved in an experiment by Gu *et al.* (2009). (Whether these features are characteristic of miRNA-target interactions in coding regions *in vivo* remains to be seen, but they do present an avenue for investigation.)

As experimental evidence sheds further light on these mechanisms, this expands the space that can be explored by bioinformatics, which may in turn point to promising avenues for further experimentation. To follow the metaphor that introduced this discussion, while we may be restricted to looking within the extent of the streetlight, we can at least work to expand the reach of its illumination.

Chapter 6: Conclusion

6.1 Conclusion

In conclusion, in the first selective pressure analysis of placental genes that includes multi-gene families, estimated levels of positive selection in placental genes were found to be higher than previous estimates, with nearly half of such genes undergoing positive selection in the ancestral Eutherian lineage. By comparison with available functional annotation, it has been possible to establish in this thesis that at least some of these are attributable to functional shift in these genes at the time of the Eutherian ancestor.

Similarly, 14 placental miRNAs arose in the Eutherian clade and remain conserved throughout Eutheria (or nearly so), as did their miRNA-target interactions (although the individual sites may differ); this pattern has also been followed by the C19MC cluster of placental miRNAs. Whether the miRNA-target interactions predicted in this thesis are confirmed to be *bona fide* awaits experimental validation, but in any case great care was taken to choose the miRNA-target prediction method with the highest specificity for a reasonable level of sensitivity.

Of the 9 miRNA-target prediction methods benchmarked here, this was found to be TargetScan, although it should be noted that its performance in ROC analysis was not significantly better, at the 5% significance level, than miRanda, miRmap or Hitsensor. It is debatable whether TargetScan would perform better than these three other methods with a different set of validated miRNA-target interactions. In any case, there can be little doubt that as the set of validated miRNA-target interactions increases in size, our ability to distinguish the performance of existing miRNA-target prediction methods will grow too.

6.2 Future Work

In any future comparison of miRNA-target prediction methods, the notable omission of miRNA-target prediction methods such as PicTar (Chen *et al.* 2006, Lall *et al.* 2006, Grün *et al.* 2005, Krek *et al.* 2005), EIMMo (Gaidatzis *et al.* 2007), and DIANA-microT (Reczko *et al.* 2012, Reczko *et al.* 2011, Maragkakis *et al.* 2009a, Maragkakis *et al.* 2009b), should be remedied. One possible way of including these methods would be to perform automated queries to the web interface of each method. However, this will not provide a remedy in all cases, since many such web query systems simply provide access to precalculated miRNA-target predictions, rather than allowing novel predictions with user-specified miRNA and target sequences.

A second avenue for improvement of the miRNA-target prediction benchmark would be to filter validated miRNA-target interactions to identify only those for which a miRNA-target prediction method was not used in the process of selecting validation candidates. This subset of validated miTIs would facilitate the exclusion of that potential source of bias, as was done by Sethupathy *et al.* in 2006, with the smaller set of validations available at that time.

In relation to the placenta-specific miRNA-target interactions, *in vitro* validation of predicted miRNA-target interactions would be of benefit in confirming the accuracy of those miRNA-target predictions. Further benefit may be obtained through analysis of the predicted miRNA-target interactions from the perspective of the miRNA-target interaction network, to tease apart the patterns of parallel regulation by one miRNA of many genes and coordinated control of genes by multiple miRNAs.

Chapter 7: Bibliography

Aas, G. (2012) LWP::UserAgent. 6.04 ed.: CPAN.

Abril, J. F., Castelo, R. and Guigó, R. (2005) Comparison of splice sites in mammals and chicken. *Genome Research,* 15**,** 111-119.

Akaike, H. (1973) Information theory and an extension of the maximum likelihood principle. *Second International Symposium on Information Theory.* Budapest.

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

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

Alexiou, P., Maragkakis, M., Papadopoulos, G. L., Reczko, M. and Hatzigeorgiou, A. G. (2009) Lost in translation: an assessment and perspective for computational microRNA target identification. *Bioinformatics,* 25**,** 3049-3055.

Allen, E., Xie, Z., Gustafson, A. M., Sung, G.-H., Spatafora, J. W. and Carrington, J. C. (2004) Evolution of microRNA genes by inverted duplication of target gene sequences in Arabidopsis thaliana. *Nature Genetics,* 36**,** 1282-1290.

Altekar, G., Dwarkadas, S., Huelsenbeck, J. P. and Ronquist, F. (2004) Parallel Metropolis coupled Markov chain Monte Carlo for Bayesian phylogenetic inference. *Bioinformatics,* 20**,** 407-415.

Altenhoff, A. M. and Dessimoz, C. (2009) Phylogenetic and functional assessment of orthologs inference projects and methods. *PLoS Computational Biology,* 5**,** e1000262.

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**,** e1002514.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. and Lipman, D. J. (1990) Basic local alignment search tool. *Journal of Molecular Biology,* 215**,** 403-410.

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D. J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research,* 25**,** 3389-3402.

Altuvia, Y. (2005) Clustering and conservation patterns of human microRNAs. *Nucleic Acids Research,* 33**,** 2697-2706.

Anderson, D. R. (2008) *Model based inference in the life sciences: a primer on evidence.,* New York, Springer-Verlag.

Anderson, E. M., Birmingham, A., Baskerville, S., Reynolds, A., Maksimova, E., Leake, D., Fedorov, Y., Karpilow, J. and Khvorova, A. (2008) Experimental validation of the importance of seed complement frequency to siRNA specificity. *RNA,* 14**,** 853-861.

Anisimova, M., Bielawski, J. P. and Yang, Z. (2001) Accuracy and power of the likelihood ratio test in detecting adaptive molecular evolution. *Molecular Biology and Evolution,* 18**,** 1585-1592.

Anisimova, M., Cannarozzi, G. and Liberles, D. A. (2010) Finding the balance between the mathematical and biological optima in multiple sequence alignment. *Trends in Evolutionary Biology,* 2.

Anisimova, M. and Liberles, D. A. (2007) The quest for natural selection in the age of comparative genomics. *Heredity,* 99**,** 567-579.

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**,** 1229-36.

Artzi, S., Kiezun, A. and Shomron, N. (2008) miRNAminer: a tool for homologous microRNA gene search. *BMC Bioinformatics,* 9**,** 39.

Arvey, A., Larsson, E., Sander, C., Leslie, C. S. and Marks, D. S. (2010) Target mRNA abundance dilutes microRNA and siRNA activity. *Molecular Systems Biology,* 6**,** 1-7.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M. and Sherlock, G. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genetics,* 25**,** 25-29.

Asher, R. J., Bennett, N. and Lehmann, T. (2009) The new framework for understanding placental mammal evolution. *BioEssays,* 31**,** 853-864.

Aukerman, M. J. and Sakai, H. (2003) Regulation of flowering time and floral organ identity by a MicroRNA and its APETALA2-like target genes. *The Plant cell,* 15**,** 2730-2741.

Axtell, M. J., Westholm, J. O. and Lai, E. C. (2011) Vive la différence: biogenesis and evolution of microRNAs in plants and animals. *Genome Biology,* 12**,** 221.

Baek, D., Villén, J., Shin, C., Camargo, F. D., Gygi, S. P. and Bartel, D. P. (2008) The impact of microRNAs on protein output. *Nature,* 455**,** 64-71.

Bairoch, A. and Boeckmann, B. (1992) The SWISS-PROT protein sequence data bank. *Nucleic Acids Research,* 20**,** 2019.

Bandyopadhyay, S. and Mitra, R. (2009) TargetMiner: microRNA target prediction with systematic identification of tissue-specific negative examples. *Bioinformatics,* 25**,** 2625-2631.

Bandyopadhyay, S., Saha, S., Maulik, U. and Deb, K. (2008) A simulated annealing-based multiobjective optimization algorithm: AMOSA. *Evolutionary Computation, IEEE Transactions on,* 12**,** 269-283.

Bar, M., Wyman, S. K., Fritz, B. R., Qi, J., Garg, K. S., Parkin, R. K., Kroh, E. M., Bendoraite, A., Mitchell, P. S., Nelson, A. M., Ruzzo, W. L., Ware, C., Radich, J. P., Gentleman, R., Ruohola-Baker, H. and Tewari, M. (2008) MicroRNA discovery and profiling in human embryonic stem cells by deep sequencing of small RNA libraries. *Stem Cells,* 26**,** 2496-2505.

Barash, Y., Calarco, J. A., Gao, W., Pan, Q., Wang, X., Shai, O., Blencowe, B. J. and Frey, B. J. (2010) Deciphering the splicing code. *Nature,* 465**,** 53-59.

Bartel, D. P. (2004) MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell,* 116**,** 281-297.

Bartel, D. P. (2009) MicroRNAs: target recognition and regulatory functions. *Cell,* 136**,** 215-233.

Bartel, D. P. and Chen, C. Z. (2004) Micromanagers of gene expression: the potentially widespread influence of metazoan microRNAs. *Nature Reviews Genetics,* 5**,** 396-400.

Bateman, A., Birney, E., Cerruti, L., Durbin, R., Etwiller, L., Eddy, S. R., Griffiths-Jones, S., Howe, K. L., Marshall, M. and Sonnhammer, E. L. L. (2002) The Pfam protein families database. *Nucleic Acids Research,* 30**,** 276-280.

Batzer, M. A. and Deininger, P. L. (2002) Alu repeats and human genomic diversity. *Nature Reviews Genetics,* 3**,** 370-379.

Beaumont, M. A. and Rannala, B. (2004) The Bayesian revolution in genetics. *Nature Reviews Genetics,* 5**,** 251-261.

Benirschke, K. (2007) *Comparative placentation* [Online]. San Diego: UCSD. Available: http://placentation.ucsd.edu/ [Accessed 2011].

Benirschke, K. and Kaufman, P. (2000) *Pathology of the human placenta,* New York, Springer-Verlag.

Benjamini, Y. and Hochberg, Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society B: Methodological***,** 289-300.

Benton, M. J. and Donoghue, P. C. J. (2007) Paleontological evidence to date the tree of life. *Molecular Biology and Evolution,* 24**,** 26-53.

Bentwich, I., Avniel, A., Karov, Y., Aharonov, R., Gilad, S., Barad, O., Barzilai, A., Einat, P., Einav, U., Meiri, E., Sharon, E., Spector, Y. and Bentwich, Z. (2005) Identification of hundreds of conserved and nonconserved human microRNAs. *Nature Genetics,* 37**,** 766-770.

Berezikov, E. (2011) Evolution of microRNA diversity and regulation in animals. *Nature Reviews Genetics,* 12**,** 846-860.

Berezikov, E., Chung, W.-J., Willis, J., Cuppen, E. and Lai, E. C. (2007) Mammalian mirtron genes. *Molecular Cell,* 28**,** 328-336.

Berezikov, E., Liu, N., Flynt, A. S., Hodges, E., Rooks, M., Hannon, G. J. and Lai, E. C. (2010) Evolutionary flux of canonical microRNAs and mirtrons in Drosophila. *Nature Genetics,* 42**,** 6-9.

Berg, J., Willmann, S. and Lässig, M. (2004) Adaptive evolution of transcription factor binding sites. *BMC Evolutionary Biology,* 4**,** 42.

Bernstein, E., Caudy, A. A., Hammond, S. M. and Hannon, G. J. (2001) Role for a bidentate ribonuclease in the initiation step of RNA interference. *Nature,* 409**,** 363-366.

Betel, D., Koppal, A., Agius, P., Sander, C. and Leslie, C. (2010) Comprehensive modeling of microRNA targets predicts functional non-conserved and non-canonical sites. *Genome Biology,* 11**,** R90.

Betel, D., Wilson, M., Gabow, A., Marks, D. S. and Sander, C. (2008) The microRNA.org resource: targets and expression. *Nucleic Acids Research,* 36**,** D149-53.

Betts, M. J. and Russell, R. B. (2003) Amino acid properties and consequences of substitutions. Wiley New Jersey.

Bielawski, J. P., Dunn, K. A. and Yang, Z. (2000) Rates of nucleotide substitution and mammalian nuclear gene evolution. Approximate and maximum-likelihood methods lead to different conclusions. *Genetics,* 156**,** 1299-1308.

Bielawski, J. P. and Yang, Z. (2003) Maximum likelihood methods for detecting adaptive evolution after gene duplication. *Journal of Structural and Functional Genomics,* 3**,** 201-12.

Birmingham, A., Anderson, E. M., Reynolds, A., Ilsley-Tyree, D., Leake, D., Fedorov, Y., Baskerville, S., Maksimova, E., Robinson, K., Karpilow, J., Marshall, W. S. and Khvorova, A. (2006) 3′ UTR seed matches, but not overall identity, are associated with RNAi off-targets. *Nature Methods,* 3**,** 199-204.

Borchert, G. M., Lanier, W. and Davidson, B. L. (2006) RNA polymerase III transcribes human microRNAs. *Nature Structural & Molecular Biology,* 13**,** 1097-1101.

Bortolin-Cavaille, M. L., Dance, M., Weber, M. and Cavaille, J. (2009) C19MC microRNAs are processed from introns of large Pol-II, non-protein-coding transcripts. *Nucleic Acids Research,* 37**,** 3464-3473.

Braun, J. T. and Conway, D. (2012) Parse::RecDescent. 1.967009 ed.: CPAN.

Brennecke, J., Hipfner, D. R., Stark, A., Russell, R. B. and Cohen, S. M. (2003) *bantam* encodes a developmentally regulated microRNA that controls cell proliferation and regulates the proapoptotic gene *hid* in Drosophila. *Cell,* 113**,** 25-36.

Brennecke, J., Stark, A., Russell, R. B. and Cohen, S. M. (2005) Principles of microRNA-target recognition. *PLoS Biology,* 3**,** e85.

Breving, K. and Esquela-Kerscher, A. (2010) The complexities of microRNA regulation: mirandering around the rules. *The International Journal of Biochemistry & Cell Biology,* 42**,** 1316-1329.

Brown, T. A. (2002) *Genomes,* New York, Wiley-Liss.

Brown, W. M., Prager, E. M., Wang, A. and Wilson, A. C. (1982) Mitochondrial DNA sequences of primates: tempo and mode of evolution. *Journal of Molecular Evolution,* 18**,** 225-239.

Bruno, W. J. and Halpern, A. L. (1999) Topological bias and inconsistency of maximum likelihood using wrong models. *Molecular Biology and Evolution,* 16**,** 564-566.

Buneman, O. P. (1971) The recovery of trees from measures of dissimilarity. *Mathematics the the archeological and historical sciences.* Edinburgh: Edinburgh University Press.

Burnham, K. P. and Anderson, D. R. (2002) *Model selection and multimodel inference: a practical information-theoretic approach.,* New York, Springer.

Butlin, R. K., Galindo, J. and Grahame, J. W. (2008) Review. Sympatric, parapatric or allopatric: the most important way to classify speciation? *Philosophical Transactions of the Royal Society B: Biological Sciences,* 363**,** 2997-3007.

Calin, G. A., Dumitru, C. D., Shimizu, M., Bichi, R., Zupo, S., Noch, E., Aldler, H., Rattan, S., Keating, M., Rai, K., Rassenti, L., Kipps, T., Negrini, M., Bullrich, F. and Croce, C. M. (2002) Frequent deletions and down-regulation of micro- RNA genes miR15 and miR16 at 13q14 in chronic lymphocytic leukemia. *Proceedings of the National Academy of Sciences of the United States of America,* 99**,** 15524-15529.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K. and Madden, T. L. (2009) BLAST+: architecture and applications. *BMC Bioinformatics,* 10**,** 421.

Campo-Paysaa, F., Sémon, M., Cameron, R. A., Peterson, K. J. and Schubert, M. (2011) MicroRNA complements in deuterostomes: origin and evolution of microRNAs. *Evolution & Development,* 13**,** 15-27.

Cancherini, D. V., França, G. S. and de Souza, S. J. (2010) The role of exon shuffling in shaping protein-protein interaction networks. *BMC Genomics,* 11 Suppl 5**,** S11.

Cao, H., Yang, C.-s. and Rana, T. M. (2008) Evolutionary emergence of microRNAs in human embryonic stem cells. *PLoS ONE,* 3**,** e2820.

Cavalli-Sforza, L. L. and Edwards, A. W. (1967) Phylogenetic analysis. Models and estimation procedures. *American Journal of Human Genetics,* 19**,** 233-257.

Chang, J., Nicolas, E., Marks, D., Sander, C., Lerro, A., Buendia, M. A., Xu, C., Mason, W. S., Moloshok, T., Bort, R., Zaret, K. S. and Taylor, J. M. (2004) miR-122, a mammalian liver-specific microRNA, is processed from hcr mRNA and may downregulate the high affinity cationic amino acid transporter CAT-1. *RNA Biology,* 1**,** 106-113.

Chao, A., Tsai, C.-L., Wei, P.-C., Hsueh, S., Chao, A.-S., Wang, C.-J., Tsai, C.-N., Lee, Y.-S., Wang, T.-H. and Lai, C.-H. (2010) Decreased expression of microRNA-199b increases protein levels of SET (protein phosphatase 2A inhibitor) in human choriocarcinoma. *Cancer Letters,* 291**,** 99-107.

Cheloufi, S., Dos Santos, C. O., Chong, M. M. W. and Hannon, G. J. (2010) A Dicer-independent miRNA biogenesis pathway that requires Ago catalysis. *Nature,* 465**,** 584-589.

Chen, C. Z. (2004a) microRNAs modulate hematopoietic lineage differentiation. *Science,* 303**,** 83-86.

Chen, K. and Rajewsky, N. (2006) Natural selection on human microRNA binding sites inferred from SNP data. *Nature Genetics,* 38**,** 1452-1456.

Chen, K. and Rajewsky, N. (2007) The evolution of gene regulation by transcription factors and microRNAs. *Nature Reviews Genetics,* 8**,** 93-103.

Chen, K., Wang, Y. L., Dewey, C. N., Sood, P., Colombo, T., Bray, N., MacMenamin, P., Kao, H. L., Gunsalus, K. C. and Pachter, L. (2006) A genome-wide map of conserved microRNA targets in C. elegans. *Current Biology,* 16**,** 460-471.

Chen, X. (2004b) A microRNA as a translational repressor of APETALA2 in Arabidopsis flower development. *Science,* 303**,** 2022-2025.

Chi, S. W., Hannon, G. J. and Darnell, R. B. (2012) An alternative mode of microRNA target recognition. *Nature Structural & Molecular Biology,* 19**,** 321-327.

Chi, S. W., Zang, J. B., Mele, A. and Darnell, R. B. (2009) Argonaute HITS-CLIP decodes microRNA-mRNA interaction maps. *Nature,* 460**,** 479-486.

Church, D. M., Schneider, V. A., Graves, T., Auger, K., Cunningham, F., Bouk, N., Chen, H.-C., Agarwala, R., McLaren, W. M., Ritchie, G. R. S., Albracht, D., Kremitzki, M., Rock, S., Kotkiewicz, H., Kremitzki, C., Wollam, A., Trani, L., Fulton, L., Fulton, R., Matthews, L., Whitehead, S., Chow, W., Torrance, J., Dunn, M., Harden, G., Threadgold, G., Wood, J., Collins, J., Heath, P., Griffiths, G., Pelan, S., Grafham, D., Eichler, E. E., Weinstock, G., Mardis, E. R., Wilson, R. K., Howe, K., Flicek, P. and Hubbard, T. (2011) Modernizing reference genome assemblies. *PLoS Biology,* 9**,** e1001091.

Cifuentes, D., Xue, H., Taylor, D. W., Patnode, H., Mishima, Y., Cheloufi, S., Ma, E., Mane, S., Hannon, G. J., Lawson, N. D., Wolfe, S. A. and Giraldez, A. J. (2010) A novel miRNA processing pathway independent of Dicer requires Argonaute2 catalytic activity. *Science,* 328**,** 1694-1698.

Claeskens, G. and Hjort, N. (2008) *Model selection and model averaging.,* Cambridge, Cambridge University Press.

Cohen, S. M., Brennecke, J. and Stark, A. (2006) Denoising feedback loops by thresholding — a new role for microRNAs. *Genes & Development,* 20**,** 2769-2772.

Comeron, J. M. (1995) A method for estimating the numbers of synonymous and nonsynonymous substitutions per site. *Journal of Molecular Evolution,* 41**,** 1152-1159.

Comeron, J. M. (1999) K-Estimator: calculation of the number of nucleotide substitutions per site and the confidence intervals. *Bioinformatics,* 15**,** 763-764.

Conant, G. C. and Wolfe, K. H. (2008) Turning a hobby into a job: how duplicated genes find new functions. *Nature Reviews Genetics,* 9**,** 938-950.

Coyne, J. A. and Orr, H. A. (2004) *Speciation,* Massachusetts, Sinauer Associates.

Crandall, K. A. and Hillis, D. M. (1997) Rhodopsin evolution in the dark. *Nature,* 387**,** 667-668.

Crespi, B. and Semeniuk, C. (2004) Parent‐offspring conflict in the evolution of vertebrate reproductive mode. *The American Naturalist,* 163**,** 635-653.

Cross, J. C., Anson-Cartwright, L. and Scott, I. C. (2002) Transcription factors underlying the development and endocrine functions of the placenta. *Recent Progress in Hormone Research,* 57**,** 221-234.

Curtis, S. E. and Clegg, M. T. (1984) Molecular evolution of chloroplast DNA sequences. *Molecular Biology and Evolution,* 1**,** 291-301.

Darriba, D., Taboada, G. L., Doallo, R. and Posada, D. (2011) ProtTest 3: fast selection of best-fit models of protein evolution. *Bioinformatics,* 27**,** 1164-1165.

Darwin, C. (1859) *On the origins of species by means of natural selection,* London, Murray.

Davidson, E. H. and Erwin, D. H. (2006) Gene regulatory networks and the evolution of animal body plans. *Science,* 311**,** 796-800.

Davis, E., Caiment, F., Tordoir, X., Cavaillé, J., Ferguson-Smith, A., Cockett, N., Georges, M. and Charlier, C. (2005) RNAi-mediated allelic trans-interaction at the imprinted Rtl1/Peg11 locus. *Current Biology,* 15**,** 743-749.

Dayhoff, M. O. and Schwartz, R. M. (1978) A model of evolutionary change in proteins. *Atlas of Protein Sequence and Structure*.

de Wit, E., Linsen, S. E. V., Cuppen, E. and Berezikov, E. (2009) Repertoire and evolution of miRNA genes in four divergent nematode species. *Genome Research,* 19**,** 2064-2074.

DeBry, R. W. and Marzluff, W. F. (1994) Selection on silent sites in the rodent H3 histone gene family. *Genetics,* 138**,** 191-202.

Degnan, J. H. and Rosenberg, N. A. (2009) Gene tree discordance, phylogenetic inference and the multispecies coalescent. *Trends in Ecology & Evolution***,** 1-9.

DeLong, E. R., DeLong, D. M. and Clarke-Pearson, D. L. (1988) Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics***,** 837-845.

Denli, A. M., Tops, B. B. J., Plasterk, R. H. A., Ketting, R. F. and Hannon, G. J. (2004) Processing of primary microRNAs by the Microprocessor complex. *Nature,* 432**,** 231-235.

Desper, R. and Gascuel, O. (2002) Fast and accurate phylogeny reconstruction algorithms based on the minimum-evolution principle. *Journal of Computational Biology,* 9**,** 687-705.

Dessimoz, C. and Gil, M. (2010) Phylogenetic assessment of alignments reveals neglected tree signal in gaps. *Genome Biology,* 11**,** R37.

Di Giammartino, D. C., Nishida, K. and Manley, J. L. (2011) Mechanisms and consequences of alternative polyadenylation. *Molecular Cell,* 43**,** 853-866.

Didiano, D. and Hobert, O. (2006) Perfect seed pairing is not a generally reliable predictor for miRNA-target interactions. *Nature Structural & Molecular Biology,* 13**,** 849-851.

Doench, J. G. and Sharp, P. A. (2004) Specificity of microRNA target selection in translational repression. *Genes & Development,* 18**,** 504-511.

Doolittle, W. F. (1999) Phylogenetic classification and the universal tree. *Science,* 284**,** 2124-2129.

Durrett, R. and Schmidt, D. (2007) Waiting for regulatory sequences to appear. *The Annals of Applied Probability,* 17**,** 1-32.

Duursma, A. M., Kedde, M., Schrier, M., le Sage, C. and Agami, R. (2008) miR-148 targets human DNMT3b protein coding region. *RNA,* 14**,** 872-877.

Dweep, H., Sticht, C., Pandey, P. and Gretz, N. (2011) miRWalk — database: prediction of possible miRNA binding sites by "walking" the genes of three genomes. *Journal of Biomedical Informatics,* 44**,** 839-847.

Ebert, M. S., Neilson, J. R. and Sharp, P. A. (2007) microRNA sponges: competitive inhibitors of small RNAs in mammalian cells. *Nature Methods,* 4**,** 721-726.

Ebert, M. S. and Sharp, P. A. (2010) Emerging roles for natural microRNA sponges. *Current Biology,* 20**,** R858-61.

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

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

Edwards, A. W. F. (1972) *Likelihood,* Cambridge Eng., University Press.

Egan, J. P. (1975) *Signal detection theory and ROC analysis,* New York, Academic Press.

Elcheva, I., Goswami, S., Noubissi, F. K. and Spiegelman, V. S. (2009) CRD-BP protects the coding region of βTrCP1 mRNA from miR-183-mediated degradation. *Molecular Cell,* 35**,** 240-246.

Elliot, M. G. and Crespi, B. J. (2009) Phylogenetic evidence for early hemochorial placentation in eutheria. *Placenta***,** 1-19.

Ellwanger, D. C., Buttner, F. A., Mewes, H. W. and Stumpflen, V. (2011) The sufficient minimal set of miRNA seed types. *Bioinformatics,* 27**,** 1346-1350.

Emery, J. F., Floyd, S. K., Alvarez, J., Eshed, Y., Hawker, N. P., Izhaki, A., Baum, S. F. and Bowman, J. L. (2003) Radial patterning of Arabidopsis shoots by class III HD-ZIP and KANADI genes. *Current Biology,* 13**,** 1768-1774.

Endo, T., Ikeo, K. and Gojobori, T. (1996) Large-scale search for genes on which positive selection may operate. *Molecular Biology and Evolution,* 13**,** 685-690.

Enright, A. J., John, B., Gaul, U., Tuschl, T., Sander, C. and Marks, D. S. (2003) MicroRNA targets in Drosophila. *Genome Biology,* 5**,** R1.

Eppig, J. T., Blake, J. A., Bult, C. J., Kadin, J. A., Richardson, J. E. and Group, M. G. D. (2012) The Mouse Genome Database (MGD): comprehensive resource for genetics and genomics of the laboratory mouse. *Nucleic Acids Research,* 40**,** D881-6.

Esau, C., Kang, X., Peralta, E., Hanson, E., Marcusson, E. G., Ravichandran, L. V., Sun, Y., Koo, S., Perera, R. J., Jain, R., Dean, N. M., Freier, S. M., Bennett, C. F., Lollo, B. and Griffey, R. (2004) MicroRNA-143 regulates adipocyte differentiation. *The Journal of Biological Chemistry,* 279**,** 52361-52365.

Eulenstein, O., Mirkin, B. and Vingron, M. (1998) Duplication-based measures of difference between gene and species trees. *Journal of Computational Biology,* 5**,** 135-48.

Eyre-Walker, A. and Keightley, P. D. (2007) The distribution of fitness effects of new mutations. *Nature Reviews Genetics,* 8**,** 610-618.

Fàbrega, C., Farrow, M. A., Mukhopadhyay, B., de Crécy-Lagard, V., Ortiz, A. R. and Schimmel, P. (2001) An aminoacyl tRNA synthetase whose sequence fits into neither of the two known classes. *Nature,* 411**,** 110-114.

Fang, Z. and Rajewsky, N. (2011) The impact of miRNA target sites in coding sequences and in 3′ UTRs. *PLoS ONE,* 6**,** e18067.

Fares, M. A. (2004) SWAPSC: sliding window analysis procedure to detect selective constraints. *Bioinformatics,* 20**,** 2867-2868.

Fares, M. A., Elena, S. F., Ortiz, J., Moya, A. and Barrio, E. (2002) A sliding window-based method to detect selective constraints in protein-coding genes and its application to RNA viruses. *Journal of Molecular Evolution,* 55**,** 509-521.

Farh, K. K.-H., Grimson, A., Jan, C., Lewis, B. P., Johnston, W. K., Lim, L. P., Burge, C. B. and Bartel, D. P. (2005) The widespread impact of mammalian MicroRNAs on mRNA repression and evolution. *Science,* 310**,** 1817-1821.

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

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

Felsenstein, J. (1996a) Inferring phylogenies from protein sequences by parsimony, distance, and likelihood methods. *Methods in Enzymology,* 266**,** 418-27.

Felsenstein, J. (1996b) *The Newick tree format.* [Online]. Available: http://evolution.genetics.washington.edu/phylip/newicktree.html [Accessed May 18th 2013 2013].

Felsenstein, J. (2004) *Inferring phylogenies,* Massachusetts, Sinauer Associates.

Ferguson-Smith, A. C. (2011) Genomic imprinting: the emergence of an epigenetic paradigm. *Nature Publishing Group,* 12**,** 565-575.

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

Fitch, W. M. (1971) Toward defining the course of evolution: minimum change for a specific tree topology. *Systematic Biology,* 20**,** 406-416.

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

Fitch, W. M. and Margoliash, E. (1967) Construction of phylogenetic trees. *Science,* 155**,** 279-284.

Flemming, A. F. and Blackburn, D. G. (2003) Evolution of placental specializations in viviparous African and South American lizards. *Journal of Experimental Zoology,* 299A**,** 33-47.

Fletcher, W. and Yang, Z. (2010) The effect of insertions, deletions, and alignment errors on the branch-site test of positive selection. *Molecular Biology and Evolution,* 27**,** 2257-2267.

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., Kähäri, 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. S., 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., Fernández-Suárez, X. M., Harrow, J., Herrero, J., Hubbard, T. J. P., Parker, A., Proctor, G., Spudich, G., Vogel, J., Yates, A., Zadissa, A. and Searle, S. M. J. (2012) Ensembl 2012. *Nucleic Acids Research,* 40**,** D84-90.

Flicek, P., Amode, M. R., Barrell, D., Beal, K., Brent, S., Chen, Y., Clapham, P., Coates, G., Fairley, S., Fitzgerald, S., Gordon, L., Hendrix, M., Hourlier, T., Johnson, N., Kähäri, A., Keefe, D., Keenan, S., Kinsella, R., Kokocinski, F., Kulesha, E., Larsson, P., Longden, I., McLaren, W., Overduin, B., Pritchard, B., Riat, H. S., Rios, D., Ritchie, G. R. S., Ruffier, M., Schuster, M., Sobral, D., Spudich, G., Tang, Y. A., Trevanion, S., Vandrovcova, J., Vilella, A. J., White, S., Wilder, S. P., Zadissa, A., Zamora, J., Aken, B. L., Birney, E., Cunningham, F., Dunham, I., Durbin, R., Fernández-Suárez, X. M., Herrero, J., Hubbard, T. J. P., Parker, A., Proctor, G., Vogel, J. and Searle, S. M. J. (2011) Ensembl 2011. *Nucleic Acids Research,* 39**,** D800-6.

Flor, I. and Bullerdiek, J. (2012) The dark side of a success story: microRNAs of the C19MC cluster in human tumours. *The Journal of Pathology,* 227**,** 270-274.

Forman, J. J., Legesse-Miller, A. and Coller, H. A. (2008) A search for conserved sequences in coding regions reveals that the let-7 microRNA targets Dicer within its coding sequence. *Proceedings of the National Academy of Sciences,* 105**,** 14879-14884.

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

Foster, P. G. (2004) Modeling compositional heterogeneity. *Systematic Biology,* 53**,** 485-495.

Foster, P. G. and Hickey, D. A. (1999) Compositional bias may affect both DNA-based and protein-based phylogenetic reconstructions. *Journal of Molecular Evolution,* 48**,** 284-290.

Franco-Zorrilla, J. M., Valli, A., Todesco, M., Mateos, I., Puga, M. I., Rubio-Somoza, I., Leyva, A., Weigel, D., García, J. A. and Paz-Ares, J. (2007) Target mimicry provides a new mechanism for regulation of microRNA activity. *Nature Genetics,* 39**,** 1033-1037.

Freyer, C. and Renfree, M. B. (2009) The mammalian yolk sac placenta. *Journal of Experimental Zoology B: Molecular and Developmental Evolution,* 312**,** 545-54.

Friedman, R. and Hughes, A. L. (2007) Likelihood-ratio tests for positive selection of human and mouse duplicate genes reveal nonconservative and anomalous properties of widely used methods. *Molecular Phylogenetics and Evolution,* 42**,** 388-393.

Friedman, R. C., Farh, K. K.-H., Burge, C. B. and Bartel, D. P. (2009) Most mammalian mRNAs are conserved targets of microRNAs. *Genome Research,* 19**,** 92-105.

Fritts, T. H., Leasman-Tanner, D. and Geological Survey (U.S.) (2001) The brown tree snake on Guam: how the arrival of one invasive species damaged the ecology, commerce, electrical systems, and human health on Guam, a comprehensive information source. 1.0. ed. Fort Collins, Colo.: U.S. Dept. of the Interior, U.S. Geological Survey.

Fuellen, G. (2008) Homology and phylogeny and their automated inference. *Naturwissenschaften,* 95**,** 469-481.

Gaidatzis, D., van Nimwegen, E., Hausser, J. and Zavolan, M. (2007) Inference of miRNA targets using evolutionary conservation and pathway analysis. *BMC Bioinformatics,* 8**,** 69.

Galtier, N. and Duret, L. (2007) Adaptation or biased gene conversion? Extending the null hypothesis of molecular evolution. *Trends in Genetics,* 23**,** 273-277.

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

Garcia, D. M., Baek, D., Shin, C., Bell, G. W., Grimson, A. and Bartel, D. P. (2011) Weak seed-pairing stability and high target-site abundance decrease the proficiency of lsy-6 and other microRNAs. *Nature Structural & Molecular Biology,* 18**,** 1139-1146.

Gascuel, O. (1997) BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. *Molecular Biology and Evolution,* 14**,** 685-695.

Gavel, Y. and von Heijne, G. (1990) Sequence differences between glycosylated and non-glycosylated Asn-X-Thr/Ser acceptor sites: implications for protein engineering. *Protein Engineering,* 3**,** 433-42.

Geng, Y., Yu, Q., Sicinska, E., Das, M., Schneider, J. E., Bhattacharya, S., Rideout, W. M., Bronson, R. T., Gardner, H. and Sicinski, P. (2003) Cyclin E ablation in the mouse. *Cell,* 114**,** 431-443.

Gerlt, J. A. and Babbitt, P. C. (2000) Can sequence determine function. *Genome Biology,* 1**,** 1-10.

Geyer, C. J. (1991) Markov chain Monte Carlo maximum likelihood. *In:* KERAMIDAS, E. M. (ed.) *Computing Science and Statistics: Proceedings of the 23rd Symposium on the Interface.* New York: Interface Foundation.

Girard, M., Jacquemin, E., Munnich, A., Lyonnet, S. and Henrion-Caude, A. (2008) miR-122, a paradigm for the role of microRNAs in the liver. *Journal of Hepatology,* 48**,** 648-656.

Gojobori, T., Li, W. H. and Graur, D. (1982) Patterns of nucleotide substitution in pseudogenes and functional genes. *Journal of Molecular Evolution,* 18**,** 360-369.

Goldman, N. (1993) Statistical tests of models of DNA substitution. *Journal of Molecular Evolution,* 36**,** 182-198.

Goldman, N. and Yang, Z. (1994) A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Molecular Biology and Evolution,* 11**,** 725-736.

Gopalakrishna, A. and Karim, K. B. (1979) Fetal membranes and placentation in Chiroptera. *Journal of Reproduction and Fertility,* 56**,** 417-429.

Gossmann, T. I., Keightley, P. D. and Eyre-Walker, A. (2012) The effect of variation in the effective population size on the rate of adaptive molecular evolution in eukaryotes. *Genome Biology and Evolution,* 4**,** 658-667.

Green, P. J. (1995) Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika,* 82**,** 711-732.

Gregory, R. I., Chendrimada, T. P., Cooch, N. and Shiekhattar, R. (2005) Human RISC couples microRNA biogenesis and posttranscriptional gene silencing. *Cell,* 123**,** 631-640.

Gregory, R. I., Yan, K.-p., Amuthan, G., Chendrimada, T., Doratotaj, B., Cooch, N. and Shiekhattar, R. (2004) The Microprocessor complex mediates the genesis of microRNAs. *Nature,* 432**,** 235-240.

Griffiths-Jones, S. (2004) The microRNA Registry. *Nucleic Acids Research,* 32**,** 109D-111.

Griffiths-Jones, S. (2006) miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Research,* 34**,** D140-D144.

Griffiths-Jones, S., Saini, H. K., van Dongen, S. and Enright, A. J. (2008) miRBase: tools for microRNA genomics. *Nucleic Acids Research,* 36**,** D154-8.

Grimson, A., Farh, K. K.-H., Johnston, W. K., Garrett-Engele, P., Lim, L. P. and Bartel, D. P. (2007) MicroRNA targeting specificity in mammals: determinants beyond seed pairing. *Molecular Cell,* 27**,** 91-105.

Grimson, A., Srivastava, M., Fahey, B., Woodcroft, B. J., Chiang, H. R., King, N., Degnan, B. M., Rokhsar, D. S. and Bartel, D. P. (2008) Early origins and evolution of microRNAs and Piwi-interacting RNAs in animals. *Nature,* 455**,** 1193-1197.

Gruber, A. R., Lorenz, R., Bernhart, S. H., Neubock, R. and Hofacker, I. L. (2008) The Vienna RNA Websuite. *Nucleic Acids Research,* 36**,** W70-W74.

Grün, D., Wang, Y. L., Langenberger, D., Gunsalus, K. C. and Rajewsky, N. (2005) microRNA target predictions across seven Drosophila species and comparison to mammalian targets. *PLoS Computational Biology,* 1**,** e13.

Gu, S., Jin, L., Zhang, F., Sarnow, P. and Kay, M. A. (2009) Biological basis for restriction of microRNA targets to the 3′ untranslated region in mammalian mRNAs. *Nature Structural & Molecular Biology,* 16**,** 144-150.

Guan, Y., Dunham, M. J. and Troyanskaya, O. G. (2007) Functional analysis of gene duplications in Saccharomyces cerevisiae. *Genetics,* 175**,** 933-943.

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

Ha, I., Wightman, B. and Ruvkun, G. (1996) A bulged lin-4/lin-14 RNA duplex is sufficient for Caenorhabditis elegans lin-14 temporal gradient formation. *Genes & Development,* 10**,** 3041-3050.

Hall, M. A. and Smith, L. A. (1997) Feature subset selection: a correlation based filter approach. *Proceedings of the 1997 International Conference on Neural Information Processing and Intelligent Information Systems.* New Zealand: Springer.

Hamlett, W. C. (2005) Ultrastructure of the maternal-fetal interface of the yolk sac placenta in sharks. *Italian Journal of Anatomy and Embryology,* 110**,** 175-81.

Hammell, M. (2010) Computational methods to identify miRNA targets. *Seminars in Cell and Developmental Biology,* 21**,** 738-744.

Hammell, M., Long, D., Zhang, L., Lee, A., Carmack, C. S., Han, M., Ding, Y. and Ambros, V. (2008) mirWIP: microRNA target prediction based on microRNA-containing ribonucleoprotein–enriched transcripts. *Nature Methods,* 5**,** 813-819.

Hammock, E. A. D. and Young, L. J. (2005) Microsatellite instability generates diversity in brain and sociobehavioral traits. *Science,* 308**,** 1630-1634.

Hamosh, A., Scott, A. F., Amberger, J. S., Bocchini, C. A. and McKusick, V. A. (2005) Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Research,* 33**,** D514-7.

Han, J. (2004) The Drosha-DGCR8 complex in primary microRNA processing. *Genes & Development,* 18**,** 3016-3027.

Hannan, E. and Quinn, B. (1979) The determination of the order of an autoregression. *Journal of the Royal Statistical Society B: Methodological,* 41**,** 190-195.

Hasegawa, M., Kishino, H. and Yano, T. (1985) Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *Journal of Molecular Evolution,* 22**,** 160-174.

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

Heimberg, A. M., Sempere, L. F., Moy, V. N., Donoghue, P. C. J. and Peterson, K. J. (2008) MicroRNAs and the advent of vertebrate morphological complexity. *Proceedings of the National Academy of Sciences,* 105**,** 2946-2950.

Hendrickson, D. G., Hogan, D. J., McCullough, H. L., Myers, J. W., Herschlag, D., Ferrell, J. E. and Brown, P. O. (2009) Concordant regulation of translation and mRNA abundance for hundreds of targets of a human microRNA. *PLoS Biology,* 7**,** e1000238.

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**,** 10915-10919.

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

Henson, M. C. and Castracane, V. D. (2006) Leptin in pregnancy: an update. *Biology of Reproduction,* 74**,** 218-229.

Hertel, J., Lindemeyer, M., Missal, K., Fried, C., Tanzer, A., Flamm, C., Hofacker, I. L., Stadler, P. F. and 2005, S. o. B. C. L. a. (2006) The expansion of the metazoan microRNA repertoire. *BMC Genomics,* 7**,** 25.

Hittinger, C. T. and Carroll, S. B. (2007) Gene duplication and the adaptive evolution of a classic genetic switch. *Nature,* 449**,** 677-681.

Hoel, P. G. (1962) Likelihood ratio tests. *Introduction to mathematical statistics.* 3d ed. New York,: Wiley.

Hofacker, I. L. (2003) Vienna RNA secondary structure server. *Nucleic Acids Research,* 31**,** 3429-3431.

Hofacker, I. L., Fontana, W., Stadler, P. F., Bonhoeffer, L. S., Tacker, M. and Schuster, P. (1994) Fast folding and comparison of RNA secondary structures. *Monatshefte für chemie,* 125**,** 167-188.

Hokamp, K., McLysaght, A. and Wolfe, K. H. (2003) The 2R hypothesis and the human genome sequence. *Journal of Structural and Functional Genomics,* 3**,** 95-110.

Holmes, I. and Rubin, G. M. (2002) An expectation maximization algorithm for training hidden substitution models. *Journal of Molecular Biology,* 317**,** 753-764.

Hon, L. S. and Zhang, Z. (2007) The roles of binding site arrangement and combinatorial targeting in microRNA repression of gene expression. *Genome Biology,* 8**,** R166.

Hornstein, E. and Shomron, N. (2006) Canalization of development by microRNAs. *Nature Genetics,* 38**,** S20-S24.

Hou, Z., Romero, R., Uddin, M., Than, N. G. and Wildman, D. E. (2009) Adaptive history of single copy genes highly expressed in the term human placenta. *Genomics,* 93**,** 33-41.

Hsu, S. D., Lin, F. M., Wu, W. Y., Liang, C., Huang, W. C., Chan, W. L., Tsai, W. T., Chen, G. Z., Lee, C. J., Chiu, C. M., Chien, C. H., Wu, M. C., Huang, C. Y., Tsou, A. P. and Huang, H. D. (2010) miRTarBase: a database curates experimentally validated microRNA-target interactions. *Nucleic Acids Research,* 39**,** D163-D169.

Hu, Y., Li, P., Hao, S., Liu, L., Zhao, J. and Hou, Y. (2009) Differential expression of microRNAs in the placentae of Chinese patients with severe pre-eclampsia. *Clinical Chemistry and Laboratory Medicine,* 47**,** 923-929.

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. *Proceedings of the National Academy of Sciences,* 109**,** 2966-2971.

Huang, Y., Zou, Q., Song, H., Song, F., Wang, L., Zhang, G. and Shen, X. (2010) A study of miRNAs targets prediction and experimental validation. *Protein & Cell,* 1**,** 979-986.

Hubisz, M. J., Pollard, K. S. and Siepel, A. (2011) PHAST and RPHAST: phylogenetic analysis with space/time models. *Briefings in Bioinformatics,* 12**,** 41-51.

Huelsenbeck, J. P. and Crandall, K. A. (1997) Phylogeny estimation and hypothesis testing using maximum likelihood. *Annual Review of Ecology and Systematics***,** 437-466.

Huelsenbeck, J. P. and Hillis, D. M. (1993) Success of phylogenetic methods in the four-taxon case. *Systematic Biology,* 42**,** 247-264.

Huelsenbeck, J. P., Larget, B., Miller, R. E. and Ronquist, F. (2002) Potential applications and pitfalls of Bayesian inference of phylogeny. *Systematic Biology,* 51**,** 673-688.

Huelsenbeck, J. P. and Ronquist, F. (2001) MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics,* 17**,** 754-755.

Huelsenbeck, J. P., Ronquist, F., Nielsen, R. and Bollback, J. P. (2001) Bayesian inference of phylogeny and its impact on evolutionary biology. *Science,* 294**,** 2310-2314.

Huerta-Cepas, J., Dopazo, J., Huynen, M. A. and Gabaldón, T. (2011) Evidence for short-time divergence and long-time conservation of tissue-specific expression after gene duplication. *Briefings in Bioinformatics,* 12**,** 442-448.

Hughes, A. L. (2007) Looking for Darwin in all the wrong places: the misguided quest for positive selection at the nucleotide sequence level. *Heredity,* 99**,** 364-373.

Hughes, A. L. and Friedman, R. (2005) Variation in the pattern of synonymous and nonsynonymous difference between two fungal genomes. *Molecular Biology and Evolution,* 22**,** 1320-1324.

Hughes, A. L. and Friedman, R. (2010) Myths and realities of gene duplication. *In:* DITTMAR, K. & LIBERLES, D. A. (eds.) *Evolution after gene duplication.* Hoboken, NJ: Wiley-Blackwell.

Hughes, A. L. and Nei, M. (1988) Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. *Nature,* 335**,** 167-70.

Huminiecki, L. and Heldin, C.-H. (2010) 2R and remodeling of vertebrate signal transduction engine. *BMC Biology,* 8**,** 146.

Huminiecki, L., Lloyd, A. T. and Wolfe, K. H. (2003) Congruence of tissue expression profiles from Gene Expression Atlas, SAGEmap and TissueInfo databases. *BMC Genomics,* 4**,** 31.

Hurst, L. D. (2009) Fundamental concepts in genetics: genetics and the understanding of selection. *Nature Publishing Group,* 10**,** 83-93.

Hurst, L. D. and Pál, C. (2001) Evidence for purifying selection acting on silent sites in BRCA1. *Trends in Genetics,* 17**,** 62-65.

Hutvagner, G., McLachlan, J., Pasquinelli, A. E., Bálint, E., Tuschl, T. and Zamore, P. D. (2001) A cellular function for the RNA-interference enzyme Dicer in the maturation of the let-7 small temporal RNA. *Science,* 293**,** 834-838.

Huxley, J. (1942) *Evolution: the modern synthesis,* London,, G. Allen & Unwin ltd.

Jackson, A. L. (2006) Position-specific chemical modification of siRNAs reduces "off-target" transcript silencing. *RNA,* 12**,** 1197-1205.

Jackson, A. L., Burchard, J., Schelter, J., Chau, B. N., Cleary, M., Lim, L. and Linsley, P. S. (2006) Widespread siRNA "off-target" transcript silencing mediated by seed region sequence complementarity. *RNA,* 12**,** 1179-1187.

Jensen, L. J., Julien, P., Kuhn, M., von Mering, C., Muller, J., Doerks, T. and Bork, P. (2008) eggNOG: automated construction and annotation of orthologous groups of genes. *Nucleic Acids Research,* 36**,** D250-4.

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

Ji, Q., Luo, Z.-X., Yuan, C.-X., Wible, J. R., Zhang, J.-P. and Georgi, J. A. (2002) The earliest known Eutherian mammal. *Nature,* 416**,** 816-822.

Jiang, Q., Wang, Y., Hao, Y., Juan, L., Teng, M., Zhang, X., Li, M., Wang, G. and Liu, Y. (2009) miR2Disease: a manually curated database for microRNA deregulation in human disease. *Nucleic Acids Research,* 37**,** D98-104.

Johannsen, W. (1911) The genotype conception of heredity. *American Naturalist***,** 129-159.

John, B., Enright, A. J., Aravin, A., Tuschl, T., Sander, C. and Marks, D. S. (2004) Human microRNA targets. *PLoS Biology,* 2**,** e363.

Johnson, J. M., Castle, J., Garrett-Engele, P., Kan, Z., Loerch, P. M., Armour, C. D., Santos, R., Schadt, E. E., Stoughton, R. and Shoemaker, D. D. (2003) Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays. *Science,* 302**,** 2141-2144.

Johnston, R. J. and Hobert, O. (2003) A microRNA controlling left/right neuronal asymmetry in Caenorhabditis elegans. *Nature,* 426**,** 845-9.

Jones, D. T., Taylor, W. R. and Thornton, J. M. (1992) The rapid generation of mutation data matrices from protein sequences. *Computer Applications in the Biosciences,* 8**,** 275-282.

Jukes, T. and Cantor, C. (1969) *Evolution of protein molecules,* New York, Academic Press.

Junier, T. and Zdobnov, E. M. (2010) The Newick utilities: high-throughput phylogenetic tree processing in the UNIX shell. *Bioinformatics,* 26**,** 1669-1670.

Kadota, K., Nishimura, S. I., Bono, H., Nakamura, S., Hayashizaki, Y., Okazaki, Y. and Takahashi, K. (2003) Detection of genes with tissue-specific expression patterns using Akaike's information criterion procedure. *Physiological Genomics,* 12**,** 251-259.

Kaessmann, H. (2010) Origins, evolution, and phenotypic impact of new genes. *Genome Research,* 20**,** 1313-1326.

Karres, J. S., Hilgers, V., Carrera, I., Treisman, J. and Cohen, S. M. (2007) The conserved microRNA miR-8 tunes atrophin levels to prevent neurodegeneration in Drosophila. *Cell,* 131**,** 136-145.

Kasahara, M. (2007) The 2R hypothesis: an update. *Current opinion in Immunology,* 19**,** 547-552.

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 Evolutionary Biology,* 6**,** 29.

Keeling, P. J. and Palmer, J. D. (2008) Horizontal gene transfer in eukaryotic evolution. *Nature Reviews Genetics,* 9**,** 605-618.

Kertesz, M., Iovino, N., Unnerstall, U., Gaul, U. and Segal, E. (2007) The role of site accessibility in microRNA target recognition. *Nature Genetics,* 39**,** 1278-1284.

Ketting, R. F. (2001) Dicer functions in RNA interference and in synthesis of small RNA involved in developmental timing in C. elegans. *Genes & Development,* 15**,** 2654-2659.

Khan, A. A., Betel, D., Miller, M. L., Sander, C., Leslie, C. S. and Marks, D. S. (2009) Transfection of small RNAs globally perturbs gene regulation by endogenous microRNAs. *Nature Biotechnology*.

Kheradpour, P., Stark, A., Roy, S. and Kellis, M. (2007) Reliable prediction of regulator targets using 12 Drosophila genomes. *Genome Research,* 17**,** 1919-1931.

Khvorova, A., Reynolds, A. and Jayasena, S. D. (2003) Functional siRNAs and miRNAs exhibit strand bias. *Cell,* 115**,** 209-216.

Kim, P., Yoon, S., Kim, N., Lee, S., Ko, M., Lee, H., Kang, H., Kim, J. and Lee, S. (2010) ChimerDB 2.0 — a knowledgebase for fusion genes updated. *Nucleic Acids Research,* 38**,** D81-5.

Kim, V. N. (2004) MicroRNA precursors in motion: exportin-5 mediates their nuclear export. *Trends in Cell Biology,* 14**,** 156-159.

Kim, V. N., Han, J. and Siomi, M. C. (2009) Biogenesis of small RNAs in animals. *Nature Reviews Molecular Cell Biology,* 10**,** 126-139.

Kimchi-Sarfaty, C., Oh, J. M., Kim, I.-W., Sauna, Z. E., Calcagno, A. M., Ambudkar, S. V. and Gottesman, M. M. (2007) A "silent" polymorphism in the MDR1 gene changes substrate specificity. *Science,* 315**,** 525-528.

Kimura, M. (1968) Evolutionary rate at the molecular level. *Nature,* 217**,** 624-626.

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

Kimura, M. (1981) Estimation of evolutionary distances between homologous nucleotide sequences. *Proceedings of the National Academy of Sciences of the United States of America,* 78**,** 454-458.

King, B. F. and Enders, A. C. (1993) Comparative development of the mammalian yolk sac. *In:* NOGALES, F. F. (ed.) *The human yolk sac and yolk sac tumors.* Berlin: Springer-Verlag.

Kiriakidou, M., Nelson, P. T., Kouranov, A., Fitziev, P., Bouyioukos, C., Mourelatos, Z. and Hatzigeorgiou, A. (2004) A combined computational-experimental approach predicts human microRNA targets. *Genes & Development,* 18**,** 1165-1178.

Kloosterman, W. P., Wienholds, E., Ketting, R. F. and Plasterk, R. H. A. (2004) Substrate requirements for let-7 function in the developing zebrafish embryo. *Nucleic Acids Research,* 32**,** 6284-6291.

Klosterman, P. S., Uzilov, A. V., Bendaña, Y. R., Bradley, R. K., Chao, S., Kosiol, C., Goldman, N. and Holmes, I. (2006) XRate: a fast prototyping, training and annotation tool for phylo-grammars. *BMC Bioinformatics,* 7**,** 428.

Knox, K. and Baker, J. C. (2008) Genomic evolution of the placenta using co-option and duplication and divergence. *Genome Research,* 18**,** 695-705.

Kononenko, I. (1994) Estimating attributes: analysis and extensions of RELIEF. *Machine Learning: ECML-94***,** 171-182.

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

Koonin, E. V., Makarova, K. S. and Aravind, L. (2001) Horizontal gene transfer in prokaryotes: quantification and classification. *Annual Review of Microbiology,* 55**,** 709-742.

Kosiol, C., Vinař, 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 Genetics,* 4**,** e1000144.

Kotzot, D. (2004) Maternal uniparental disomy 14 dissection of the phenotype with respect to rare autosomal recessively inherited traits, trisomy mosaicism, and genomic imprinting. *Annales de génétique,* 47**,** 251-260.

Kozomara, A. and Griffiths-Jones, S. (2011) miRBase: integrating microRNA annotation and deep-sequencing data. *Nucleic Acids Research,* 39**,** D152-7.

Krause, A. and Vingron, M. (1998) A set-theoretic approach to database searching and clustering. *Bioinformatics,* 14**,** 430-438.

Krek, A., Grün, D., Poy, M. N., Wolf, R., Rosenberg, L., Epstein, E. J., MacMenamin, P., da Piedade, I., Gunsalus, K. C., Stoffel, M. and Rajewsky, N. (2005) Combinatorial microRNA target predictions. *Nature Genetics,* 37**,** 495-500.

Krüger, J. and Rehmsmeier, M. (2006) RNAhybrid: microRNA target prediction easy, fast and flexible. *Nucleic Acids Research,* 34**,** W451-4.

Krützfeldt, J., Rajewsky, N., Braich, R., Rajeev, K. G., Tuschl, T., Manoharan, M. and Stoffel, M. (2005) Silencing of microRNAs in vivo with 'antagomirs'. *Nature,* 438**,** 685-689.

Kuchenbauer, F., Mah, S. M., Heuser, M., McPherson, A., Rüschmann, J., Rouhi, A., Berg, T., Bullinger, L., Argiropoulos, B., Morin, R. D., Lai, D., Starczynowski, D. T., Karsan, A., Eaves, C. J., Watahiki, A., Wang, Y., Aparicio, S. A., Ganser, A., Krauter, J., Döhner, H., Döhner, K., Marra, M. A., Camargo, F. D., Palmqvist, L., Buske, C. and Humphries, R. K. (2011) Comprehensive analysis of mammalian miRNA* species and their role in myeloid cells. *Blood,* 118**,** 3350-3358.

Kumar, S. and Filipski, A. (2007) Multiple sequence alignment: In pursuit of homologous DNA positions. *Genome Research,* 17**,** 127-135.

Kuzniar, A., van Ham, R. C. H. J., Pongor, S. and Leunissen, J. A. M. (2008) The quest for orthologs: finding the corresponding gene across genomes. *Trends in Genetics,* 24**,** 539-551.

Lagos-Quintana, M. (2003) New microRNAs from mouse and human. *RNA,* 9**,** 175-179.

Lagos-Quintana, M., Rauhut, R., Lendeckel, W. and Tuschl, T. (2001) Identification of novel genes coding for small expressed RNAs. *Science,* 294**,** 853-858.

Lal, A., Navarro, F., Maher, C. A., Maliszewski, L. E., Yan, N., O'Day, E., Chowdhury, D., Dykxhoorn, D. M., Tsai, P., Hofmann, O., Becker, K. G., Gorospe, M., Hide, W. and Lieberman, J. (2009) miR-24 inhibits cell proliferation by targeting E2F2, MYC, and other cell-cycle genes via binding to "seedless" 3′UTR microRNA recognition elements. *Molecular Cell,* 35**,** 610-625.

Lall, S., Grün, D., Krek, A., Chen, K., Wang, Y.-L., Dewey, C. N., Sood, P., Colombo, T., Bray, N., MacMenamin, P., Kao, H.-L., Gunsalus, K. C., Pachter, L., Piano, F. and Rajewsky, N. (2006) A genome-wide map of conserved microRNA targets in C. elegans. *Current Biology,* 16**,** 460-471.

Lanave, C., Preparata, G., Saccone, C. and Serio, G. (1984) A new method for calculating evolutionary substitution rates. *Journal of Molecular Evolution,* 20**,** 86-93.

Landan, G. and Graur, D. (2007) Heads or tails: a simple reliability check for multiple sequence alignments. *Molecular Biology and Evolution,* 24**,** 1380-1383.

Lander, E. S. and Waterman, M. S. (1988) Genomic mapping by fingerprinting random clones: a mathematical analysis. *Genomics,* 2**,** 231-239.

Landgraf, P., Rusu, M., Sheridan, R., Sewer, A., Iovino, N., Aravin, A., Pfeffer, S., Rice, A., Kamphorst, A. O., Landthaler, M., Lin, C., Socci, N. D., Hermida, L., Fulci, V., Chiaretti, S., Foà, R., Schliwka, J., Fuchs, U., Novosel, A., Müller, R.-U., Schermer, B., Bissels, U., Inman, J., Phan, Q., Chien, M., Weir, D. B., Choksi, R., De Vita, G., Frezzetti, D., Trompeter, H.-I., Hornung, V., Teng, G., Hartmann, G., Palkovits, M., Di Lauro, R., Wernet, P., Macino, G., Rogler, C. E., Nagle, J. W., Ju, J., Papavasiliou, F. N., Benzing, T., Lichter, P., Tam, W., Brownstein, M. J., Bosio, A., Borkhardt, A., Russo, J. J., Sander, C., Zavolan, M. and Tuschl, T. (2007) A mammalian microRNA expression atlas based on small RNA library sequencing. *Cell,* 129**,** 1401-1414.

Landthaler, M., Yalcin, A. and Tuschl, T. (2004) The human DiGeorge syndrome critical region gene 8 and Its D. melanogaster homolog are required for miRNA biogenesis. *Current Biology,* 14**,** 2162-2167.

Larget, B. and Simon, D. L. (1999) Markov chain Monte Carlo algorithms for the Bayesian analysis of phylogenetic trees. *Molecular Biology and Evolution,* 16, 750-759.

Larkin, M. A., Blackshields, G., Brown, N. P., Chenna, R., McGettigan, P. A., McWilliam, H., Valentin, F., Wallace, I. M., Wilm, A., Lopez, R., Thompson, J. D., Gibson, T. J. and Higgins, D. G. (2007) Clustal W and Clustal X version 2.0. *Bioinformatics,* 23, 2947-2948.

Lartillot, N. (2004) A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Molecular Biology and Evolution,* 21, 1095-1109.

Lartillot, N. (2012) Interaction between selection and biased gene conversion in mammalian protein-coding sequence evolution revealed by a phylogenetic covariance analysis. *Molecular Biology and Evolution,* 30, 356-368.

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

Le, S. Q. and Gascuel, O. (2008) An improved general amino acid replacement matrix. *Molecular Biology and Evolution,* 25, 1307-1320.

Lee, I., Ajay, S. S., Yook, J. I., Kim, H. S., Hong, S. H., Kim, N. H., Dhanasekaran, S. M., Chinnaiyan, A. M. and Athey, B. D. (2009) New class of microRNA targets containing simultaneous 5′-UTR and 3′-UTR interaction sites. *Genome Research,* 19, 1175-1183.

Lee, R. C., Feinbaum, R. L. and Ambros, V. (1993) The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell,* 75, 843-854.

Lee, T. I. and Young, R. A. (2000) Transcription of eukaryotic protein-coding genes. *Annual Review of Genetics,* 34, 77-137.

Lee, Y., Ahn, C., Han, J., Choi, H., Kim, J., Yim, J., Lee, J., Provost, P., Rådmark, O., Kim, S. and Kim, V. N. (2003) The nuclear RNase III Drosha initiates microRNA processing. *Nature,* 425, 415-419.

Lee, Y., Kim, M., Han, J., Yeom, K.-H., Lee, S., Baek, S. H. and Kim, V. N. (2004) MicroRNA genes are transcribed by RNA polymerase II. *The EMBO Journal,* 23, 4051-4060.

Lehnert, S., Van Loo, P., Thilakarathne, P. J., Marynen, P., Verbeke, G. and Schuit, F. C. (2009) Evidence for co-evolution between human microRNAs and Alu-repeats. *PLoS ONE,* 4**,** e4456.

Levasseur, A., Gouret, P., Lesage-Meessen, L., Asther, M., 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 Evolutionary Biology,* 6**,** 92.

Levine, M. and Tjian, R. (2003) Transcription regulation and animal diversity. *Nature,* 424**,** 147-151.

Lewin, B. (2008) *Genes IX,* Sudbury, Mass., Jones and Bartlett Publishers.

Lewis, B. P., Burge, C. B. and Bartel, D. P. (2005) Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell,* 120**,** 15-20.

Lewis, B. P., Shih, I.-h., Jones-Rhoades, M. W., Bartel, D. P. and Burge, C. B. (2003) Prediction of mammalian microRNA targets. *Cell,* 115**,** 787-798.

Li, H. (2007) TreeBeST.

Li, L., Stoeckert, C. J. and Roos, D. S. (2003) OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Research,* 13**,** 2178-2189.

Li, M., Yee, D., Magnuson, T. R., Smithies, O. and Caron, K. M. (2006) Reduced maternal expression of adrenomedullin disrupts fertility, placentation, and fetal growth in mice. *Journal of Clinical Investigation,* 116**,** 2653-2662.

Li, W. H. (1993) Unbiased estimation of the rates of synonymous and nonsynonymous substitution. *Journal of Molecular Evolution,* 36**,** 96-99.

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. *Molecular Biology and Evolution,* 2**,** 150-174.

Liang, Y., Ridzon, D., Wong, L. and Chen, C. (2007) Characterization of microRNA expression profiles in normal human tissues. *BMC Genomics,* 8**,** 166.

Lim, L. P., Lau, N. C., Garrett-Engele, P., Grimson, A., Schelter, J. M., Castle, J., Bartel, D. P., Linsley, P. S. and Johnson, J. M. (2005) Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs. *Nature,* 433**,** 769-773.

Lin, M. F., Kheradpour, P., Washietl, S., Parker, B. J., Pedersen, J. S. and Kellis, M. (2011) Locating protein-coding sequences under selection for additional, overlapping functions in 29 mammalian genomes. *Genome Research,* 21**,** 1916-1928.

Lin, S.-L., Miller, J. D. and Ying, S.-Y. (2006) Intronic microRNA (miRNA). *Journal of Biomedicine and Biotechnology,* 2006**,** 26818.

Linsley, P. S., Schelter, J., Burchard, J., Kibukawa, M., Martin, M. M., Bartz, S. R., Johnson, J. M., Cummins, J. M., Raymond, C. K., Dai, H., Chau, N., Cleary, M., Jackson, A. L., Carleton, M. and Lim, L. (2007) Transcripts targeted by the microRNA-16 family cooperatively regulate cell cycle progression. *Molecular and Cellular Biology,* 27**,** 2240-2252.

Liu, H., Yue, D., Chen, Y., Gao, S.-J. and Huang, Y. (2010) Improving performance of mammalian microRNA target prediction. *BMC Bioinformatics,* 11**,** 476.

Liu, J., Carmell, M. A., Rivas, F. V., Marsden, C. G., Thomson, J. M., Song, J.-J., Hammond, S. M., Joshua-Tor, L. and Hannon, G. J. (2004) Argonaute2 is the catalytic engine of mammalian RNAi. *Science,* 305**,** 1437-1441.

Lockhart, P., Novis, P., Milligan, B. G., Riden, J., Rambaut, A. and Larkum, T. (2006) Heterotachy and tree building: a case study with plastids and eubacteria. *Molecular Biology and Evolution,* 23**,** 40-45.

Long, D., Lee, R., Williams, P., Chan, C. Y., Ambros, V. and Ding, Y. (2007) Potent effect of target structure on microRNA function. *Nature Publishing Group,* 14**,** 287-294.

Long, M. and Langley, C. H. (1993) Natural selection and the origin of jingwei, a chimeric processed functional gene in Drosophila. *Science,* 260**,** 91-95.

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. *Molecular Biology and Evolution,* 29**,** 2039-2046.

Löytynoja, A. and Goldman, N. (2008) Phylogeny-aware gap placement prevents errors in sequence alignment and evolutionary analysis. *Science,* 320**,** 1632-1635.

Lu, J., Shen, Y., Wu, Q., Kumar, S., He, B., Shi, S., Carthew, R. W., Wang, S. M. and Wu, C.-I. (2008) The birth and death of microRNA genes in Drosophila. *Nature Genetics,* 40**,** 351-355.

Luo, L., Ye, G., Nadeem, L., Fu, G., Yang, B. B., Honarparvar, E., Dunk, C., Lye, S. and Peng, C. (2012) MicroRNA-378a-5p promotes trophoblast cell survival, migration and invasion by targeting Nodal. *Journal of Cell Science,* 125**,** 3124-3132.

Luo, S. S., Ishibashi, O., Ishikawa, G., Ishikawa, T., Katayama, A., Mishima, T., Takizawa, T., Shigihara, T., Goto, T. and Izumi, A. (2009) Human villous trophoblasts express and secrete placenta-specific microRNAs into maternal circulation via exosomes. *Biology of Reproduction,* 81**,** 717-729.

Lynch, M. and Conery, J. S. (2000) The evolutionary fate and consequences of duplicate genes. *Science,* 290**,** 1151-1155.

Lytle, J. R., Yario, T. A. and Steitz, J. A. (2007) Target mRNAs are repressed as efficiently by microRNA-binding sites in the 5′ UTR as in the 3′ UTR. *Proceedings of the National Academy of Sciences of the United States of America,* 104**,** 9667-9672.

MacArthur, S. and Brookfield, J. F. Y. (2004) Expected rates and modes of evolution of enhancer sequences. *Molecular Biology and Evolution,* 21**,** 1064-1073.

Maddison, W. P. (1997) Gene trees in species trees. *Systematic Biology,* 46**,** 523-536.

Maere, S., De Bodt, S., Raes, J., Casneuf, T., Van Montagu, M., Kuiper, M. and Van de Peer, Y. (2005) Modeling gene and genome duplications in eukaryotes. *Proceedings of the National Academy of Sciences of the United States of America,* 102**,** 5454-5459.

Majoros, W. H. and Ohler, U. (2007) Spatial preferences of microRNA targets in 3′ untranslated regions. *BMC Genomics,* 8**,** 152.

Mallet, J. (2005) Hybridization as an invasion of the genome. *Trends in Ecology & Evolution,* 20**,** 229-237.

Maragkakis, M., Alexiou, P., Papadopoulos, G. L., Reczko, M., Dalamagas, T., Giannopoulos, G., Goumas, G., Koukis, E., Kourtis, K., Simossis, V. A., Sethupathy, P., Vergoulis, T., Koziris, N., Sellis, T., Tsanakas, P. and Hatzigeorgiou, A. G. (2009a) Accurate microRNA target prediction correlates with protein repression levels. *BMC Bioinformatics,* 10**,** 295.

Maragkakis, M., Reczko, M., Simossis, V. A., Alexiou, P., Papadopoulos, G. L., Dalamagas, T., Giannopoulos, G., Goumas, G., Koukis, E., Kourtis, K., Vergoulis, T., Koziris, N., Sellis, T., Tsanakas, P. and Hatzigeorgiou, A. G. (2009b) DIANA-microT web server: elucidating microRNA functions through target prediction. *Nucleic Acids Research,* 37**,** W273-6.

Marco, A., Hui, J. H. L., Ronshaugen, M. and Griffiths-Jones, S. (2010) Functional shifts in insect microRNA evolution. *Genome Biology and Evolution*.

Marín, R. M., Sulc, M. and Vaníček, J. (2013) Searching the coding region for microRNA targets. *RNA,* 19**,** 467-474.

Marín, R. M. and Vaníček, J. (2011) Efficient use of accessibility in microRNA target prediction. *Nucleic Acids Research,* 39**,** 19-29.

Marín, R. M. and Vaníček, J. (2012) Optimal use of conservation and accessibility filters in microRNA target prediction. *PLoS ONE,* 7**,** e32208.

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 Research,* 21**,** 863-874.

Martin, A. P. and Burg, T. M. (2002) Perils of paralogy: using HSP70 genes for inferring organismal phylogenies. *Systematic Biology,* 51**,** 570-587.

Martin, R. D. (2008) Evolution of placentation in primates: implications of mammalian phylogeny. *Evolutionary Biology,* 35**,** 125-145.

Martinez, N. J., Ow, M. C., Reece-Hoyes, J. S., Barrasa, M. I., Ambros, V. R. and Walhout, A. J. M. (2008) Genome-scale spatiotemporal analysis of Caenorhabditis elegans microRNA promoter activity. *Genome Research,* 18**,** 2005-2015.

Mathews, D. H., Disney, M. D., Childs, J. L., Schroeder, S. J., Zuker, M. and Turner, D. H. (2004) Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure. *Proceedings of the National Academy of Sciences of the United States of America,* 101**,** 7287-7292.

Mau, B., Newton, M. A. and Larget, B. (1999) Bayesian phylogenetic inference via Markov chain Monte Carlo methods. *Biometrics,* 55**,** 1-12.

Maymó, J. L., Pérez, A. P., Gambino, Y., Calvo, J. C., Sánchez-Margalet, V. and Varone, C. L. (2011) Leptin gene expression in the placenta — regulation of a key hormone in trophoblast proliferation and survival. *Placenta,* 32 Suppl 2**,** S146-53.

Mayor-Lynn, K., Toloubeydokhti, T., Cruz, A. C. and Chegini, N. (2011) Expression profile of microRNAs and mRNAs in human placentas from pregnancies complicated by preeclampsia and preterm labor. *Reproductive Sciences,* 18**,** 46-56.

Mayr, E. (1970) *Populations, species, and evolution,* Cambridge, Mass.,, Belknap Press of Harvard University Press.

McCaskill, J. S. (1990) The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers,* 29**,** 1105-1119.

McGhee, G. R. (2011) *Convergent evolution: limited forms most beautiful,* Cambridge, Mass., MIT Press.

McNamara, J. (2012) Spreadsheet::WriteExcel. 2.37 ed.: CPAN.

Meister, G., Landthaler, M., Patkaniowska, A., Dorsett, Y., Teng, G. and Tuschl, T. (2004) Human Argonaute2 mediates RNA cleavage targeted by miRNAs and siRNAs. *Molecular Cell,* 15**,** 185-197.

Mess, A. and Carter, A. M. (2007) Evolution of the placenta during the early radiation of placental mammals. *Comparative Biochemistry and Physiology A: Molecular & Integrative Physiology,* 148**,** 769-779.

Messier, W. and Stewart, C. B. (1997) Episodic adaptive evolution of primate lysozymes. *Nature,* 385**,** 151-154.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. and Teller, E. (1953) Equation of state calculations by fast computing machines. *The Journal of Chemical Physics,* 21**,** 1087.

Metzger, K. J. and Thomas, M. A. (2010) Evidence of positive selection at codon sites localized in extracellular domains of mammalian CC motif chemokine receptor proteins. *BMC Evolutionary Biology,* 10**,** 139.

Millar, A. A. and Waterhouse, P. M. (2005) Plant and animal microRNAs: similarities and differences. *Functional & Integrative Genomics,* 5**,** 129-135.

Miller, W., Rosenbloom, K., Hardison, R. C., Hou, M., Taylor, J., Raney, B., Burhans, R., King, D. C., Baertsch, R., Blankenberg, D., Kosakovsky Pond, S. L., Nekrutenko, A., Giardine, B., Harris, R. S., Tyekucheva, S., Diekhans, M., Pringle, T. H., Murphy, W. J., Lesk, A., Weinstock, G. M., Lindblad-Toh, K., Gibbs, R. A., Lander, E. S., Siepel, A., Haussler, D. and Kent, W. J. (2007) 28-way vertebrate alignment and conservation track in the UCSC Genome Browser. *Genome Research,* 17**,** 1797-1808.

Miranda, K. C., Huynh, T., Tay, Y., Ang, Y. S., Tam, W. L., Thomson, A. M., Lim, B. and Rigoutsos, I. (2006) A pattern-based method for the identification of MicroRNA binding sites and their corresponding heteroduplexes. *Cell,* 126**,** 1203-1217.

Mirkin, B., Muchnik, I. and Smith, T. F. (1995) A biologically consistent model for comparing molecular phylogenies. *Journal of Computational Biology,* 2**,** 493-507.

Mitra, R. and Bandyopadhyay, S. (2011) MultiMiTar: a novel multi objective optimization based miRNA-target prediction method. *PLoS ONE,* 6**,** e24583.

Miyata, T. and Yasunaga, T. (1980) Molecular evolution of mRNA: a method for estimating evolutionary rates of synonymous and amino acid substitutions from homologous nucleotide sequences and its application. *Journal of Molecular Evolution,* 16**,** 23-36.

Morgan, C. (2012) *The molecular phylogeny of placental mammals and its application to uncovering signatures of molecular adaptation.* PhD Thesis, Dublin City University: School of Biotechnology.

Morozova, N., Zinovyev, A., Nonne, N., Pritchard, L.-L., Gorban, A. N. and Harel-Bellan, A. (2012) Kinetic signatures of microRNA modes of action. *RNA,* 18**,** 1635-1655.

Moustafa, A. (2005) JAligner. 1.0 ed.

Mukherji, S., Ebert, M. S., Zheng, G. X. Y., Tsang, J. S., Sharp, P. A. and van Oudenaarden, A. (2011) MicroRNAs can generate thresholds in target gene expression. *Nature Publishing Group,* 43**,** 854-859.

Muller, J., Szklarczyk, D., Julien, P., Letunic, I., Roth, A., Kuhn, M., Powell, S., von Mering, C., Doerks, T., Jensen, L. J. and Bork, P. (2009) eggNOG v2.0: extending the evolutionary genealogy of genes with enhanced non-supervised orthologous groups, species and functional annotations. *Nucleic Acids Research,* 38**,** D190-D195.

Müller, T. and Vingron, M. (2000) Modeling amino acid replacement. *Journal of Computational Biology,* 7**,** 761-776.

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. (2001) Resolution of the early placental mammal radiation using Bayesian phylogenetics. *Science,* 294**,** 2348-2351.

Needleman, S. B. and Wunsch, C. D. (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology,* 48**,** 443-453.

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 Computational Biology,* 7**,** e1002073.

Nei, M. and Gojobori, T. (1986) Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Molecular Biology and Evolution,* 3**,** 418-426.

Nielsen, C. B., Shomron, N., Sandberg, R., Hornstein, E., Kitzman, J. and Burge, C. B. (2007) Determinants of targeting by endogenous and exogenous microRNAs and siRNAs. *RNA,* 13**,** 1894-1910.

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**,** 929-936.

Noguer-Dance, M., Abu-Amero, S., Al-Khtib, M., Lefevre, A., Coullin, P., Moore, G. E. and Cavaille, J. (2010) The primate-specific microRNA gene cluster (C19MC) is imprinted in the placenta. *Human Molecular Genetics,* 19**,** 3566-3582.

Nozawa, M., Suzuki, Y. and Nei, M. (2009) Reliabilities of identifying positive selection by the branch-site and the site-prediction methods. *Proceedings of the National Academy of Sciences,* 106**,** 6700-6705.

Nuel, G., Regad, L., Martin, J. and Camproux, A.-C. (2010) Exact distribution of a pattern in a set of random sequences generated by a Markov source: applications to biological data. *Algorithms for Molecular Biology,* 5**,** 15.

Ohno, S. (1970) *Evolution by gene duplication,* New York, Springer-Verlag.

Ohno, S. (1984) Birth of a unique enzyme from an alternative reading frame of the preexisted, internally repetitious coding sequence. *Proceedings of the National Academy of Sciences of the United States of America,* 81**,** 2421-2425.

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

Ohta, T. (1992) The nearly neutral theory of molecular evolution. *Annual Review of Ecology and Systematics***,** 263-286.

Okamura, K., Hagen, J. W., Duan, H., Tyler, D. M. and Lai, E. C. (2007) The mirtron pathway generates microRNA-class regulatory RNAs in Drosophila. *Cell,* 130**,** 89-100.

Okamura, K., Phillips, M. D., Tyler, D. M., Duan, H., Chou, Y.-t. and Lai, E. C. (2008) The regulatory activity of microRNA* species has substantial influence on microRNA and 3′ UTR evolution. *Nature Structural & Molecular Biology,* 15**,** 354-363.

Ong, S.-E., Blagoev, B., Kratchmarova, I., Kristensen, D. B., Steen, H., Pandey, A. and Mann, M. (2002) Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. *Molecular & Cellular Proteomics,* 1**,** 376-386.

Ørom, U. A., Nielsen, F. C. and Lund, A. H. (2008) MicroRNA-10a binds the 5′UTR of ribosomal protein mRNAs and enhances their translation. *Molecular Cell,* 30**,** 460-471.

Ostlund, G., Schmitt, T., Forslund, K., Kostler, T., Messina, D. N., Roopra, S., Frings, O. and Sonnhammer, E. L. L. (2009) InParanoid 7: new algorithms and tools for eukaryotic orthology analysis. *Nucleic Acids Research,* 38**,** D196-D203.

Overbeek, R., Fonstein, M., D'Souza, M., Pusch, G. D. and Maltsev, N. (1999) The use of gene clusters to infer functional coupling. *Proceedings of the National Academy of Sciences of the United States of America,* 96**,** 2896-2901.

Owen, R. (1843) Lectures on the comparative anatomy and physiology of the invertebrate animals.

Ozsolak, F., Poling, L. L., Wang, Z., Liu, H., Liu, X. S., Roeder, R. G., Zhang, X., Song, J. S. and Fisher, D. E. (2008) Chromatin structure analyses identify miRNA promoters. *Genes & Development,* 22**,** 3172-3183.

Page, R. D. and Charleston, M. A. (1997) From gene to organismal phylogeny: reconciled trees and the gene tree/species tree problem. *Molecular Phylogenetics and Evolution,* 7**,** 231-40.

Papadopoulos, G. L., Reczko, M., Simossis, V. A., Sethupathy, P. and Hatzigeorgiou, A. G. (2009) The database of experimentally supported targets: a functional update of TarBase. *Nucleic Acids Research,* 37**,** D155-8.

Pasquinelli, A. E. (2012) MicroRNAs and their targets: recognition, regulation and an emerging reciprocal relationship. *Nature Publishing Group,* 13**,** 271-282.

Pasquinelli, A. E., Reinhart, B. J., Slack, F., Martindale, M. Q., Kuroda, M. I., Maller, B., Hayward, D. C., Ball, E. E., Degnan, B., Muller, P., Spring, J., Srinivasan, A., Fishman, M., Finnerty, J., Corbo, J., Levine, M., Leahy, P., Davidson, E. and Ruvkun, G. (2000) Conservation of the sequence and temporal expression of let-7 heterochronic regulatory RNA. *Nature,* 408**,** 86-89.

Patthy, L. (2008) *Protein evolution,* Malden, MA ; Oxford, Blackwell Science.

Peter, I. S. and Davidson, E. H. (2011) Evolution of gene regulatory networks controlling body plan development. *Cell,* 144**,** 970-985.

Peter, M. E. (2010) Targeting of mRNAs by multiple miRNAs: the next step. *Oncogene,* 29**,** 2161-2164.

Peterson, K. J., Dietrich, M. R. and McPeek, M. A. (2009a) MicroRNAs and metazoan macroevolution: insights into canalization, complexity, and the Cambrian explosion. *BioEssays,* 31**,** 736-747.

Peterson, M. E., Chen, F., Saven, J. G., Roos, D. S., Babbitt, P. C. and Sali, A. (2009b) Evolutionary constraints on structural similarity in orthologs and paralogs. *Protein Science,* 18**,** 1306-1315.

Pevsner, J. (2003) *Bioinformatics and functional genomics,* Hoboken, N.J., Wiley-Liss, Inc.

Philippe, H., Brinkmann, H., Lavrov, D. V., Littlewood, D. T. J., Manuel, M., WOrheide, G. and Baurain, D. (2011) Resolving difficult phylogenetic questions: why more sequences are not enough. *PLoS Biology,* 9**,** e1000602.

Pineles, B. L., Romero, R., Montenegro, D., Tarca, A. L., Han, Y. M., Kim, Y. M., Draghici, S., Espinoza, J., Kusanovic, J. P., Mittal, P., Hassan, S. S. and Kim, C. J. (2007) Distinct subsets of microRNAs are expressed differentially in the human placentas of patients with pre-eclampsia. *American Journal of Obstetrics and Gynecology,* 196**,** 261.e1-6.

Podlaha, O., Webb, D. M., Tucker, P. K. and Zhang, J. (2005) Positive selection for indel substitutions in the rodent sperm protein catsper1. *Molecular Biology and Evolution,* 22**,** 1845-1852.

Poliseno, L., Tuccoli, A., Mariani, L., Evangelista, M., Citti, L., Woods, K., Mercatanti, A., Hammond, S. and Rainaldi, G. (2006) microRNAs modulate the angiogenic properties of HUVECs. *Blood,* 108**,** 3068-3071.

Pollard, K. S., Hubisz, M. J., Rosenbloom, K. R. and Siepel, A. (2010) Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Research,* 20**,** 110-121.

Pond, S. L. K., Frost, S. D. W. and Muse, S. V. (2005) HyPhy: hypothesis testing using phylogenies. *Bioinformatics,* 21**,** 676-679.

Posada, D. and Buckley, T. (2004) Model selection and model averaging in phylogenetics: advantages of Akaike information criterion and Bayesian approaches over likelihood ratio tests. *Systematic Biology,* 53**,** 793-808.

Posada, D. and Crandall, K. A. (2001) Selecting the best-fit model of nucleotide substitution. *Systematic Biology,* 50**,** 580-601.

Poulton, E. B. (1903) What is a species? *Proceedings of the Entomological Society of London***,** lxxvii–cxvi.

Poy, M. N., Eliasson, L., Krützfeldt, J., Kuwajima, S., Ma, X., Macdonald, P. E., Pfeffer, S., Tuschl, T., Rajewsky, N., Rorsman, P. and Stoffel, M. (2004) A pancreatic islet-specific microRNA regulates insulin secretion. *Nature,* 432**,** 226-230.

Press, W. H. (2007) *Numerical recipes: the art of scientific computing,* Cambridge, UK ; New York, Cambridge University Press.

Prieto, D. M. M. and Markert, U. R. (2011) MicroRNAs in pregnancy. *Journal of Reproductive Immunology,* 88**,** 106-111.

Prochnik, S. E., Rokhsar, D. S. and Aboobaker, A. A. (2006) Evidence for a microRNA expansion in the Bilaterian ancestor. *Development Genes and Evolution,* 217**,** 73-77.

Qian, W. and Zhang, J. (2009) Protein subcellular relocalization in the evolution of yeast singleton and duplicate genes. *Genome Biology and Evolution,* 1**,** 198-204.

Ragan, C., Zuker, M. and Ragan, M. A. (2011) Quantitative prediction of miRNA-mRNA interaction based on equilibrium concentrations. *PLoS Computational Biology,* 7**,** e1001090.

Rajewsky, N. (2006) microRNA target predictions in animals. *Nature Genetics,* 38**,** S8-S13.

Rannala, B. and Yang, Z. (1996) Probability distribution of molecular evolutionary trees: a new method of phylogenetic inference. *Journal of Molecular Evolution,* 43**,** 304-311.

Rannala, B., Zhu, T. and Yang, Z. (2012) Tail paradox, partial identifiability, and influential priors in Bayesian branch length inference. *Molecular Biology and Evolution,* 29**,** 325-335.

Rawn, S. M. and Cross, J. C. (2008) The evolution, regulation, and function of placenta-specific genes. *Annual Review of Cell and Developmental Biology,* 24**,** 159-181.

Reczko, M., Maragkakis, M., Alexiou, P., Grosse, I. and Hatzigeorgiou, A. G. (2012) Functional microRNA targets in protein coding sequences. *Bioinformatics,* 28**,** 771-776.

Reczko, M., Maragkakis, M., Alexiou, P., Papadopoulos, G. L. and Hatzigeorgiou, A. G. (2011) Accurate microRNA target prediction using detailed binding site accessibility and machine learning on proteomics data. *Frontiers in Genetics,* 2**,** 103.

Reeves, J. H. (1992) Heterogeneity in the substitution process of amino acid sites of proteins coded for by mitochondrial DNA. *Journal of Molecular Evolution,* 35**,** 17-31.

Rehmsmeier, M., Steffen, P., Hochsmann, M. and Giegerich, R. (2004) Fast and effective prediction of microRNA/target duplexes. *RNA,* 10**,** 1507-1517.

Reinhart, B. J., Slack, F. J., Basson, M., Pasquinelli, A. E., Bettinger, J. C., Rougvie, A. E., Horvitz, H. R. and Ruvkun, G. (2000) The 21-nucleotide let-7 RNA regulates developmental timing in Caenorhabditis elegans. *Nature,* 403**,** 901-906.

Remm, M., Storm, C. E. and Sonnhammer, E. L. (2001) Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *Journal of Molecular Biology,* 314**,** 1041-1052.

Ren, J., Jin, P., Wang, E., Marincola, F. M. and Stroncek, D. F. (2009) MicroRNA and gene expression patterns in the differentiation of human embryonic stem cells. *Journal of Translational Medicine,* 7**,** 20.

Rennison, D. J., Owens, G. L. and Taylor, J. S. (2012) Opsin gene duplication and divergence in ray-finned fish. *Molecular Phylogenetics and Evolution,* 62**,** 986-1008.

Reyes-Herrera, P. H. and Ficarra, E. (2012) One decade of development and evolution of microRNA target prediction algorithms. *Genomics, Proteomics & Bioinformatics,* 10**,** 254-263.

Reyes-Herrera, P. H., Ficarra, E., Acquaviva, A. and Macii, E. (2011) miREE: miRNA recognition elements ensemble. *BMC Bioinformatics,* 12**,** 454.

Rhoades, M. W., Reinhart, B. J., Lim, L. P., Burge, C. B., Bartel, B. and Bartel, D. P. (2002) Prediction of plant microRNA targets. *Cell,* 110**,** 513-520.

Rice, P., Longden, I. and Bleasby, A. (2000) EMBOSS: the European Molecular Biology Open Software Suite. *Trends in Genetics,* 16**,** 276-277.

Rigoutsos, I. (2009) New tricks for animal microRNAS: targeting of amino acid coding regions at conserved and nonconserved sites. *Cancer Research,* 69**,** 3245-3248.

Ritchie, W., Flamant, S. and Rasko, J. E. J. (2009) Predicting microRNA targets and functions: traps for the unwary. *Nature Methods,* 6**,** 397-398.

Robbins, L. S., Nadeau, J. H., Johnson, K. R., Kelly, M. A., Roselli-Rehfuss, L., Baack, E., Mountjoy, K. G. and Cone, R. D. (1993) Pigmentation phenotypes of variant extension locus alleles result from point mutations that alter MSH receptor function. *Cell,* 72**,** 827-834.

Rodriguez, A., Griffiths-Jones, S., Ashurst, J. L. and Bradley, A. (2004) Identification of mammalian microRNA host genes and transcription units. *Genome Research,* 14**,** 1902-1910.

Rodríguez, F., Oliver, J. L., Marín, A. and Medina, J. R. (1990) The general stochastic model of nucleotide substitution. *Journal of Theoretical Biology,* 142**,** 485-501.

Romiguier, J., Ranwez, V., Douzery, E. J. P. and Galtier, N. (2013) Genomic evidence for large, long-lived ancestors to placental mammals. *Molecular Biology and Evolution,* 30**,** 5-13.

Ronquist, F. and Huelsenbeck, J. P. (2003) MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics,* 19**,** 1572-1574.

Ronquist, F., Huelsenbeck, J. P., van der Mark, P. and Lemey, P. (2007) Bayesian phylogenetic analysis using MrBayes.

Roy, S. W. and Gilbert, W. (2006) The evolution of spliceosomal introns: patterns, puzzles and progress. *Nature Reviews Genetics,* 7**,** 211-221.

Ruby, J. G., Jan, C. H. and Bartel, D. P. (2007a) Intronic microRNA precursors that bypass Drosha processing. *Nature,* 448**,** 83-86.

Ruby, J. G., Stark, A., Johnston, W. K., Kellis, M., Bartel, D. P. and Lai, E. C. (2007b) Evolution, biogenesis, expression, and target predictions of a substantially expanded set of Drosophila microRNAs. *Genome Research,* 17**,** 1850-1864.

Russ, J. and Futschik, M. E. (2010) Comparison and consolidation of microarray data sets of human tissue expression. *BMC Genomics,* 11**,** 305.

Russo, C. A., Takezaki, N. and Nei, M. (1996) Efficiencies of different genes and different tree-building methods in recovering a known vertebrate phylogeny. *Mol Biol Evol,* 13**,** 525-36.

Rzhetsky, A. and Nei, M. (1994) METREE: a program package for inferring and testing minimum-evolution trees. *Computer Applications in the Biosciences,* 10**,** 409-412.

Saetrom, P., Heale, B. S. E., Snøve, O., Aagaard, L., Alluin, J. and Rossi, J. J. (2007) Distance constraints between microRNA target sites dictate efficacy and cooperativity. *Nucleic Acids Research,* 35**,** 2333-2342.

Saha, A., Wittmeyer, J. and Cairns, B. R. (2006) Chromatin remodelling: the industrial revolution of DNA around histones. *Nature Reviews Molecular Cell Biology,* 7**,** 437-447.

Saitou, N. and Nei, M. (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution,* 4**,** 406-425.

San Mauro, D. and Agorreta, A. (2010) Molecular systematics: a synthesis of the common methods and the state of knowledge. *Cellular Molecular Biology Letters,* 15**,** 311-341.

Sandberg, R., Neilson, J. R., Sarma, A., Sharp, P. A. and Burge, C. B. (2008) Proliferating cells express mRNAs with shortened 3′ untranslated regions and fewer microRNA target sites. *Science,* 320**,** 1643-1647.

Saunders, M. A., Liang, H. and Li, W.-H. (2007) Human polymorphism at microRNAs and microRNA target sites. *Proceedings of the National Academy of Sciences of the United States of America,* 104**,** 3300-3305.

Scannell, D. R., Frank, A. C., Conant, G. C., Byrne, K. P., Woolfit, M. and Wolfe, K. H. (2007) Independent sorting-out of thousands of duplicated gene pairs in two yeast species descended from a whole-genome duplication. *Proceedings of the National Academy of Sciences of the United States of America,* 104**,** 8397-8402.

Schadt, E. E., Edwards, S. W., GuhaThakurta, D., Holder, D., Ying, L., Svetnik, V., Leonardson, A., Hart, K. W., Russell, A., Li, G., Cavet, G., Castle, J., McDonagh, P., Kan, Z., Chen, R., Kasarskis, A., Margarint, M., Caceres, R. M., Johnson, J. M., Armour, C. D., Garrett-Engele, P. W., Tsinoremas, N. F. and Shoemaker, D. D. (2004) A comprehensive transcript index of the human genome generated using microarrays and computational approaches. *Genome Biology,* 5**,** R73.

Schmid, K. and Yang, Z. (2008) The trouble with sliding windows and the selective pressure in BRCA1. *PLoS ONE,* 3**,** e3746.

Schnall-Levin, M., Rissland, O. S., Johnston, W. K., Perrimon, N., Bartel, D. P. and Berger, B. (2011) Unusually effective microRNA targeting within repeat-rich coding regions of mammalian mRNAs. *Genome Research,* 21**,** 1395-1403.

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 Biology and Evolution,* 1**,** 114-118.

Schwarz, D. S., Ding, H., Kennington, L., Moore, J. T., Schelter, J., Burchard, J., Linsley, P. S., Aronin, N., Xu, Z. and Zamore, P. D. (2006) Designing siRNA that distinguish between genes that differ by a single nucleotide. *PLoS Genetics,* 2**,** e140.

Schwarz, D. S., Hutvágner, G., Du, T., Xu, Z., Aronin, N. and Zamore, P. D. (2003) Asymmetry in the assembly of the RNAi enzyme complex. *Cell,* 115**,** 199-208.

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

Seal, R. L., Gordon, S. M., Lush, M. J., Wright, M. W. and Bruford, E. A. (2011) genenames.org: the HGNC resources in 2011. *Nucleic Acids Research,* 39**,** D514-9.

Seitz, H. (2009) Redefining microRNA targets. *Current Biology,* 19**,** 870-873.

Seitz, H., Youngson, N., Lin, S.-P., Dalbert, S., Paulsen, M., Bachellerie, J.-P., Ferguson-Smith, A. C. and Cavaillé, J. (2003) Imprinted microRNA genes transcribed antisense to a reciprocally imprinted retrotransposon-like gene. *Nature Genetics,* 34**,** 261-262.

Sekita, Y., Wagatsuma, H., Nakamura, K., Ono, R., Kagami, M., Wakisaka, N., Hino, T., Suzuki-Migishima, R., Kohda, T., Ogura, A., Ogata, T., Yokoyama, M., Kaneko-Ishino, T. and Ishino, F. (2008) Role of retrotransposon-derived imprinted gene, Rtl1, in the feto-maternal interface of mouse placenta. *Nature Genetics,* 40**,** 243-248.

Selbach, M., Schwanhäusser, B., Thierfelder, N., Fang, Z., Khanin, R. and Rajewsky, N. (2008) Widespread changes in protein synthesis induced by microRNAs. *Nature,* 455**,** 58-63.

Sempere, L. F., Cole, C. N., McPeek, M. A. and Peterson, K. J. (2006) The phylogenetic distribution of metazoan microRNAs: insights into evolutionary complexity and constraint. *Journal of Experimental Zoology,* 306B**,** 575-588.

Sethupathy, P., Megraw, M. and Hatzigeorgiou, A. G. (2006) A guide through present computational approaches for the identification of mammalian microRNA targets. *Nature Methods,* 3**,** 881-886.

Shin, C. and Manley, J. L. (2004) Cell signalling and the control of pre-mRNA splicing. *Nature Reviews Molecular Cell Biology,* 5**,** 727-738.

Shin, C., Nam, J.-W., Farh, K. K.-H., Chiang, H. R., Shkumatava, A. and Bartel, D. P. (2010) Expanding the microRNA targeting code: functional sites with centered pairing. *Molecular Cell,* 38**,** 789-802.

Shomron, N., Golan, D. and Hornstein, E. (2009) An evolutionary perspective of animal microRNAs and their targets. *Journal of Biomedicine and Biotechnology,* 2009.

Shyamsundar, R., Kim, Y. H., Higgins, J. P., Montgomery, K., Jorden, M., Sethuraman, A., van de Rijn, M., Botstein, D., Brown, P. O. and Pollack, J. R. (2005) A DNA microarray survey of gene expression in normal human tissues. *Genome Biology,* 6**,** R22.

Sibley, C. R., Seow, Y., Saayman, S., Dijkstra, K. K., El Andaloussi, S., Weinberg, M. S. and Wood, M. J. A. (2011) The biogenesis and characterization of mammalian microRNAs of mirtron origin. *Nucleic Acids Research,* 40**,** 438-448.

Siepel, A. and Haussler, D. (2004) Phylogenetic estimation of context-dependent substitution rates by maximum likelihood. *Molecular Biology and Evolution,* 21**,** 468-488.

Smedley, D., Haider, S., Ballester, B., Holland, R., London, D., Thorisson, G. and Kasprzyk, A. (2009) BioMart — biological queries made easy. *BMC Genomics,* 10**,** 22.

Smith, A. A., Wyatt, K., Vacha, J., Vihtelic, T. S., Zigler, J. S., Jr., Wistow, G. J. and Posner, M. (2006) Gene duplication and separation of functions in αB-crystallin from zebrafish (Danio rerio). *The FEBS Journal,* 273**,** 481-90.

Smith, T. F. and Waterman, M. S. (1981) Identification of common molecular subsequences. *Journal of Molecular Biology,* 147**,** 195-197.

Snel, B. B., Bork, P. P. and Huynen, M. M. (2000) Genome evolution-gene fusion versus gene fission. *Trends in Genetics,* 16**,** 9-11.

Song, S., Liu, L., Edwards, S. V. and Wu, S. (2012) Resolving conflict in Eutherian mammal phylogeny using phylogenomics and the multispecies coalescent model. *Proceedings of the National Academy of Sciences,* 109**,** 14942-14947.

Sonnhammer, E. L. L. and Koonin, E. V. (2002) Orthology, paralogy and proposed classification for paralog subtypes. *Trends in Genetics,* 18**,** 619-620.

Sperling, E. A. and Peterson, K. J. (2009) MicroRNAs and metazoan phylogeny: big trees from little genes. *In:* MJ, T. & DTJ, L. (eds.) *Animal evolution: genomes, fossils, and trees.*

Spiegelhalter, D. J., Best, N. G., Carlin, B. P. and Van Der Linde, A. (2002) Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society B: Statistical Methodology,* 64**,** 583-639.

Spitz, F. and Furlong, E. E. M. (2012) Transcription factors: from enhancer binding to developmental control. *Nature Publishing Group,* 13**,** 613-626.

Stark, A., Brennecke, J., Bushati, N., Russell, R. B. and Cohen, S. M. (2005) Animal MicroRNAs confer robustness to gene expression and have a significant impact on 3′UTR evolution. *Cell,* 123**,** 1133-1146.

Stark, A., Brennecke, J., Russell, R. B. and Cohen, S. M. (2003) Identification of Drosophila microRNA targets. *PLoS Biology,* 1**,** E60.

Stark, A., Bushati, N., Jan, C. H., Kheradpour, P., Hodges, E., Brennecke, J., Bartel, D. P., Cohen, S. M. and Kellis, M. (2008) A single Hox locus in Drosophila produces functional microRNAs from opposite DNA strands. *Genes & Development,* 22**,** 8-13.

Stark, A., Lin, M. F., Kheradpour, P., Pedersen, J. S., Parts, L., Carlson, J. W., Crosby, M. A., Rasmussen, M. D., Roy, S., Deoras, A. N., Ruby, J. G., Brennecke, J., curators, H. F., Project, B. D. G., Hodges, E., Hinrichs, A. S., Caspi, A., Paten, B., Park, S.-W., Han, M. V., Maeder, M. L., Polansky, B. J., Robson, B. E., Aerts, S., van Helden, J., Hassan, B., Gilbert, D. G., Eastman, D. A., Rice, M., Weir, M., Hahn, M. W., Park, Y., Dewey, C. N., Pachter, L., Kent, W. J., Haussler, D., Lai, E. C., Bartel, D. P., Hannon, G. J., Kaufman, T. C., Eisen, M. B., Clark, A. G., Smith, D., Celniker, S. E., Gelbart, W. M. and Kellis, M. (2007) Discovery of functional elements in 12 Drosophila genomes using evolutionary signatures. *Nature,* 450**,** 219-232.

Stern, D. L. (1998) A role of Ultrabithorax in morphological differences between Drosophila species. *Nature,* 396**,** 463-466.

Stone, J. R. and Wray, G. A. (2001) Rapid evolution of cis-regulatory sequences via local point mutations. *Molecular Biology and Evolution,* 18**,** 1764-1770.

Studer, R. A. and Robinson-Rechavi, M. (2009) How confident can we be that orthologs are similar, but paralogs differ? *Trends in Genetics,* 25**,** 210-216.

Sturm, M., Hackenberg, M., Langenberger, D. and Frishman, D. (2010) TargetSpy: a supervised machine learning approach for microRNA target prediction. *BMC Bioinformatics,* 11**,** 292.

Su, A. I., Cooke, M. P., Ching, K. A., Hakak, Y., Walker, J. R., Wiltshire, T., Orth, A. P., Vega, R. G., Sapinoso, L. M., Moqrich, A., Patapoutian, A., Hampton, G. M., Schultz, P. G. and Hogenesch, J. B. (2002) Large-scale analysis of the human and mouse transcriptomes. *Proceedings of the National Academy of Sciences of the United States of America,* 99**,** 4465-4470.

Sukumaran, J. and Holder, M. T. (2010) DendroPy: a Python library for phylogenetic computing. *Bioinformatics,* 26**,** 1569-1571.

Suzuki, M. M. and Bird, A. (2008) DNA methylation landscapes: provocative insights from epigenomics. *Nature Publishing Group,* 9**,** 465-476.

Suzuki, Y. (2008) False-positive results obtained from the branch-site test of positive selection. *Genes & Genetic Systems,* 83**,** 331-338.

Svoboda, P. and Di Cara, A. (2006) Hairpin RNA: a secondary structure of primary importance. *Cellular and Molecular Life Sciences,* 63**,** 901-908.

Swofford, D. L., Waddell, P. J., Huelsenbeck, J. P., Foster, P. G., Lewis, P. O. and Rogers, J. S. (2001) Bias in phylogenetic estimation and its relevance to the choice between parsimony and likelihood methods. *Systematic Biology,* 50**,** 525-539.

Syvanen, M. (1985) Cross-species gene transfer; implications for a new theory of evolution. *Journal of Theoretical Biology,* 112**,** 333-343.

Takanori, K. and McNamara, J. (2012) OLE::Storage_Lite. 0.19 ed.: CPAN.

Tam, W. (2001) Identification and characterization of human BIC, a gene on chromosome 21 that encodes a noncoding RNA. *Gene,* 274**,** 157-167.

Tamura, K. (1992) Estimation of the number of nucleotide substitutions when there are strong transition-transversion and G+C-content biases. *Molecular Biology and Evolution,* 9**,** 678-687.

Tamura, K. and Nei, M. (1993) Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Molecular Biology and Evolution,* 10**,** 512-526.

Tan, L. P., Seinen, E., Duns, G., de Jong, D., Sibon, O. C. M., Poppema, S., Kroesen, B.-J., Kok, K. and van den Berg, A. (2009) A high throughput experimental approach to identify miRNA targets in human cells. *Nucleic Acids Research,* 37**,** e137.

Tanaka, T. and Nei, M. (1989) Positive Darwinian selection observed at the variable-region genes of immunoglobulins. *Molecular Biology and Evolution,* 6**,** 447-459.

Tanzer, A. and Stadler, P. F. (2004) Molecular evolution of a microRNA cluster. *Journal of Molecular Biology,* 339**,** 327-335.

Tatusov, R. L., Koonin, E. V. and Lipman, D. J. (1997) A genomic perspective on protein families. *Science,* 278**,** 631-637.

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

Tay, Y., Zhang, J., Thomson, A. M., Lim, B. and Rigoutsos, I. (2008) MicroRNAs to Nanog, Oct4 and Sox2 coding regions modulate embryonic stem cell differentiation. *Nature,* 455**,** 1124-1128.

Taylor, M. S., Ponting, C. P. and Copley, R. R. (2004) Occurrence and consequences of coding sequence insertions and deletions in Mammalian genomes. *Genome Research,* 14**,** 555-566.

Thadani, R. and Tammi, M. T. (2006) MicroTar: predicting microRNA targets from RNA duplexes. *BMC Bioinformatics,* 7**,** S20.

Thomas, P. D., Wood, V., Mungall, C. J., Lewis, S. E. and Blake, J. A. (2012) On the use of gene ontology annotations to assess functional similarity among orthologs and paralogs: a short report. *PLoS Computational Biology,* 8**,** e1002386.

Thompson, J. D., Higgins, D. G. and Gibson, T. J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research,* 22**,** 4673-4680.

Tierney, L. (1994) Markov chains for exploring posterior distributions. *The annals of statistics***,** 1701-1728.

Tompa, M., Li, N., Bailey, T. L., Church, G. M., De Moor, B., Eskin, E., Favorov, A. V., Frith, M. C., Fu, Y., Kent, W. J., Makeev, V. J., Mironov, A. A., Noble, W. S., Pavesi, G., Pesole, G., Régnier, M., Simonis, N., Sinha, S., Thijs, G., van Helden, J., Vandenbogaert, M., Weng, Z., Workman, C., Ye, C. and Zhu, Z. (2005) Assessing computational tools for the discovery of transcription factor binding sites. *Nature Biotechnology,* 23**,** 137-144.

Tsai, K.-W., Kao, H.-W., Chen, H.-C., Chen, S.-J. and Lin, W.-c. (2009) Epigenetic control of the expression of a primate-specific microRNA cluster in human cancer cells. *Epigenetics,* 4**,** 587-592.

Tyler, D. M., Okamura, K., Chung, W. J., Hagen, J. W., Berezikov, E., Hannon, G. J. and Lai, E. C. (2008) Functionally distinct regulatory RNAs generated by bidirectional transcription and processing of microRNA loci. *Genes & Development,* 22**,** 26-36.

Tyndale-Biscoe, C. H. and CSIRO Publishing. (2005) *Life of marsupials,* Collingwood, Vic., CSIRO Publishing.

Tyndale-Biscoe, C. H. and Renfree, M. B. (1987) *Reproductive physiology of marsupials,* Cambridge Cambridgeshire ; New York, Cambridge University Press.

Ueda, T. (1996) Simple method for the detection of outliers. *Japanese Journal of Applied Statistics,* 25**,** 17-26.

Ui-Tei, K., Naito, Y., Nishi, K., Juni, A. and Saigo, K. (2008) Thermodynamic stability and Watson-Crick base pairing in the seed duplex are major determinants of the efficiency of the siRNA-based off-target effect. *Nucleic Acids Research,* 36**,** 7100-7109.

UniProt Consortium, T. (2011) Ongoing and future developments at the Universal Protein Resource. *Nucleic Acids Research,* 39**,** D214-9.

Varki, A., Cummings, R. D., Esko, J. D., Freeze, H. H., Stanley, P., Bertozzi, C. R., Hart, G. W., Etzler, M. E., Akimoto, Y., Brockhausen, I., Colley, K. J., Crocker, P. R., Doering, T. L., Elbein, A. D., Ferguson, M. A., Finney, N., Gagneux, P., Haltiwanger, R. S., Hascall, V., Henrissat, B., Kannagi, R., Kimata, K., Kinoshita, T., Kornfeld, S., Lindahl, U., Linhardt, R. J., Liu, F.-T., Lowe, J. B., McEver, R. P., Mohnen, D., Mulloy, B., Nizet, V., Parodi, A. J., Rabuka, D., Raetz, C. R., Rini, J. M., Sasisekharan, R., Schachter, H., Schauer, R., Schnaar, R. L., Seeberger, P. H., Selleck, S. B., Sharon, N., Surolia, A., Suzuki, A., Taniguchi, N., Tiemeyer, M., Toole, B. P., Turco, S., Vacquier, V. D. and West, C. M. (2009) *Essentials of glycobiology.,* Cold Spring Harbor, Cold Spring Harbor Laboratory Press.

Vasudevan, S., Tong, Y. and Steitz, J. A. (2007) Switching from repression to activation: microRNAs can up-regulate translation. *Science,* 318**,** 1931-1934.

Vejnar, C. E. and Zdobnov, E. M. (2012) miRmap: Comprehensive prediction of microRNA target repression strength. *Nucleic Acids Research*.

Vella, M. C. (2004) The C. elegans microRNA let-7 binds to imperfect let-7 complementary sites from the lin-41 3′UTR. *Genes & Development,* 18**,** 132-137.

Vergara, I. A., Norambuena, T., Ferrada, E., Slater, A. W. and Melo, F. (2008) StAR: a simple tool for the statistical comparison of ROC curves. *BMC Bioinformatics,* 9**,** 265.

Vergoulis, T., Vlachos, I. S., Alexiou, P., Georgakilas, G., Maragkakis, M., Reczko, M., Gerangelos, S., Koziris, N., Dalamagas, T. and Hatzigeorgiou, A. G. (2012) TarBase 6.0: capturing the exponential growth of miRNA targets with experimental support. *Nucleic Acids Research,* 40**,** D222-9.

Vilella, A. J., Severin, J., Ureta-Vidal, A., Heng, L., Durbin, R. and Birney, E. (2009) EnsemblCompara GeneTrees: complete, duplication-aware phylogenetic trees in vertebrates. *Genome Res,* 19**,** 327-35.

Vogel, P. (2005) The current molecular phylogeny of Eutherian mammals challenges previous interpretations of placental evolution. *Placenta,* 26**,** 591-596.

Walitza, S., Wewetzer, C., Warnke, A., Gerlach, M., Geller, F., Gerber, G., Görg, T., Herpertz-Dahlmann, B., Schulz, E., Remschmidt, H., Hebebrand, J. and Hinney, A. (2002) 5-HT2A promoter polymorphism -1438G/A in children and adolescents with obsessive-compulsive disorders. *Molecular Psychiatry,* 7**,** 1054-1057.

Wall, D. P., Fraser, H. B. and Hirsh, A. E. (2003) Detecting putative orthologs. *Bioinformatics,* 19**,** 1710-1711.

Wang, Q.-H., Zhou, M., Sun, J., Ning, S.-W., Li, Y., Chen, L., Zheng, Y., Li, X., Lv, S.-l. and Li, X. (2010) Systematic analysis of human microRNA divergence based on evolutionary emergence. *FEBS Letters***,** 1-9.

Wang, W., Yu, H. and Long, M. (2004) Duplication-degeneration as a mechanism of gene fission and the origin of new genes in Drosophila species. *Nature Genetics,* 36**,** 523-527.

Wang, X. (2006) Systematic identification of microRNA functions by combining target prediction and expression profiling. *Nucleic Acids Research,* 34**,** 1646-52.

Wang, X. (2008) miRDB: a microRNA target prediction and functional annotation database with a wiki interface. *RNA,* 14**,** 1012-1017.

Wang, X. and El Naqa, I. M. (2008) Prediction of both conserved and nonconserved microRNA targets in animals. *Bioinformatics,* 24**,** 325-332.

Wasserman, W. W. and Sandelin, A. (2004) Applied bioinformatics for the identification of regulatory elements. *Nature Reviews Genetics,* 5**,** 276-287.

Webb, G. I. (2000) MultiBoosting: a technique for combining boosting and wagging. *Machine learning,* 40**,** 159-196.

Wen, J., Parker, B. J., Jacobsen, A. and Krogh, A. (2011) MicroRNA transfection and AGO-bound CLIP-seq data sets reveal distinct determinants of miRNA action. *RNA,* 17**,** 820-834.

Wheeler, B. M., Heimberg, A. M., Moy, V. N., Sperling, E. A., Holstein, T. W., Heber, S. and Peterson, K. J. (2009) The deep evolution of metazoan microRNAs. *Evolution & Development,* 11**,** 50-68.

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

Wightman, B., Burglin, T. R., Gatto, J., Arasu, P. and Ruvkun, G. (1991) Negative regulatory sequences in the lin-14 3′-untranslated region are necessary to generate a temporal switch during Caenorhabditis elegans development. *Genes & Development,* 5**,** 1813-1824.

Wightman, B., Ha, I. and Ruvkun, G. (1993) Posttranscriptional regulation of the heterochronic gene lin-14 by lin-4 mediates temporal pattern formation in C. elegans. *Cell,* 75**,** 855-862.

Wildman, D. E., Chen, C., Erez, O., Grossman, L. I., Goodman, M. and Romero, R. (2006) Evolution of the mammalian placenta revealed by phylogenetic analysis. *Proceedings of the National Academy of Sciences of the United States of America,* 103**,** 3203-3208.

Willi, Y., Van Buskirk, J. and Hoffmann, A. A. (2006) Limits to the adaptive potential of small populations. *Annual Review of Ecology, Evolution, and Systematics***,** 433-458.

Witten, I. H. and Frank, E. (2005) *Data mining: practical machine learning tools and techniques,* Amsterdam ; Boston, MA, Morgan Kaufman.

Wloch, D. M., Szafraniec, K., Borts, R. H. and Korona, R. (2001) Direct estimate of the mutation rate and the distribution of fitness effects in the yeast Saccharomyces cerevisiae. *Genetics,* 159**,** 441-452.

Wong, K. M., Suchard, M. A. and Huelsenbeck, J. P. (2008) Alignment uncertainty and genomic analysis. *Science,* 319**,** 473-476.

Wong, W. S. W., 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**,** 1041-1051.

Wray, G. A. (2007) The evolutionary significance of cis-regulatory mutations. *Nature Reviews Genetics,* 8**,** 206-216.

Wu, S., Huang, S., Ding, J., Zhao, Y., Liang, L., Liu, T., Zhan, R. and He, X. (2010) Multiple microRNAs modulate p21Cip1/Waf1 expression by directly targeting its 3′ untranslated region. *Oncogene,* 29**,** 2302-2308.

Wuchty, S., Fontana, W., Hofacker, I. L. and Schuster, P. (1999) Complete suboptimal folding of RNA and the stability of secondary structures. *Biopolymers,* 49**,** 145-165.

Xia, X. (1996) Maximizing transcription efficiency causes codon usage bias. *Genetics,* 144, 1309-1320.

Xiao, F., Zuo, Z., Cai, G., Kang, S., Gao, X. and Li, T. (2009) miRecords: an integrated resource for microRNA-target interactions. *Nucleic Acids Research,* 37, D105-D110.

Xiao, S. J., Zhang, C., Zou, Q. and Ji, Z. L. (2010) TiSGeD: a database for tissue-specific genes. *Bioinformatics,* 26, 1273-1275.

Yang, J.-S., Maurin, T., Robine, N., Rasmussen, K. D., Jeffrey, K. L., Chandwani, R., Papapetrou, E. P., Sadelain, M., O'Carroll, D. and Lai, E. C. (2010) Conserved vertebrate mir-451 provides a platform for Dicer-independent, Ago2-mediated microRNA biogenesis. *Proceedings of the National Academy of Sciences,* 107, 15163-15168.

Yang, J. S., Phillips, M. D., Betel, D., Mu, P., Ventura, A., Siepel, A. C., Chen, K. C. and Lai, E. C. (2011) Widespread regulatory activity of vertebrate microRNA* species. *RNA,* 17, 312-326.

Yang, Z. (1994) Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *Journal of Molecular Evolution,* 39, 306-314.

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

Yang, Z. (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. *Computer Applications in the Biosciences,* 13, 555-556.

Yang, Z. (1998) Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. *Molecular Biology and Evolution,* 15, 568-573.

Yang, Z. (2006) *Computational molecular evolution*, Oxford University Press, USA.

Yang, Z. (2007) PAML 4: phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution,* 24, 1586-1591.

Yang, Z., Goldman, N. and Friday, A. (1994) Comparison of models for nucleotide substitution used in maximum-likelihood phylogenetic estimation. *Molecular Biology and Evolution,* 11, 316-324.

Yang, Z. and Nielsen, R. (2000) Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Molecular Biology and Evolution,* 17, 32-43.

Yang, Z. and Nielsen, R. (2002) Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. *Molecular Biology and Evolution,* 19**,** 908-917.

Yang, Z., Nielsen, R. and Goldman, N. (2009) In defense of statistical methods for detecting positive selection. *Proceedings of the National Academy of Sciences,* 106**,** E95-author reply E96.

Yang, Z., Nielsen, R., Goldman, N. and Pedersen, A. M. (2000) Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics,* 155**,** 431-449.

Yang, Z. and Rannala, B. (2012) Molecular phylogenetics: principles and practice. *Nature Reviews Genetics,* 13**,** 303-314.

Yang, Z., Wong, W. S. W. and Nielsen, R. (2005) Bayes empirical bayes inference of amino acid sites under positive selection. *Molecular Biology and Evolution,* 22**,** 1107-1118.

Yekta, S., Shih, I.-h. and Bartel, D. P. (2004) MicroRNA-directed cleavage of HOXB8 mRNA. *Science,* 304**,** 594-596.

Yokoyama, S. (1996) Molecular evolution of retinal and nonretinal opsins. *Genes to Cells,* 1**,** 787-794.

Yokoyama, S., Tada, T., Zhang, H. and Britt, L. (2008) Elucidation of phenotypic adaptations: molecular analyses of dim-light vision proteins in vertebrates. *Proceedings of the National Academy of Sciences,* 105**,** 13480-13485.

Yu, J., Wang, F., Yang, G.-H., Wang, F.-L., Ma, Y.-N., Du, Z.-W. and Zhang, J.-W. (2006) Human microRNA clusters: genomic organization and expression profile in leukemia cell lines. *Biochemical and Biophysical Research Communications,* 349**,** 59-68.

Yuan, Z., Sun, X., Liu, H. and Xie, J. (2011) MicroRNA genes derived from repetitive elements and expanded by segmental duplication events in mammalian genomes. *PLoS ONE,* 6**,** e17666.

Zhai, W., Nielsen, R., Goldman, N. and Yang, Z. (2012) Looking for Darwin in genomic sequences — validity and success of statistical methods. *Molecular Biology and Evolution,* 29**,** 2889-2893.

Zhang, J., Nielsen, R. and Yang, Z. (2005) Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. *Molecular Biology and Evolution,* 22**,** 2472-2479.

Zhang, J., Zhang, Y. P. and Rosenberg, H. F. (2002) Adaptive evolution of a duplicated pancreatic ribonuclease gene in a leaf-eating monkey. *Nature Genetics,* 30**,** 411-5.

Zhang, R., Wang, Y.-Q. and Su, B. (2008) Molecular evolution of a primate-specific microRNA family. *Molecular Biology and Evolution,* 25**,** 1493-1502.

Zhang, Y. and Verbeek, F. J. (2010) Comparison and integration of target prediction algorithms for microRNA studies. *Journal of Integrative Bioinformatics,* 7.

Zharkikh, A. (1994) Estimation of evolutionary distances between nucleotide sequences. *Journal of Molecular Evolution,* 39**,** 315-329.

Zheng, Y. and Zhang, W. (2010) Animal microRNA target prediction using diverse sequence-specific determinants. *Journal of Bioinformatics and Computational Biology,* 8**,** 763-788.

Zhou, X., Duan, X., Qian, J. and Li, F. (2009) Abundant conserved microRNA target sites in the 5′-untranslated region and coding sequence. *Genetica,* 137**,** 159-164.

Zuber, V. and Strimmer, K. (2011) High-dimensional regression and variable selection using CAR scores. *Statistical Applications in Genetics and Molecular Biology,* 10.

Zuker, M. and Stiegler, P. (1981) Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Research,* 9**,** 133-148.

Publications

(1) Morgan, C.C., Shakya, K., Webb, A.E., 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 Evolutionary Biology. 12:114 .
doi:10.1186/1471-2148-12-114

Author contribution for Walsh TA: authorship of Codeml wrapper Perl module used in selective pressure analysis (see Section 2.3.2), co-authorship of Methods section.

(2) O’Connell, M.J.*, Loughran, N.B., Walsh, T.A., Donoghue, M.T.A., Schmid, K.J., Spillane, C. 2010. A phylogenetic approach to test for evidence of parental conflict or gene duplications associated with protein-encoding imprinted orthologous genes in placental mammals.
Mammalian Genome. 21:9-10.
doi: 10.1007/s00335-010-9283-5

Author contribution for Walsh TA: performance of Shimodaira–Hasegawa (SH) tests to compare gene phylogenies with corresponding tree phylogenies, to determine if evolutionary history, of the imprinted genes studied, differed from those of non-imprinted genes.

(3) 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 Evolutionary Biology. 10:39.
doi:10.1186/1471-2148-10-39.

Author contribution: authorship of Codeml wrapper Perl module used in selective pressure analysis (see Section 2.3.2).

BMC Evolutionary Biology

RESEARCH ARTICLE Open Access

Colon cancer associated genes exhibit signatures of positive selection at functionally significant positions

Claire C Morgan[1,2†], Kabita Shakya[1,2,3†], Andrew Webb[1,2], Thomas A Walsh[1,2], Mark Lynch[1,2,4], Christine E Loscher[4], Heather J Ruskin[2,3] and Mary J O'Connell[1,2*]

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 suggests 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

* Correspondence: mary.oconnell@dcu.ie
†Equal contributors
[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
Full list of author information is available at the end of the article

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 combine to act as a 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].

To assess if there is evidence for protein functional shift from the patterns of substitution in colon cancer associated genes we have carried out selective pressure analyses using codon models of lineage-specific rate heterogeneity.

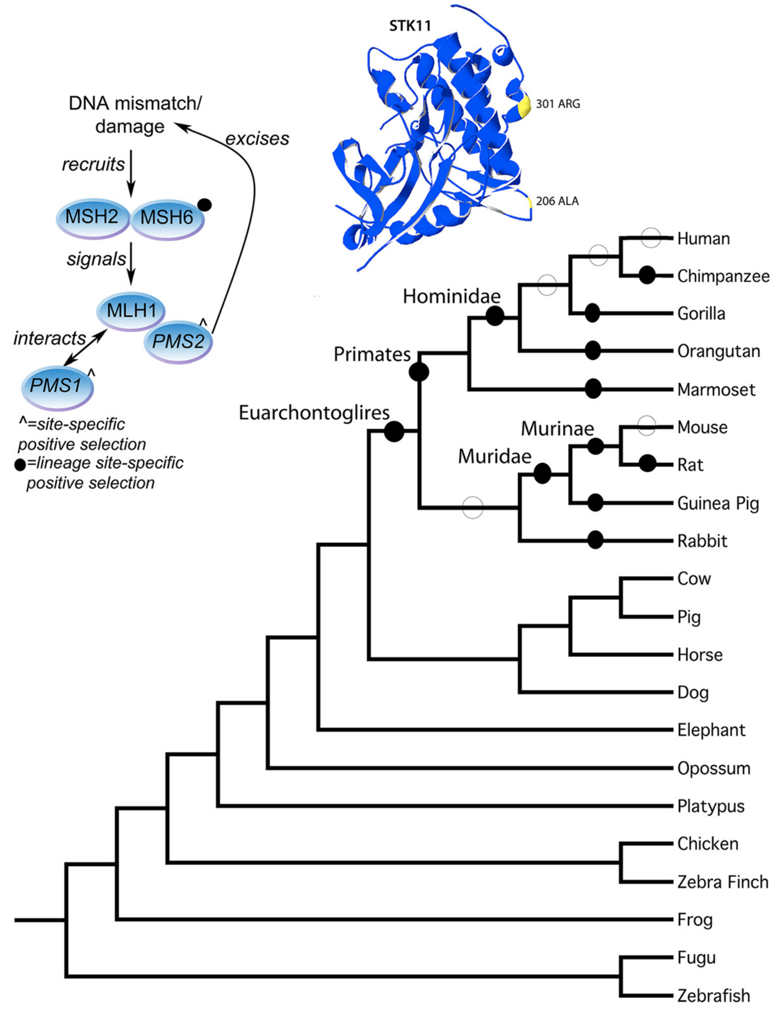


Figure 1 Phylogeny of animal species used in this study. The ancestral lineages tested in the analysis are labeled with their corresponding names as used throughout the text. Those lineages where positive selection was detected are labeled with filled circles, no evidence of positive selection is denoted with an empty circle.

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 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 genome coverage. These included representatives from 3 of the 4 main lineages of Eutheria, namely Afrotheria, Euarchontoglires, and Laurasiatheria, as well as outgroup species including platypus, zebrafish, and zebra finch (see Additional file 1).

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]. 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 reduced the power [34]. To increase the statistical power of the analyses performed here we have therefore only considered single gene families containing 6 or more taxa, and alignment lengths of greater than 500 amino acids.

Recombination analysis

Recombination events can result in the incorrect detection of positive selection. To reduce the detection of potential false positives from our analyses, we have implemented GENECONV (version 1.81a) [35]. GENECONV detects gene conversion events between ancestors of sequences in a multiple sequence alignment. Default parameters where employed, and 10,000 randomly permuted datasets were generated for each Single Gene Orthologous family and global inner fragments were listed if P-value <= 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; 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 1a (Nearly Neutral), Model 2a (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 1a (Nearly Neutral) with Model 2a (Selection) is greater than the critical value determined from the χ^2 distribution, Model 1 a can be rejected. If $\omega_1 > 1$ under Model 2a, positive selection may be inferred. Additional file 3 shows the set of LRTs used for selective pressure 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 these have been shown to be more robust [37].

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], [17]
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], [17]
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 rounded parentheses.

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). 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 setting up and interpreting 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, it 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 criteria 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 were 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 analyses 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, i.e. - 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 are briefly summarized on a gene-by-gene basis. All positively selected sites were assessed using functional information from the Uniprot database [41]. The model of best fit along with associated parameter estimates are described and a summary table 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. 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. In the ancestral Muridae lineage there is evidence for 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.

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 sites 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.

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 clades as depicted in

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

Table 2 Summary of parameter estimates and likelihood scores for the model of best fit showing evidence of positive selection *(Continued)*

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.

Figure 1. The STK11 alignment consists of 18 taxa and was the only gene that showed evidence of positive selection in this ancestral 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 [43,44],). Mutations in STK11 cause Peutz-Jeghers syndrome (PJS), this is 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 Posterior Probability (PP) = 0.889 is a hydrophobic Alanine or Valine in Euarchontoglires species and is 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 analyses showed evidence of lineage-specific 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 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, with ω >>1. See Figure 2(a) and Table 2. The positively selected positions were 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.

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.

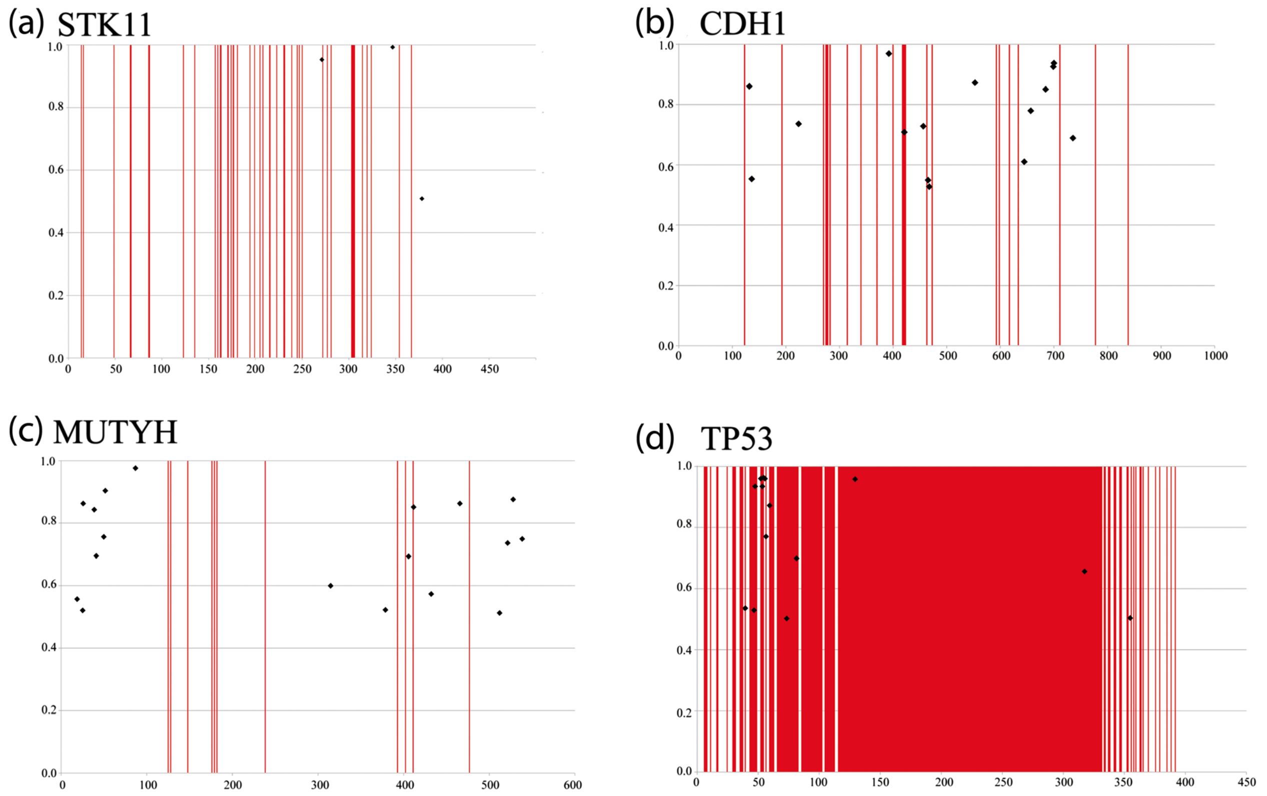


Figure 2 Positive selection analysis for 4 genes: (a) STK11, (b) CDH1, (c) MUTYH, and (d) TP53. The x-axis depicts the gene from start to end of alignment. The Y-axis is the posterior probability. The vertical red bars on each graph represent the known cancer causing variants from human populations. The black dots on each graph represent the positively selected sites identified in this study.

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 where 1.2% of sites have $\omega >> 1$, and also in the Marmoset lineage where 5.5% sites have $\omega >> 1$. 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. In marmoset, positively selected site 803 (PP = 0.551) coincides with Human colorectal cancer variants D803G [51] and V800A [52]. 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 greater predisposition 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 the potential functional impact 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, iHS [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. 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

The 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 substitution 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 consists 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 (ω = 3.59). 15/51 positively selected sites were identified as occurring within 10 amino acid positions of the metal binding site 127, 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 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 (ω = 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 [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 (ω = 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 has 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 ($\omega >> 1$) and 0.12% of sites in guinea pig ($\omega >> 1$). 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 where 1.2% of the sites have $\omega >> 1$. 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 analyses revealed 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 within the extracellular topological domain (positions 155–709). Many of these positively selected sites 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. 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 consists 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 ideal as 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, ω = 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 under positive selection, ω = 1.97. See Figure 2(d) and Table 2 for detailed analyses. 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 MY 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 ancestral branches and in specific extant lineages. In the ancestral lineages 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 force in 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.

Additional files

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: 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 Likelihood ratio tests performed and their associated significance values.

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: 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).

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 RS2000172) 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. MJO'C would also like to thank SFI Travel fellowship 10/RFP/ EOB2673 - STTF 11. 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.

Author details

[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.

Received: 1 February 2012 Accepted: 22 June 2012
Published: 12 July 2012

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.
16. Loughran NB, Hinde S, *et al*: **Functional consequence of positive selection revealed through rational mutagenesis of human myeloperoxidase.** *Mol Biol Evol* 2012, (Published advance access March 28th 2012, page numbers not currently available): doi:10.1093/molbev/mss073.
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 Research* 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.

doi:10.1186/1471-2148-12-114
Cite this article as: Morgan *et al.*: Colon cancer associated genes exhibit signatures of positive selection at functionally significant positions. *BMC Evolutionary Biology* 2012 **12**:114.

This text redacted due to 3rd party copyright
This text redacted due to 3rd party copyright
This text redacted due to 3rd party copyright
This text redacted due to 3rd party copyright
This text redacted due to 3rd party copyright
This text redacted due to 3rd party copyright
This text redacted due to 3rd party copyright
This text redacted due to 3rd party copyright
This text redacted due to 3rd party copyright
This text redacted due to 3rd party copyright
This text redacted due to 3rd party copyright
This text redacted due to 3rd party copyright
This text redacted due to 3rd party copyright
This text redacted due to 3rd party copyright
This text redacted due to 3rd party copyright
This text redacted due to 3rd party copyright
This text redacted due to 3rd party copyright
This text redacted due to 3rd party copyright
This text redacted due to 3rd party copyright
This text redacted due to 3rd party copyright
This text redacted due to 3rd party copyright
This text redacted due to 3rd party copyright
This text redacted due to 3rd party copyright
This text redacted due to 3rd party copyright
This text redacted due to 3rd party copyright

This text redacted due to 3rd party copyright

This text redacted due to 3rd party copyright

This text redacted due to 3rd party copyright

This text redacted due to 3rd party copyright

This text redacted due to 3rd party copyright

This text redacted due to 3rd party copyright

This text redacted due to 3rd party copyright

This text redacted due to 3rd party copyright

This text redacted due to 3rd party copyright

This text redacted due to 3rd party copyright

This text redacted due to 3rd party copyright

This text redacted due to 3rd party copyright

This text redacted due to 3rd party copyright

This text redacted due to 3rd party copyright

This text redacted due to 3rd party copyright

This text redacted due to 3rd party copyright

This text redacted due to 3rd party copyright

This text redacted due to 3rd party copyright

This text redacted due to 3rd party copyright

This text redacted due to 3rd party copyright

This text redacted due to 3rd party copyright

This text redacted due to 3rd party copyright

This text redacted due to 3rd party copyright

This text redacted due to 3rd party copyright

This text redacted due to 3rd party copyright

This text redacted due to 3rd party copyright
This text redacted due to 3rd party copyright
This text redacted due to 3rd party copyright
This text redacted due to 3rd party copyright
This text redacted due to 3rd party copyright
This text redacted due to 3rd party copyright
This text redacted due to 3rd party copyright
This text redacted due to 3rd party copyright
This text redacted due to 3rd party copyright
This text redacted due to 3rd party copyright
This text redacted due to 3rd party copyright
This text redacted due to 3rd party copyright
This text redacted due to 3rd party copyright
This text redacted due to 3rd party copyright
This text redacted due to 3rd party copyright
This text redacted due to 3rd party copyright
This text redacted due to 3rd party copyright
This text redacted due to 3rd party copyright
This text redacted due to 3rd party copyright
This text redacted due to 3rd party copyright
This text redacted due to 3rd party copyright
This text redacted due to 3rd party copyright
This text redacted due to 3rd party copyright
This text redacted due to 3rd party copyright
This text redacted due to 3rd party copyright

This text redacted due to 3rd party copyright
This text redacted due to 3rd party copyright
This text redacted due to 3rd party copyright
This text redacted due to 3rd party copyright
This text redacted due to 3rd party copyright
This text redacted due to 3rd party copyright
This text redacted due to 3rd party copyright
This text redacted due to 3rd party copyright
This text redacted due to 3rd party copyright
This text redacted due to 3rd party copyright
This text redacted due to 3rd party copyright
This text redacted due to 3rd party copyright
This text redacted due to 3rd party copyright
This text redacted due to 3rd party copyright
This text redacted due to 3rd party copyright
This text redacted due to 3rd party copyright
This text redacted due to 3rd party copyright
This text redacted due to 3rd party copyright
This text redacted due to 3rd party copyright
This text redacted due to 3rd party copyright
This text redacted due to 3rd party copyright
This text redacted due to 3rd party copyright
This text redacted due to 3rd party copyright
This text redacted due to 3rd party copyright

This text redacted due to 3rd party copyright
This text redacted due to 3rd party copyright
This text redacted due to 3rd party copyright
This text redacted due to 3rd party copyright
This text redacted due to 3rd party copyright
This text redacted due to 3rd party copyright
This text redacted due to 3rd party copyright
This text redacted due to 3rd party copyright
This text redacted due to 3rd party copyright
This text redacted due to 3rd party copyright
This text redacted due to 3rd party copyright
This text redacted due to 3rd party copyright
This text redacted due to 3rd party copyright
This text redacted due to 3rd party copyright
This text redacted due to 3rd party copyright
This text redacted due to 3rd party copyright
This text redacted due to 3rd party copyright
This text redacted due to 3rd party copyright
This text redacted due to 3rd party copyright
This text redacted due to 3rd party copyright
This text redacted due to 3rd party copyright
This text redacted due to 3rd party copyright
This text redacted due to 3rd party copyright
This text redacted due to 3rd party copyright
This text redacted due to 3rd party copyright

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

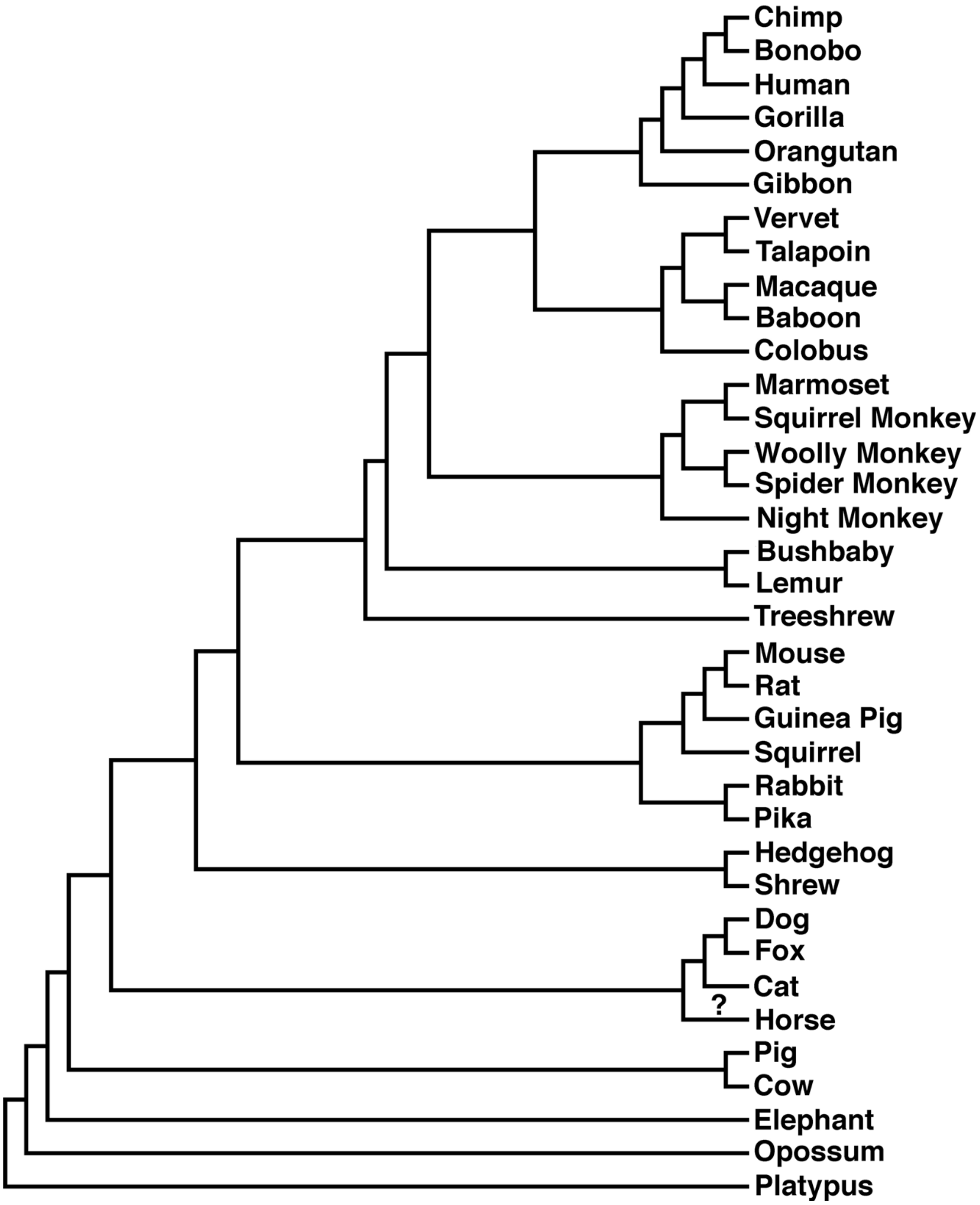


Figure 1 Canonical mammalian species phylogeny. Shown here is a representation of the agreed relationships amongst the *mammalia* for the species used in this analysis. The "?" on the lineage leading to horse indicates controversy over the position of this lineage on the phylogeny.

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

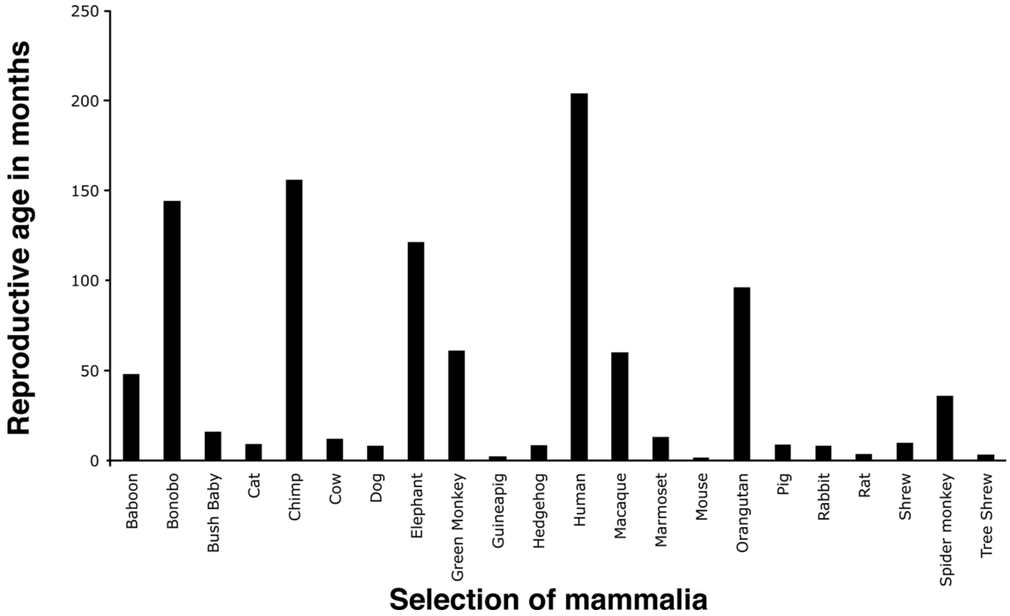


Figure 2 Selection of mammalia used in the analysis and the time to reproductive age in months. Species are shown on the X-axis in alphabetical order. On the Y-axis is the number of months it takes for each species to reach reproductive age.

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

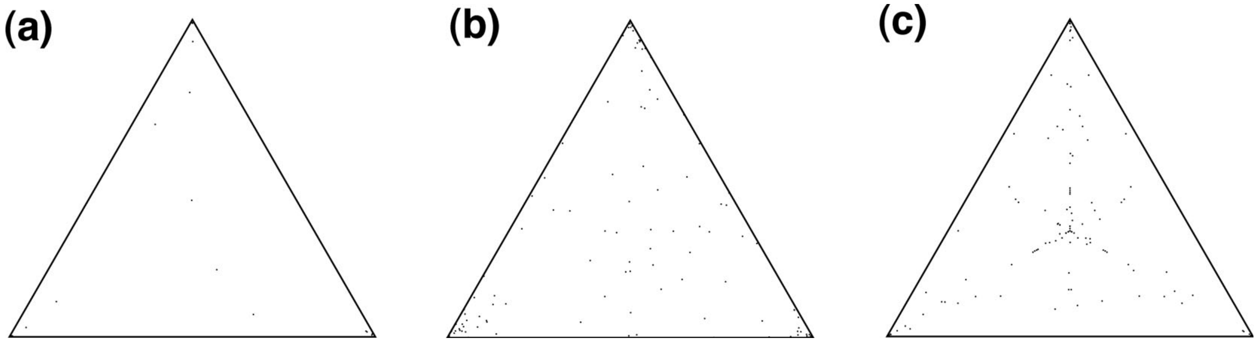


Figure 3 Example of likelihood mapping categories. **(a), category 1** genes with strong phylogenetic signal the example given here is the ZP2 gene, **(b) category 2** genes with intermediate levels of phylogenetic signal, the example given here is the Prkar2a gene, and **(c)**, **category 3** genes 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

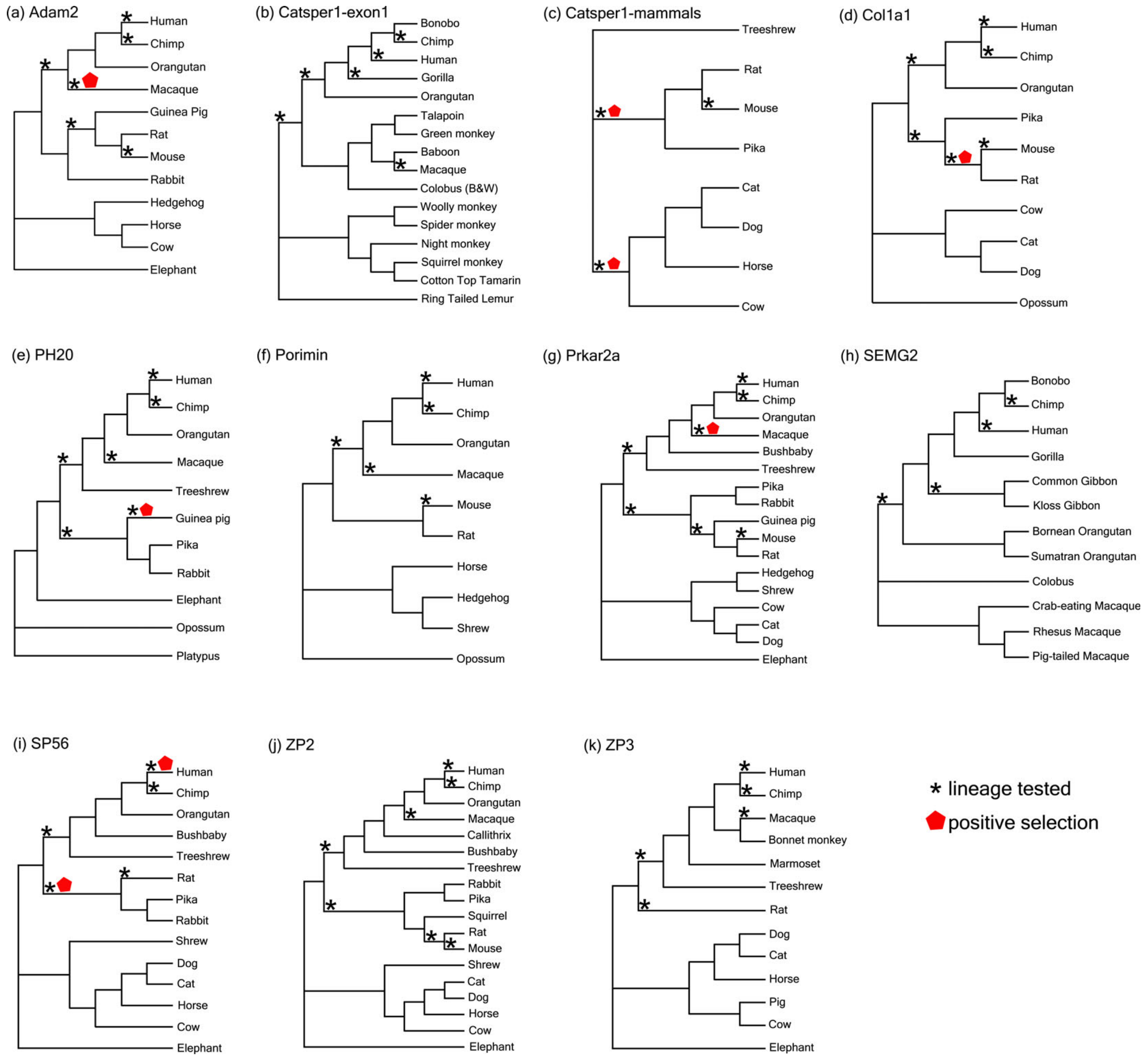


Figure 4 Results of lineage specific positive selection analysis on 11 datasets. The phylogeny used for each gene is a reduced version of the species phylogeny. The lineages labeled as foreground in the analysis are denoted in the diagram with asterix symbols. These are the results following LRT analysis. 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

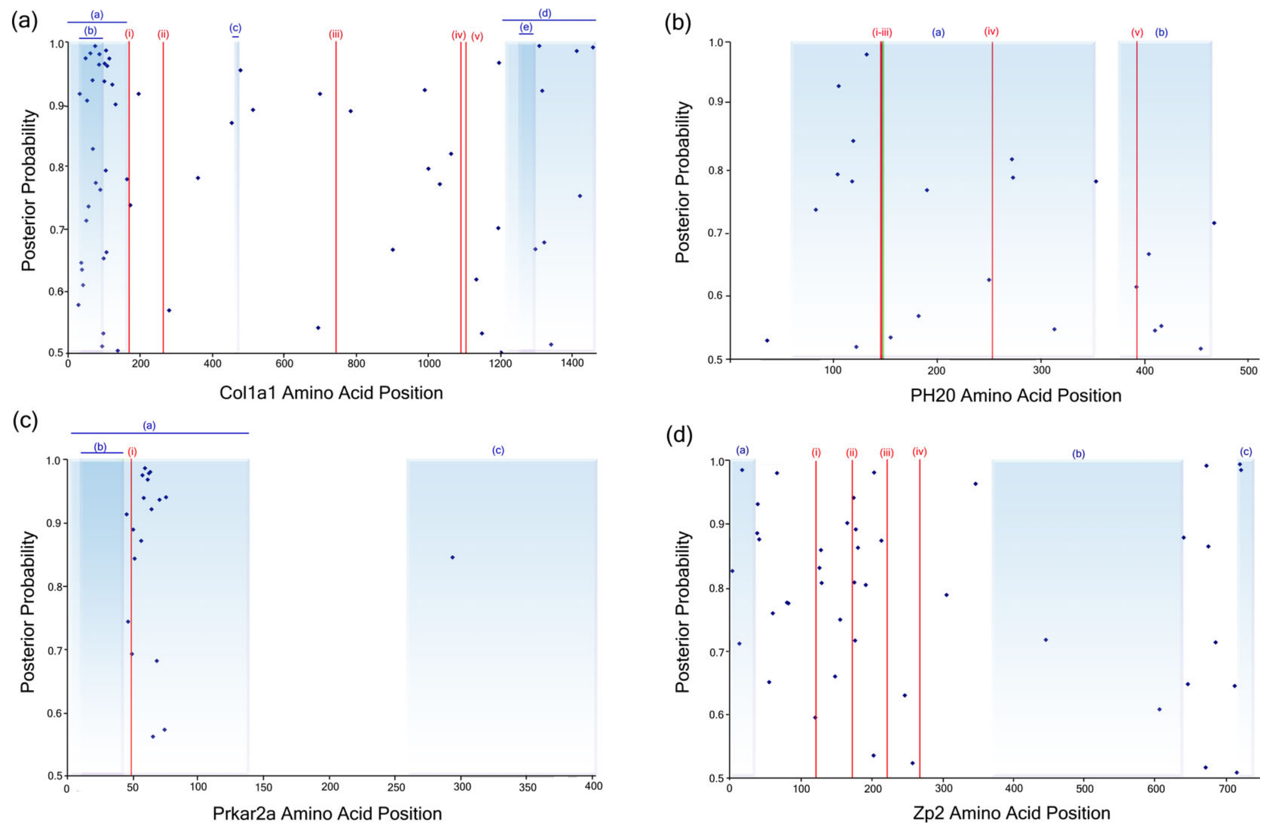


Figure 5 Results of positive selection analysis for 4 genes. Each of the four graphs represents the CDS of a gene from position 1 to the stop codon (X-axis). The Y-axis is the posterior probability of each of the sites belonging to the positively selected category. The dark blue data points are sites 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 ω = 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 $\omega = 3.13$. 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 11th 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.