EVOLUTIONARY ANALYSIS OF MAMMALIAN GENOMES AND ASSOCIATIONS TO HUMAN DISEASE

JESSICA JANAKI VAMATHEVAN

Department of Biology

University College London

2008

Submitted to University College London

for the degree of Doctor of Philosophy

AUTHORSHIP DECLARATION

I, Jessica Janaki Vamathevan, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

ABSTRACT

Statistical models of DNA sequence evolution for analysing protein-coding genes can be used to estimate rates of molecular evolution and to detect signals of natural selection. Genes that have undergone positive selection during evolution are indicative of functional adaptations that drive species differences.

Genes that underwent positive selection during the evolution of humans and four mammals used to model human diseases (mouse, rat, chimpanzee and dog) were identified, using maximum likelihood methods. I show that genes under positive selection during human evolution are implicated in diseases such as epithelial cancers, schizophrenia, autoimmune diseases and Alzheimer's disease. Comparisons of humans with great apes have shown such diseases to display biomedical disease differences, such as varying degrees of pathology, differing symptomatology or rates of incidence.

The chimpanzee lineage was found to have more adaptive genes than any of the other lineages. In addition, evidence was found to support the hypothesis that positively selected genes tend to interact with each other. This is the first such evidence to be detected among mammalian genes and may be important in identifying molecular pathways causative of species differences.

The genome scan analysis spurred an in-depth evolutionary analysis of the nuclear receptors, a family of transcription factors. 12 of the 48 nuclear receptors were found to be under positive selection in mammalia. The androgen receptor was found to have undergone positive selection along the human lineage. Positively selected sites were found to be present in the major activation domain, which has implications for ligand recognition and binding.

Studying the evolution of genes which are associated with biomedical disease differences between species is an important way to gain insight into the molecular causes of diseases and may provide a method to predict when animal models do not mirror human biology.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGEMENTS

First and foremost, I would like to thank God for his continuous blessings, for which I will be forever grateful.

I would like to thank my supervisor Ziheng Yang for making it possible for me to undertake this PhD programme. I am extremely appreciative of his guidance, his expertise and patience throughout the project.

I am also extremely grateful to Joanna Holbrook, my industrial supervisor at GlaxoSmithKline (GSK), for her advice and support. I would also like to especially thank Richard Emes for his direction, encouragement and enthusiasm.

The genome scan study presented in Chapters 2 – 5 resulted in the following publications:

Vamathevan J, Hasan S, Emes RD, Amrine-Madsen H, Rajagopalan D, Topp SD, Kumar V, Word M, Simmons MD, Foord SM, Sanseau P, Yang Z, Holbrook JD (2008) The Role of Positive Selection in Determining the Molecular Cause of Species Differences in Disease. ***BMC Evolutionary Biology*** *(accepted).*

Vamathevan J, Holbrook JD and Emes RD (2007) The Mouse Genome as a Rodent Model in Evolutionary Studies. ***Encyclopedia of Life Sciences 2007****. John Wiley & Sons, Ltd.*

This 3 year project would not have been possible without the financial support of a GSK studentship and the funding provided from BBSRC and NERC to Ziheng Yang.

Special thanks also go to my mother and father, Samuel, Veronica, Rupert, Dayalan, Vasanthi, Chloe, Andrew, Varsha, Mareeni, Amelia, Sanjay, Ralph and Maria. I am very grateful for your help and support.

Finally my heartfelt thanks go to all my family members and friends worldwide, for their love and support during the last three years.

ABBREVIATIONS AND DEFINITIONS

Abbreviations

BEB	Bayes empirical Bayes
DBD	DNA-Binding Domain
CSAC	Chimpanzee Sequencing and Analysis Consortium
LBD	Ligand-Binding Domain
LRT	Likelihood-Ratio Test
MHC	Major Histocompatibility Complex
NEB	Naïve empirical Bayes
NR	Nuclear Receptor
PSG	Positively Selected Gene

Definitions of major symbols

Symbol	Definition
$\boldsymbol{S}$	Number of synonymous sites in a sequence
$\boldsymbol{N}$	Number of nonsynonymous sites in a sequence
$\boldsymbol{S_d}$	Number of synonymous differences between two sequences
$\boldsymbol{N_d}$	Number of nonsynonymous differences between two sequences
$\boldsymbol{d_\mathrm{S}}$	Number of synonymous substitutions per synonymous site
$\boldsymbol{d_\mathrm{N}}$	Number of nonsynonymous substitutions per nonsynonymous site
$\boldsymbol{\kappa}$	Transition/transversion (mutation) rate ratio
$\boldsymbol{\omega}$ or $\boldsymbol{d_\mathrm{N} / d_\mathrm{S}}$	Nonsynonymous /synonymous substitution rate ratio
$\boldsymbol{t}$	Time or branch length, measured as the expected number of (nucleotide) substitutions per codon
$\boldsymbol{\pi_j}$	Equilibrium frequency of codon j

Chapter 1

Introduction

1.1 NATURAL SELECTION

> "*Natural selection is daily, hourly, scrutinising the slightest variations, rejecting those that are bad, preserving and adding up all those that are good.*"
>
> – Charles Darwin, The Origin of Species (1859)

The nineteenth-century attempt by Darwin to explain the mechanistic action by which evolutionary processes bring about variation between species was, in hindsight, only partially correct. The 1920s and 1930s brought forth the theory of neo-Darwinism from pioneers such as R.A. Fisher, J.B.S. Haldane and S. Wright, where mutation was recognised as the major source of genetic variation and natural selection was the dominant factor in shaping genetic makeup. The discovery of DNA in 1953 and the advancement of molecular techniques enabled the search for evidence of adaptation at the molecular level and challenged the Darwinian concept of natural selection.

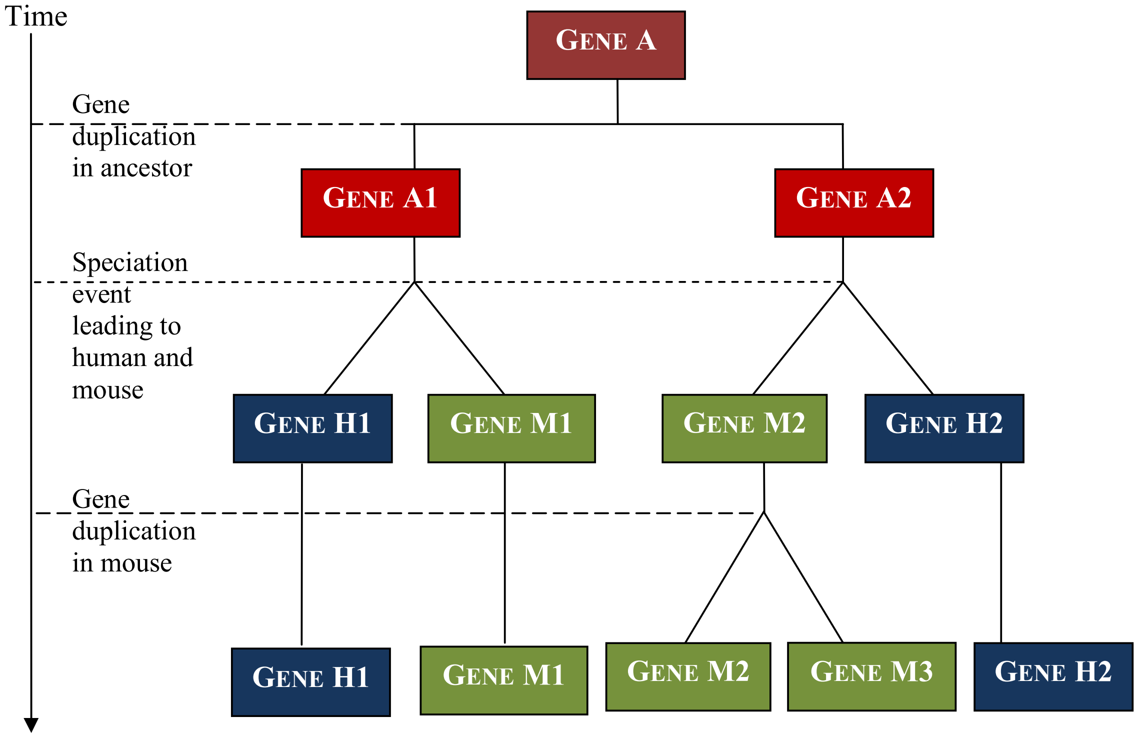
The generally accepted present day viewpoint is that the great majority of evolutionary changes at the molecular level are caused by random drift of selectively neutral or nearly neutral mutations. These neutral mutations have a fitness coefficient which is equal to that of the common allele in the population. The relevant consequence of such a mechanism is that mutations become fixed at a constant rate within the population. This is known as the neutral theory of evolution (Kimura, 1968; King and Jukes, 1969). However, a large proportion of mutations that occur confer a selective disadvantage to the individual. Lethal mutations and mutations that reduce the fitness level of the carrier within a population are removed over time, due to the decreased reproductive success of their carriers. This process is known as negative or purifying selection. In general, genes that are crucial for the basic functions needed for the sustenance of the cell are under purifying selection. A very small proportion of mutations confer a selective advantage to the individual. These mutations, said to be under positive selection, are driven to fixation within the population at rates higher than for neutral mutations. The majority of variation that we see today between species is a result of the interaction of these complex processes, as well as the effects of recombination.

Positive selection, the fixation of advantageous mutations, is an exciting topic as it is ultimately responsible for differences in protein function between species and hence genes involved in adaptation. Positive selection leading to functional divergence in homologous genes may help explain at the molecular level the divergence in the anatomy, biology and cognitive abilities of mammals.

1.2 HOMOLOGOUS GENES

Homology is defined as similarity between a pair of genes that is the result of inheritance from a common ancestor. The accurate identification and analysis of homologies is key to the study of phylogenetic systematics. Homologous genes are subdivided into orthologues and paralogues (Figure 1.1). Orthologous genes are homologous genes in two or more organisms that are the result of speciation and not gene duplication (Fitch, 2000). Paralogous genes are homologous genes that are the result of gene duplication. These are further classified as inparalogues and outparalogues: the term inparalogues indicates paralogues in a given lineage that all evolved by gene duplications that happened after the speciation event that separated the given lineage from the other lineage under consideration, whereas outparalogues are paralogs in the given lineage that evolved by gene duplications that happened before the speciation event.

Figure 1.1 Relationships between orthologous and paralogous genes
Genes A1 and A2 are inparalogues, arising from a duplication of Gene A. Genes M1 and H1 are orthologues since they arose from the same ancestral gene A1. Similarly, genes M2 and H2 are orthologous as well. Genes H1 and M2 are outparalogues, since gene M2 arose from gene duplication before speciation. Genes M2 and M3 are inparalogues since gene M3 is a duplication of gene M2. Genes H2 and M3 are also orthologues since M3 is a duplication of M2.


Time
GENE A
Gene duplication in ancestor
GENE A1
GENE A2
Speciation event leading to human and mouse
GENE H1
GENE M1
GENE M2
GENE H2
Gene duplication in mouse
GENE H1
GENE M1
GENE M2
GENE M3
GENE H2

1.3 MEASURING SELECTION PRESSURE ON A PROTEIN

The availability of DNA sequence information from closely related organisms allows the direct comparison of their encoded protein sequences. Thus, nucleotide differences between homologous proteins can be used to infer the number and type of mutations that have occurred between two species since they last shared a common ancestor. Nucleotide differences can be of two types: a nonsynonymous nucleotide change is one which results in an amino acid change in the protein sequence whereas a synonymous nucleotide substitution leads to no change in amino acid, due to the degeneracy of the genetic code (Figure 1.2).

Figure 1.2 Examples of nonsynonymous and synonymous mutations

Nonsynonymous (replacement) substitution:

Original sequence:	UUU CAU CGU
Mutation:	UUU CA**G** CGU
Original protein sequence:	Phe His Arg
New protein sequence:	Phe **Gln** Arg

Synonymous (silent) substitution:

Original sequence:	UUU CAU CGU
Mutation:	UUU CA**C** CGU
Original protein sequence:	Phe His Arg
New protein sequence:	Phe His Arg

Since synonymous mutations have no effect on protein sequence, they are not subject to natural selection acting on the protein. However, in mammals, synonymous mutations have been found to have an effect on fitness. Such mutations can, for example, disrupt splicing of alternatively spliced exons, can interfere with miRNA binding (Charmary, Parmley and Hurst, 2006). They have also been found to modify protein abundance most probably mediated by alteration in mRNA stability and modification of protein structure and activity probably mediated by translational pausing (Parmley and Hurst, 2007).

The rate of fixation of nonsynonymous mutations is also monitored by selection. Thus, the comparison of the fixation rates of synonymous and nonsynonymous mutations can be used to understand the action of selective pressure on protein coding sequences. Selection pressure can be measured by contrasting the number of nonsynonymous substitutions per non-synonymous site (d_N), with the number of synonymous substitutions per synonymous site (d_S) (Miyata and Yasunaga, 1980).

The d_N/d_S rate ratio is known as ω, and in this measure $\omega = 1$ ($d_N = d_S$) indicates neutral evolution (Yang and Bielawski, 2000). If one or more amino acid substitutions reduce the fitness of the carrier, these changes are likely to be removed by negative selection. This results in the nonsynonymous substitution rate to be less than the synonymous substitution rate ($d_N < d_S$). Hence $\omega < 1$ is indicative of negative selection. If amino acid substitutions confer fitness advantages, the nonsynonymous substitution rate will be greater than the synonymous substitution rate ($d_N > d_S$) leading to $\omega > 1$, an indication of positive Darwinian selection. The resulting alterations in the protein sequence can lead to variations in the secondary structure of the protein as seen in human MHC class 1 molecules or a change in protein conformation exemplified by positive selection in the insulin gene in caviomorph rodents (Opazo et al., 2005). Changes in protein sequence can also result in modifications in the substrate protein binding mechanism such as MRGX2, a G-protein coupled receptor (Yang S et al., 2005) or the means by which it interacts with other proteins as seen in genes in the HOX cluster.

1.4 ESTIMATION OF SYNONYMOUS AND NONSYNONYMOUS SUBSTITUTION RATES

I will first describe the early counting methods which were used to estimate d_N and d_S between a pair of homologous, protein-coding sequences, and then the more realistic approach using maximum likelihood methods under a codon substitution model.

1.4.1 Counting methods: pairwise methods

Perler et al. (1980) developed a simple statistical method to estimate synonymous substitutions. Miyata and Yasunaga (1980) and Li et al. (1985) proposed methods to incorporate different weights for two or more possible evolutionary pathways between a pair of codons. Both these methods were fairly complicated so Nei and Gojobori (1986) devised a simpler method based on the Miyata-Yasunaga method (1980) that gave essentially the same results. The Nei-Gojobori method estimates d_N and d_S in three steps:

Step One:

The numbers of nonsynonymous sites (n) and synonymous sites (s) in each codon are calculated. For a codon, let i be the number of possible synonymous changes at this site. For example, the codon TTA can undergo two synonymous substitutions: one substitution at the first nucleotide T (T $\rightarrow$ C) and one substitution at the third nucleotide A (A $\rightarrow$ G). Therefore for codon TTA, $i = 2$. This site is then counted as having $i/3$ synonymous sites and $(9-i)/3$ nonsynonymous sites, which gives $s = 2/3$ and $n = 7/3$ for codon TTA. The values of s and n from each codon in the sequence are summed to give S and N

for the sequence. To obtain S and N for two sequences which are being compared, the mean values of S and N of each sequence are used.

Step Two:

The numbers of nonsynonymous (N_d) and synonymous (S_d) differences between the two sequences are counted. This step is fairly straightforward if there is only one nucleotide difference between the two codons (e.g. there is one synonymous difference between the codons GTA (Val) and GTT (Val)). However, if the two codons differ by more than one nucleotide, all possible evolutionary pathways between the two codons will have to be evaluated. For example, in the comparison of codons TTT and GTA, the two pathways are:

Pathway I: TTT (Phe) ↔ GTT (Val) ↔ GTA (Val)

Pathway II: TTT (Phe) ↔ TTA (Leu) ↔ GTA (Val)

Pathway I includes one nonsynonymous change and one synonymous change, whereas Pathway II includes two nonsynonymous changes. Pathways I and II can be assumed to occur with equal probability or be weighted for synonymous and nonsynonymous changes. If equal weights are assumed, $S_d = 0.5$ and $N_d = 1.5$ for this pair of codons. However, in almost all genes the synonymous substitution rate is higher than the nonsynonymous substitution rate so, to improve the model, a larger weight is given to synonymous substitutions than nonsynonymous substitutions. The weights are then multiplied by the number of differences to give S_d and N_d. Incorporating weights for different pathways was implemented by Li et al. (1985).

Step Three:

The proportion of different sites at synonymous sites (p_S) is now S_d/S and the proportion of different sites at the nonsynonymous sites (p_N) is N_d/N (see Table of Definitions). p_S and p_N are actually underestimates of the distance between the two sequences (expected number of substitutions) because multiple substitutions could have occurred at the same site, not reflected in the observed sequences. Multiple hits include parallel substitutions, convergent substitutions, and back substitutions. The Jukes and Cantor distance formula (Jukes and Cantor, 1969) applies a correction to p_S and p_N to account for multiple substitutions which results in the number of synonymous substitutions per site (d_S) and the nonsynonymous substitutions per site (d_N) as:

$$d_S = -\frac{3}{4}\log(1-\frac{4}{3}p_s)$$

$$d_N = -\frac{3}{4}\log(1-\frac{4}{3}p_N)$$

The estimates of d_N and d_S obtained by this method can be used to calculate $\omega = d_N/d_S$.

The Nei-Gojobori method (Nei and Gojobori, 1986) was later improved by Ina (1995), which accounts for the transition/transversion bias in nucleotide substitutions. Transversions are substitutions for a purine for a pyrimidine or vice versa which changes the chemical structure of DNA dramatically. It is well known that nucleotide substitutions that are transitions (T ↔ C, and A ↔ G) are more common than transversions (T, C ↔ A, G). Ignoring the transition/transversion bias causes underestimation of S, overestimation of d_S and underestimation of the d_N/d_S ratio (Li, 1993; Pamilo and Bianchi, 1993).

All of the counting methods to estimate the d_N/d_S are relatively simple. However there are several disadvantages:

- The pairwise method averages d_N/d_S over all the sites of the protein and over time which leads to selection pressure being drastically underestimated. For example, Endo et al. (1996) performed a large-scale search of 3595 genes using the Nei and Gojobori method (1986) to estimate d_N and d_S, and only identified 17 proteins under positive selection, a very small proportion of 0.47% probably due to the lack of power in the methods used.

- Most proteins only have very few sites that have undergone positive selection, with most of the protein under strong purifying selection. Also, as adaptive evolution may occur over a very small period of evolutionary time, pairwise methods do not reliably infer the action of positive selection.

- Use of the pairwise method cannot determine the particular codon that is under selection.

- The pairwise method ignores the effects of codon usage bias and unequal nucleotide frequencies. In real data, base compositions and codon usage are quite biased, which implies that the substitution rates are not symmetrical and will affect the counting of sites and differences. Assuming codon frequency to be equal was shown to cause overestimation of S, underestimation of d_S and overestimation of the d_N/d_S ratio by Yang and Nielsen (2000), who implemented a counting method that takes into account unequal nucleotide frequencies.

- Lastly, the Jukes and Cantor distance correction to account for multiple substitutions used in the third step of the pairwise method is not very accurate as the correction procedure is based on correction of multiple hits within nucleotide sequences, not codons.

The last of the counting methods was formulated by Yang and Nielsen (2000). They devised a new approximate method incorporating the transition/transversion bias and unequal base frequencies in their algorithm assuming the HKY85 nucleotide substitution model (Hasegawa, 1985). This method was shown to produce estimates of d_N and d_S very close to the true values even for data with strong transition/transversion and codon biases.

1.4.2 Maximum likelihood estimation methods based on a codon-substitution model

Maximum likelihood is a major statistical inference tool used in a variety of fields. The likelihood of a phylogenetic tree is the probability of observing the data under a given tree and a specific substitution model (such as a codon substitution model), $P(\text{data}|\text{tree})$ (Felsenstein, 1981). Nucleotide substitution models were already in existence so a natural extension was to create the slightly more complex codon-substitution model. Models that focus on detecting selection at the level of individual codons have been shown to fit data better and to produce more reliable estimates of d_N/d_S than nucleotide models (Goldman and Yang, 1994). Within the codon-substitution model, d_N/d_S is a parameter which is estimated along with other parameters by the maximum likelihood method. Similar to pairwise methods the numbers of synonymous and nonsynonymous sites and the numbers of synonymous and nonsynonymous differences are calculated for each codon. Ancestral reconstruction allows these values to be computed across the entire tree generated from a multiple alignment instead of simply for two sequences as in pairwise methods. For any given codon site, $d_N = d_S$ under the null hypothesis of neutrality. Each codon site in the alignment is taken in turn and positive selection inferred when the null hypothesis can be rejected ($d_N > d_S$).

Initial codon-substitution models were developed by Goldman and Yang (1994) and another, a slightly simpler version, by Muse and Gaut (1994). Later, a maximum parsimony method (Fitch, 1971) that required a phylogenetic tree to initially infer the ancestral codon for every node in the tree was developed.

The Goldman-Yang model (1994) is described below and modified versions used in the analyses in this study. The model specifies the probability that codon *i* changes to codon *j* during evolution along the segment of the tree of length *t* (in time units). A first-order Markov model, which assumes that the state at time *t* depends only on the previous state at time *t*-1, is used to model the substitutions within a codon. There are 61 states in the Markov model which correspond to the 61 sense codons in the Universal genetic code (stop codons are not considered). Each codon has a maximum of nine neighbours to which it may change to instantaneously (Figure 1.3).

Codon substitution probabilities

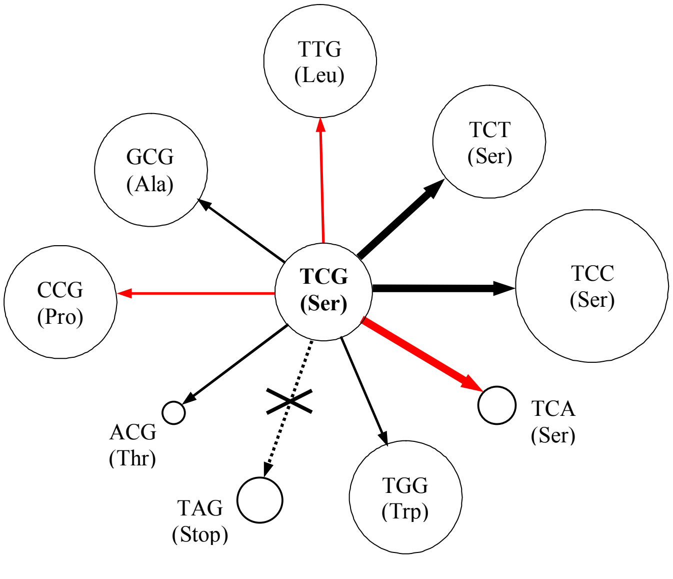
From the substitution model, the instantaneous substitution rate from codon *i* to codon *j* ($i \neq j$) is q_{ij} and can be specified as follows:

$$q_{ij} = \begin{cases} 0, & \text{if } i \text{ and } j \text{ differ at two or three codon positions,} \\ \pi_j, & \text{if } i \text{ and } j \text{ differ by a synonymous transversion,} \\ \kappa\pi_j, & \text{if } i \text{ and } j \text{ differ by a synonymous transition,} \\ \omega\pi_j, & \text{if } i \text{ and } j \text{ differ by a nonsynonymous transversion,} \\ \omega\kappa\pi_j, & \text{if } i \text{ and } j \text{ differ by a nonsynonymous transition.} \end{cases}$$

The substitution rate is proportional to the equilibrium frequency (π_j) of the codon being changed to (codon *j*), calculated using the observed frequencies in the data, which allows codon usage information to be incorporated into the analyses. Alternatively the equilibrium frequency of the codon can be calculated using the observed nucleotide frequencies at the three codon positions. Substitutions that are transitions are multiplied by κ, the transition/transversion rate ratio, and substitutions that result in a nonsynonymous substitution are multiplied by ω.

Figure 1.3 A codon's neighbours

Example of other codons a codon (TCG) may instantaneously evolve into through a single nucleotide substitution. Black arrows represent transversions and red arrows represent transitions. Substitutions that result in no change in amino acid are marked with thicker arrows. Circle size represents the frequency (equilibrium) of that codon (in pooled α- and β-globin sequences). Adapted from Goldman and Yang, 1994.



$Q = \{q_{ij}\}$ is the 61 x 61 rate matrix. The Markov process is reversible so $\pi_i Q_{ij} = \pi_j Q_{ji}$. The probability $p_{ij}(t)$ that any codon i will become codon j after time t can be calculated from q_{ij}. A standard numerical algorithm is used to obtain the eigenvalues and eigenvectors of the rate matrix $Q = \{q_{ij}\}$, to calculate the transition probability matrix for a branch of length t:

$$P(t) = \{p_{ij}(t)\} = e^{Qt}$$

When the parameter ω is different among branches in the tree (see Section 1.5.1), it is necessary to perform this calculation for each branch with a different ω.

Advantages and disadvantages of the model

The model assumes substitutions at the three positions occur independently, that is, only single-nucleotide substitutions can occur during an infinitely small time interval Δt. The model ignores related changes that occur in more than one position and also multiple substitutions in a position as these would occur in time $(\Delta t)^2$. The calculation of the transition probability from one codon to another eliminates the need to explicitly weight evolutionary pathways between codons. All of these factors enable the use of more realistic models in the estimation process. Since sites which contain gaps are ignored by the model, one of the disadvantages of this model is that it does not incorporate any processes for insertions or deletions.

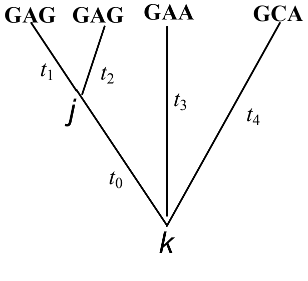
Averaging over all possible ancestral sequences

The model assumes that since the ancestral sequence separated into two descendent sequences when the species diverged, each sequence evolved independently of the other sequence, and each codon evolved according to the Markov model independent of other codons in the sequence. Since the ancestral sequence is unknown, the likelihood method averages over all possible ancestral sequences at each interior node in the tree. The probability of observing the two descendent codons at a site is given by summing over all ancestral codons in the common ancestor. First the probability at each site is calculated as described below. Then the probabilities are multiplied together to obtain the likelihood of all the sites in the alignment.

Calculating probability at each site

Given a set of aligned DNA sequences we can calculate the probability at each site separately. The process calculates the probability of each possible reconstruction, depending on assumptions made about the process of nucleotide substitution, branch lengths, rate of substitution and evolutionary time. For example, given a tree of four species (Figure 1.4), we have to sum over all 61x61 combinations of ancestral states (*j* and *k*) at the two internal nodes, to calculate the probability of data at each site.

Figure 1.4 An example of one site in an unrooted 4-taxon tree



So the probability of observing data at site *i* (with codons GAG, GAG, GAA, GCA) is:

$$\begin{aligned} p_i &= \mathrm{prob}(\{GAG, GAG, GAA, GCA\} \mid t_0, t_1, t_2, t_3, t_4) \\ &= \sum_k \sum_j \left[\pi_k p_{kj}(t_0) p_{jGAG}(t_1) p_{jGAG}(t_2) p_{kGAA}(t_3) p_{kGCA}(t_4) \right] \end{aligned}$$

This averages over all possible values of the ancestral codons, *j* and *k*. The probability of observing data at site *i* is equal to the probability that the codon at the root is *k*, which is given by the equilibrium frequency π_k, multiplied by the five transition probabilities along the five branches of the phylogeny.

This calculation can be computationally intensive as an unrooted tree of *n* species has (*n* – 2) ancestral nodes, so the probability at each site will be a sum over $61^{(n-2)}$ possible ancestral reconstructions. The computation can be speeded up by the use of Felsenstein’s “pruning” algorithm (Felsenstein, 1981), by which conditional probabilities are calculated in a bottom-up manner. The probabilities for tips and daughter nodes are calculated before ancestral nodes with the probability for the root calculated last, hence reducing computation time.

Likelihood of all sites in alignment

The likelihood *L*, for the given set of sequences of length *n* codons, that is, the probability of the data across all the sites in the alignment, is the product of the probabilities across all the sites:

$$L = p_1 \times p_2 \times \ldots \times p_i \times \ldots \times p_n = \prod_{i=1}^{n} p_i \ .$$

As opposed to pairwise methods, the sequences from all the given species can be considered simultaneously. The values for the parameters (branch lengths in the tree, t, κ, and ω) are estimated such that the likelihood value is maximised, using a numerical iteration algorithm. For ease of use, the log-likelihood (the log of the likelihood value) is calculated and is a sum over all sites:

$$\ell = \log(L) = \log(p_1) + \log(p_2) + \ldots \log(p_{i)} + \ldots + \log(p_n) = \sum_{i=1}^{n} \log(p_i) \, .$$

Likelihood-ratio tests

The log-likelihood value can be used as an optimality criterion to evaluate different models of evolution. The model with the highest likelihood value represents the model that best fits the data. Two models can be compared using a likelihood ratio test (LRT). However, both models must be nested with respect to each other, that is, one model must be a simpler form (H_0, the null hypothesis) of the other model (H_1, the alternate hypothesis).

The test statistic from the LRT is $2\Delta\ell = 2(\ell_1 - \ell_0)$ where ℓ_1 is the log-likelihood from the complex alternative model H_1 with p parameters and ℓ_0 is the log-likelihood from the simpler null model H_0, with q parameters. If H_0 is true, then twice the log likelihood difference, $2\Delta\ell = 2(\ell_1 - \ell_0)$, is approximately χ^2 distributed with degree of freedom $= p - q$. If the observed value of the test statistic $2\Delta\ell$ is greater than the χ^2 critical value, we reject H_0 and accept H_1.

Computer simulations suggest that the maximum likelihood method has good power and accuracy in detecting positive selection over a wide range of parameter values (Wong et al., 2004). The type I error occurs if H_0 is rejected when it is true. A test is accurate if the type I error rate is less than the significance threshold chosen for the test, α. The type II error of a test occurs if the test fails to reject H_0 when it is false. The power of the LRT is defined as

1 – type II error rate and is equal to the probability of rejecting H_0 given that H_0 is wrong and H_1 is correct. Using computer simulations, Anisimova et al., (2001) showed that the power of the LRT increases with sequence length, sequence divergence and the strength of positive selection.

1.5 MODELS IMPLEMENTED WITHIN THE MAXIMUM LIKELIHOOD FRAMEWORK

The basic Goldman-Yang codon-substitution model (1994) for likelihood analysis can be modified to account for different levels of heterogeneity in the d_N/d_S ratio among lineages and among sites. The simplest model assumes the same ω for all branches in the phylogeny (model M0). The most general model assumes an independent ω for each branch in the phylogeny and is referred to as the "free-ratio" model (Yang, 1998). This model has as many ω parameters as the number of branches in the tree.

1.5.1 Models of variable selective pressures among branches: branch models

The first model assumed ω to be the same for all branches in the phylogeny. However for some genes, the evolutionary rate may be expected to be higher at specific points in the lineage i.e. along a particular branch. This branch is labelled as the foreground branch. Models that can accommodate a different ω value for a pre-specified branch were first illustrated by an analysis of the enzyme lysozyme (Yang, 1998).

Lysozyme is a bacteriolytic enzyme that can cleave the glycosidic bonds in the cell wall peptidoglycans of bacteria. Due to its lytic activity, the enzyme is part of the antibacterial defence mechanisms of many animals; it is found primarily in the tears and saliva of mammals and in the eggs of birds. In foregut fermenting animals, where ingested plant material is subject to bacterial fermentation, lysozyme is also secreted in the digestive system, permitting the retrieval of the nutrients from lysed bacterial cells. The ruminant artiodactyls (e.g. cow, deer) and the leaf-eating colobine monkeys (e.g. langur) have independently recruited lysozyme as a means of digesting bacteria. The difference between saliva and gastric lysozymes is that gastric lysozymes are active at low pH and are unusually resistant to cleavage by pepsin. In most mammals the omega values for lysozyme will be similar, other than in the artiodactyl ruminants, leaf-eating monkeys and a leaf-eating bird, where lysozyme is thought to have gone through positive selection followed by increased purifying selection.

Based on the phylogeny given, and from the above biological knowledge, we can formulate hypotheses that can be tested using maximum likelihood methods. Previous molecular work had demonstrated that the branch ancestral to the colobines and the branch ancestral to the hominids might be under positive selection or relaxed purifying selection (Messier and Stewart, 1997).

In the analysis by Yang (1998), the null hypothesis assumed the evolutionary rate (ω) of the branches of interest is equal to that of the background branches. Two alternative hypotheses were then formulated: one which assumed that ω was the same for all branches except the branch leading to the colobines; and the other which assumed ω was different only for the hominid branch. Both alternative hypotheses were tested against the null hypothesis in LRTs with 1 degree of freedom. LRTs to examine whether the foreground branch ω (branch leading to colobines or the hominid branch) was greater than 1 were also performed.

Maximum likelihood analyses found the background ω ratio was approximately 0.57 indicating negative selection for lysozyme during primate evolution. The LRTs resulted in the inferred omega of the lineage leading to the hominids to be significantly greater than 1, with approximately 9 nonsynonymous substitutions and 0 synonymous substitutions occurring along this branch. The lack of synonymous substitutions results in the value of ω being infinity. The ω of the branch leading to the colobines was significantly greater than the background ω but the second LRT resulted in the inferred omega to not be significantly greater than 1, with 9 nonsynonymous substitutions and 1 synonymous substitution having occurred along this branch. Therefore, it was concluded that lysozymes have evolved under positive selection possibly as an adaptation to the ruminant diet.

1.5.2 Models of variable selective pressures among sites: site models

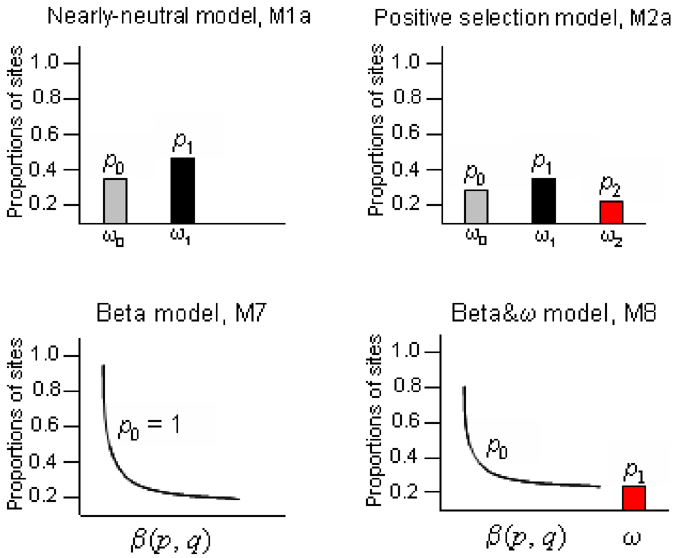
The codon-based maximum likelihood model can also allow for categories of sites to evolve with different values of ω. This is a good model to detect adaptive evolution that affects only a few amino acids in functionally distinct regions of the gene. In almost all proteins where positive selection has been shown to operate, only a few amino acid sites were found to be responsible (Hughes and Nei, 1992). Averaging the estimates for ω across the entire sequence may result in values less than 1, therefore failing to detect positive selection. Site models have higher power if positive selection had occurred over a long time period.

Site models allow different proportions of sites to be under purifying selection ($\omega < 1$), neutral evolution ($\omega = 1$) or positive selection ($\omega > 1$). The neutral (M1) and selection (M2) models were described by Nielsen and Yang (1998), using the HIV-1 *env* gene as an example. These were later superseded by the nearly-neutral (M1a) and positive-selection models (M2a) (Yang et al., 2000). The nearly-neutral model incorporates two categories of sites (Table 1.1). The first category of sites is of proportion p_0 with $0 < \omega_0 < 1$, and the second of proportion p_1 ($p_1 = 1 - p_0$) with $\omega_1 = 1$ (Figure 1.5). The positive-selection model M2a has an extra category of sites of proportion p_2 ($p_2 = 1 - p_0 - p_1$). The inferred value of ω (ω_2) for this category must be greater than 1 for positive selection to be inferred.

Table 1.1 Models of ω ratio variation among sites used for analysis

Model	Parameters	Number of free parameters	Free parameters
M1a (neutral)	$p_{0,} p_1(= 1 - p_0)$ $0 < \omega_0 < 1, \omega_1 = 1$	2	$p_0, \ \omega_0 < 1$
M2a (positive selection)	$p_0, p_1, p_2 (= 1 - p_0 - p_1)$ $\omega_0 < 1, \omega_1 = 1, \ \omega_2 > 1$	4	$p_0, p_1,$ $\omega_0 < 1, \omega_2 > 1$
M7 (beta)	p, q	2	p, q
M8 (beta&ω)	$p_0 (p_1 = 1 - p_0)$ $p, q, \omega_s > 1$	4	$p_0, p, q, \omega_s > 1$

Figure 1.5 Examples of nested site models used in likelihood ratio tests for detecting positive selection



Two other models, the M7 (beta) and M8 (beta&ω) models use the beta distribution to accommodate the shape of the ω distribution that is likely to occur in real data (Yang et al., 2000). The null model M7 (beta) assumes a beta distribution for ω in the interval (0, 1). The alternative M8 (beta&ω) model adds an extra class of sites under positive selection with $\omega_s > 1$ (Figure 1.5).

If positive selection is detected, a Bayes empirical Bayes procedure (BEB) (Yang et al., 2005) is used to calculate the posterior probability of a site belonging to a particular category. The BEB method replaced the naïve empirical Bayes (NEB) method used in earlier models as NEB failed to account for sampling errors in the maximum likelihood estimates of the model parameters, such as the proportions and estimates of ω for the site classes. The BEB method resolves sampling errors by assigning a prior to the model parameters and integrating over their uncertainties.

The human major histocompatibility complex (MHC) class 1 molecules is a good illustration of genes that have several sites that are extremely polymorphic within the antigen recognition site (ARS), whilst the immunoglobin domain is subject to purifying selection. It is advantageous for a population exposed to an array of pathogens to be polymorphic at the MHC loci, because a heterozygote will be able to detect a broader array of antigens, and thus resist a broader array of pathogens. Fixed-site models using *a priori* information to partition the sites in the MHC into two classes, those in the ARS and those outside, have been used to detect position selection. An analysis of 192 alleles of the A, B and C loci of human class I MHC demonstrated sites within the ARS were under positive selection, with an ω of 1.9, whereas non-ARS sites were under purifying selection with an ω of 0.23 (Yang and Swanson, 2002). The sites were scattered among the primary sequence but are clustered together at the ARS in the crystal structure of the protein. Many other convincing examples of positive selection have been detected in a variety of organisms and functional classes (Yang and Bielawski, 2000).

1.5.3 Models of variable selective pressures among branches and sites: branch-site models

The branch-site models are a composite of the branch and site models in that they allow ω to vary both among sites and among lineages (Yang and Nielsen, 2002). With this approach, positive selection can be identified for only a few sites in the protein along pre-specified lineages, which is likely to be a more effective method to detect positive selection as the evolutionary rate will have varied at different sites and at specific times in the gene's evolution.

After reports of false positives (Zhang, 2004), models were developed that were more robust against violations of model assumptions (Zhang et al., 2005). The alternative branch-site model (also known as Model A) has four codon site categories. The first two classes are for sites evolving under purifying selection and neutral evolution on all the lineages and the additional two allow for sites under positive selection on the foreground branch and either purifying selection or neutral evolution on the background branches (Table 1.2). The null model restricts sites on the foreground lineage to be undergoing neutral evolution. The branch-site model has been shown to be very conservative with a low false positive rate and more sensitivity than a lineage model (Zhang et al., 2005). The branch-site, site and branch models have been implemented in the PAML package (Yang, 1997; 2007).

Table 1.2 Parameters in the branch-site model A

		Branch-site alternative model		**Branch-site null model**	
Site class	**Proportion**	**Background**	**Foreground**	**Background**	**Foreground**
0	p_0	$0 < \omega_0 < 1$	$0 < \omega_0 < 1$	$0 < \omega_0 < 1$	$0 < \omega_0 < 1$
1	p_1	$\omega_1 = 1$	$\omega_1 = 1$	$\omega_1 = 1$	$\omega_1 = 1$
2a	$(1-p_0-p_1)p_0$ $/(p_0+p_1)$	$0 < \omega_0 < 1$	$\omega_2 \geq 1$	$0 < \omega_0 < 1$	$\omega_2 = 1$
2b	$(1-p_0-p_1)p_1$ $/(p_0+p_1)$	$\omega_1 = 1$	$\omega_2 \geq 1$	$\omega_1 = 1$	$\omega_2 = 1$

1.6 GENOME SCANS FOR POSITIVE SELECTION

The developments of the branch, site and branch-site models have facilitated the evolutionary analyses of many genes in numerous species. Genome-wide analyses to look for positive selection can provide great insights into the underlying factors that contribute to biological differences between species. The action of positive selection pressure during orthologue evolution is indicative of divergence of gene function between species (Yang and Bielawski, 2000). Therefore researchers have been interested in comparing human genes with their close mammalian relatives to discover genes that have been subjected to positive selection during mammalian evolution.

Another interesting aspect is the relationship between selective pressures affecting gene evolution and the genes involved in disease processes. Enrichment of positive selection signals in disease genes may be due to adaptive changes in response to the environment of early hominids which are maladaptive in the dramatically different conditions we live in today (Young et al., 2005). Another reason for positively selected disease genes is that perhaps highly derived functions that have developed since the separation of chimpanzees and humans cause higher rates of disease than functions that have been subject to purifying selection for millions of years. This poses questions such as whether humans are more susceptible to psychiatric disease than other animals due to our specialisation for higher cognitive function (Keller and Miller, 2006).

1.6.1 Mammalian genome sequences

The availability of genome sequences have made feasible large scale analyses such as genome scans for positively selected genes. Since the advent of whole genome shotgun sequencing and advances in BAC-based sequencing, the complete sequence for the human genome has become available along with increasing numbers of genomes of model organisms. The availability of sequence information from various organisms makes it possible to compare precise nucleotide differences between genes to infer the number and types of changes that have occurred since they last shared a common ancestor (Miller et al., 2004). The comparative genomics study in this thesis compares human with four other mammalian species, mouse, rat, chimpanzee and dog, as they are common models for human disease studies and as complete, good quality genome sequences were available for these species.

The more similar the genomes under comparison (for example, human and chimpanzee are thought to have separated 6 million years ago), the more fitting they are for finding major sequence differences that could account for differences between species (Hardison, 2003). Only homologous genes can be compared in this manner. In the four species selected for this study, over 80% of the protein-coding genes have clear 1:1 orthologues in human (Table 1.3).

Table 1.3 Numbers of genes and orthologues of human present in the species analysed in this study

Genome	Known and novel protein-coding genes*	Number of 1:1 orthologues of human*
Human	22740	-
Chimp	20543	18133
Rat	22503	13912
Mouse	23493	15048
Dog	19305	14700

Note: *Data obtained from Ensembl (May 2008)

1.6.2 Detection of adaptive evolution: a historical perspective

Even prior to the availability of whole genome data, biologists have been interested in comparing the evolutionary parameters of protein-coding genes. A very early study by Makalowski and Boguski (1998) defined 1880 unique human/rodent orthologue pairs using a phylogenetic approach. They found a strong relationship between substitution rates and the coding status of DNA, showing that non-coding sequences evolve approximately five times faster than coding sequences.

Since the availability of large scale genomic data many studies of evolution have been conducted. Here I provide a summary of some of the key studies in chronological order. Following the publication of the human and mouse genomes, Clark et al. (2003) conducted an in-depth analysis of 7645 orthologues from human, chimpanzee and mouse. The authors categorized genes and pathways along the human lineage which had undergone positive selection. More importantly, when differences existed between human and chimpanzee, comparison with the mouse sequence allowed the inference of the primate ancestral state. Using maximum likelihood methods they tested their gene set with the branch-site models (Yang and Nielsen, 2002) and detected 125 human genes evolving with $\omega > 1$ ($p < 0.01$).

Certain functional classes of proteins had significant evidence of positive selection along the human lineage, including olfactory receptors and associated genes and genes involved in processes such as amino acid catabolism, developmental processes, reproduction, neurogenesis and hearing. The excess of positively selected genes (PSGs) in such functional categories are thought to be the result of the differing dietary habits and skeletal development of humans compared with chimpanzees. Genes associated with speech such as the FOXP2 transcription factor, and hearing development were found to be subject to positive selection, consistent with the fact that speech is a human-specific characteristic (White et al., 2006). Interestingly, they also found an over-representation of PSGs within the Online Mendelian Inheritance in Man (OMIM), a repository of genes associated with human disease (www.ncbi.nlm.nih.gov/omim). One possible conclusion from this result is that genomic differences in humans (in comparison with our primate relatives) are associated with human specific diseases.

To investigate the hypothesis that there may be a relationship between disease genes and human diseases, Smith and Eyre-Walker (2003) compared 387 human-rodent alignments for disease genes (defined as genes with mutations known to cause human disease) with 2,024 human-rodent alignments for non-disease genes. They found that higher d_N/d_S ratios were detected for disease genes ($p < 0.001$) suggesting disease genes are either under weaker purifying selection compared to non-disease genes, or are often subject to positive selection.

Huang et al. (2004) also studied human-rodent alignments but found d_N/d_S ratios for their 844 disease genes were only modestly higher than those of non-disease genes ($p = 0.035$). They also found that genes implicated in different diseases had different evolutionary characteristics; in particular, genes involved in neurological disease were well conserved between primates and tend to exhibit low d_N/d_S ratios. However as this analysis was performed using pairwise counting methods and d_N/d_S ratios were estimated across the entire length of the coding sequence, it is possible that any positive selection on a subset of residues within the gene would not be detected as the signal would be swamped by the purifying selection acting on the majority of residues.

Another study by Bustamante et al. (2005) also found that human genes subject to positive Darwinian selection, as a result of the McDonald-Kreitman test (McDonald and Kreitman, 1991), were over-represented in OMIM. Their study tested 11,624 loci and found 304 to be positively selected, many of which were involved in apoptosis, gametogenesis and immunity-defence.

The availability of the chimpanzee genome sequence in 2005 (CSAC, 2005) finally allowed DNA sequence comparisons with our closest relative. The analysis by the Chimpanzee Sequencing and Analysis Consortium (CSAC) of 7,043 quartet orthologues found no significant difference in the mutation rates of disease genes in the human and chimpanzee lineages. Their functional analysis using GO categories revealed a wide range of processes including intracellular signaling, metabolism, neurogenesis and synaptic transmission, which were under strong purifying selection. They also found that the omega values of each of the GO categories that showed an over-representation of adaptive genes were highly correlated between hominid and murid orthologue pairs, suggesting that the positive selection acting on particular functional categories has been largely similar in hominid and murid evolution. The chimpanzee genome sequence (CSAC, 2005) also allowed the estimation of the genome-wide nucleotide divergence between the human and chimpanzee genomes to be 1.23%, with the proportion of divergent sites to be less than 1.06%. This means that a small proportion of sites identified as being under positive selection can potentially be a polymorphic site in either genome.

Nielsen et al. (2005) investigated 8079 human-chimpanzee alignments. The outcome of a LRT using the branch models was 35 genes ($p < 0.05$) for which the null hypothesis ($\omega = 1$) was rejected. This study also used the

PANTHER database (Thomas et al., 2003) to identify functional groups of genes that showed an over-representation of positively selected genes. Again groups that had the most candidates for positive selection were involved in immune-defence and sensory perception. Genes involved in apoptosis and spermatogenesis were also under positive selection, probably due to genomic conflict caused by the natural process of elimination of germ cells by apoptosis during spermatogenesis. Cancer-related genes that function in tumour suppression, apoptosis and cell cycle control also had strong evidence for positive selection. An investigation of gene expression patterns for genes under positive selection found that genes maximally expressed in the brain showed little or no evidence of positive selection. In contrast, genes with maximal expression in the testis were enriched with positively selected genes. They also found that genes on the X chromosome also had an increased tendency to be under positive selection.

In an attempt to differentiate between positive selection and relaxed selective constraint, Arbiza et al. (2006) performed lineage-specific tests on the human, chimp and hominid branches of the phylogeny. They compared 9,674 human, chimpanzee, mouse, rat and dog orthologues using the branch-site methods (Zhang et al., 2005). Using both Test 1 and Test II, they could distinguish between cases of positive selection as opposed to cases of relaxed selection constraint. The more stringent Test II generated 108 and 577 PSGs in the human and chimpanzee lineages respectively. These numbers are after correction for multiple testing, which was employed for all comparisons unlike other studies which only applied correction in some cases. Interestingly, the same sets of biological functions were over-represented by human and chimpanzee

PSGs but not due to an overlap of genes. GO terms such as G-protein coupled receptor (GPCR), sensory perception, electron transport, integrin-mediated signalling pathway and inflammatory response were augmented by human PSGs. Genes that were exclusively in Test I and not in Test II are likely cases of relaxed selective constraint (122 in human, 245 in chimpanzee and 287 hominid genes). Again G-protein coupled receptors were increased in representation in both the human and chimp lineages, which suggests that the process of relaxed selective constraint in G-protein coupled receptors occurred in both species.

From the above studies, it seems that some functional categories are consistently found enriched for positive selection. However, an analysis of 10,376 human-chimpanzee-rhesus alignments by the Rhesus Macaque Genome Sequencing Consortium (Gibbs et al., 2007) found new categories such as iron ion binding and oxidoreductase activity which are encoded by keratin proteins to be enriched among human PSGs. These genes were proposed to have come under selection due to climate change or mate selection. The finding of new functional classes enhanced for PSGs perhaps indicates that the use of more primate species has the potential to uncover human specific neo-functionalisation in genes.

With this is mind, more recently, Bakewell et al. (2007) used the macaque sequence to root 13,888 human and chimpanzee orthologous pairs and investigated the evolution of disease genes since the separation of humans and chimpanzees. They found 9.7% of genes that are positively selected on the human lineage are represented in OMIM compared with 6.1% for the chimpanzee lineage. Therefore there seems to be evidence that disease-causing genes have been prone to positive selection pressure during human evolution.

1.6.3 Functional classification of positively selected genes

Genome scans for selection pressure have attempted to identify molecular functions that are enriched for positively selected genes using the public domain Gene Ontology (Gene Ontology Consortium, 2008) or the PANTHER ontology (Thomas et al., 2003). The classes highlighted by seven recent reports are summarised in Holbrook and Sanseau (2007). Although all the studies used slightly different comparisons and methodologies, it can be seen that there is some consensus among the broad ontological categories identified as enriched for positively selected genes in human evolution, namely: defence/immunity, signal transduction, reproduction, apoptosis, nucleotide metabolism, sensory perception, transcription, subcellular transport, cellular structure, metabolism and development. It is not surprising that genes involved in immune defence are repeatedly identified as evolving by positive selection as the speed at which pathogens evolve has resulted in a co-evolutionary arms race between host cells and pathogens (e.g. MHC molecules).

All these studies have potentially high false positive error rate associated with the results, due to the difficulties of identifying orthologues and aligning sequences in a high throughput manner as well as the reliability of the positive selection detection methodologies. Also, analyses using genomic sequence encounter extra methodological problems, as predicting open reading frames is performed in an automated manner. Errors in any of these processes can impair the accurate detection of positive selection pressure. Despite these computational difficulties it is remarkable that most genome scans have found similar functional categories enriched for adaptive human genes.

1.6.4 Studies of functionally related genes

Some analyses for positive selection focused on specific groups of genes. One such study was by Dorus et al. (2004) who selected 214 genes to cover nervous system biology as broadly as possible. On average these genes had substantially higher d_N/d_S ratios in primates than rodents ($p < 0.0001$) suggesting adaptive evolution in primates. Sub-classification of these genes showed that ones involved in nervous system development had a greater d_N/d_S disparity between primates (humans and macaques) and rodents (mice and rats) compared with house-keeping genes.

Yu et al. (2006) performed an analysis on 2633 human genes with maximal expression in the brain, and gave evidence for the rhesus macaque as a better outgroup than mouse in identifying human selection. They identified 47 candidate genes showing strong evidence of positive selection in the human lineage.

Another study on a specific set of genes was performed on genes related to skin (Izagirre et al., 2006). Analyses for positive selection in 81 candidate loci for skin pigmentation using both population and phylogenetic methods found two genes, *MYO7A* and *PGR* as being under positive selection.

1.7 WHY SELECTION PRESSURE MATTERS TO DRUG DISCOVERY

The discovery of new drugs to treat human diseases is a difficult business with very low success rates. One reason for this is the transition from pre-clinical R&D to clinical trials in humans is reliant on successfully translating experimental results in model organisms such as mice, rats, non-human primates and dogs, to humans. Animal models are used during the phases of target selection and validation (Pravenec and Kurtz, 2007), drug efficacy studies (Priest and Kaczorowski, 2007) and drug safety studies (Valentin et al., 2005). One of the major causes of the observed high levels of attrition in R&D pipelines is lack of human efficacy and safety since animal models of efficacy are notoriously unpredictive (Kola and Landis, 2004). The two therapeutic areas with very high attrition rates, oncology and the central nervous system, are also the areas in which animal models are often not predictive of true human pathophysiology. Difficulties lie in determining the different susceptibilities of various diseases in animal models, especially in studies of higher cognitive abilities. For example, some aspects of aging in humans (Alzheimer's disease) do not develop naturally in nonhuman primates or do not follow the same course of natural development in monkeys (menopause), therefore it is necessary to use experimental models of these conditions for study. However, difficulties in the interpretation of animal experiments to predict human drug response can also be caused by biomedical differences between humans and model species in the biology of the drug target or proteins interacting with the drug target or in drug metabolism. On the molecular level, the complete absence of an evolutionary orthologue of the human drug target in a model species is an extreme example of this difficulty (Norgren, 2004; Holbrook and Sanseau, 2007). However a more likely cause is variations in drug target function between species as a result of positive selection acting on the gene in one of the species. Studying the evolutionary history of the genes encoding drug targets could help elucidate species differences prior to choosing animal models for pre-clinical tests and allow better interpretation of experimental results from model species.

1.8 PROJECT AIMS

This thesis describes a comparative genomics project which applies maximum likelihood models of DNA sequence evolution to detect episodic periods of evolution along the human lineage and lineages of model organisms. This information can be used to identify the biological processes that have been subject to adaptive evolution in the species under investigation. I also explore whether genes under positive selection show significant associations with human disease. Detection of selective pressures which indicate functional shifts are also important in the pharmaceutical industry for establishing species differences that affect drug-discovery assays and the choice of animal models.

The study begins with the identification of strict 1:1 orthologues to human in the four mammalian species: chimpanzee, mouse, rat and dog. The resulting 3079 high-quality gene sets were scanned for positive selection signals during mammalian evolution comparing the five species together. In contrast to previous studies which have tended to focus on human evolution, the objective of this study was to determine genes which have undergone adaptive evolution in both humans and animal models. Chapter 2 describes the orthologue identification and alignment pipeline and use of the branch-site model to test all extent and ancestral lineages on the species phylogeny for evidence of positive selection.

Functional classes such as nucleic acid metabolism, neuronal activities, and immunity and defence were found to be the most enriched by primate genes under positive selection, as explained in Chapter 3. I also provide evidence to support the hypothesis that genes under positive selection tend to interact more with each other than other genes.

The chimpanzee lineage was found to have more genes under positive selection than any of the other lineages. In Chapter 4, I show that positive selection in these genes is unique to the chimpanzee lineage, explore the effects of taxon sampling on the detection of positive selection and finally offer some hypotheses for the high number of divergent chimpanzee genes.

Chapter 5 illustrates how genes that have been subject to positive selection pressure during human evolution are implicated in diseases which have uniquely human pathogenic mechanisms. Epithelial cancers, schizophrenia, autoimmune diseases and Alzheimer’s disease are some diseases which differ in incidence or severity between humans and apes (Olson and Varki, 2003; Varki and Altheide, 2005). Biomedical differences between species could be due to functional shifts in gene involved in the molecular mechanisms of the disease and hence can be attributed to positively selected genes.

Further work, covered in Chapter 6, explores selection pressure in transcription factors, such as nuclear receptors. An in-depth analysis of the 48 human nuclear receptors and their mammalian orthologues using the site and branch-site models, demonstrates the variation of selection in functionally distinct regions of these genes.

Chapter 2

Genome Scan Methodology

The processes described in this chapter form the data sources and analyses for results described in Chapters 3, 4 and 5. Additional methods specific to a single chapter are described within that chapter. Methods pertaining to Chapter 6 are at the beginning of that chapter.

The Bioinformatics groups at GSK designed and carried out some of the procedures described in this chapter. Simon Topp, Vinod Kumar, Mike Word and Mark Simmons designed and analysed the data collection processes and orthologue calling pipeline described in Sections 2.1 – 2.3. Samiul Hasan wrote the Perl programs and carried out the data analysis in Sections 2.6 and 2.8. Dilip Rajagoplan designed the co-evolution experiments in Section 2.12.

2.1 SOURCES OF DATA FOR HUMAN AND MODEL SPECIES GENES

The study of positive selection in orthologous genes is wholly dependent on careful collation of true orthologues of human in the species selected for analysis and their alignment. Human genes from NCBI Entrez Gene (accessed in September 2006) (http://www.ncbi.nlm.nih.gov/sites/entrez?db=gene) that had been annotated as protein coding and had an identified DNA transcript were selected for evaluation. The DNA transcripts and corresponding peptides were downloaded from GenBank (http://www.ncbi.nlm.nih.gov/Genbank/), RefSeq (http://www.ncbi.nlm.nih.gov/RefSeq/) and Ensembl (http://www.ensembl.org/) with the exception of RefSeq predicted transcripts and peptides (commonly recognised by accession numbers prefixed by XM and XP), which were excluded from the analysis pipeline due to the potential for poor quality gene prediction. To produce a non-redundant list of peptides, the peptide sequence files for each locus were clustered based on identity and length via the blastclust program (blastclust -i sequence_file -p T -L 1 -b T -S 100) (http://www.ncbi.nlm.nih.gov/). For each cluster, DNA sequences were generated using the longest open reading frame of the longest peptide in each cluster. The transcript sequences were then named MELgeneID_clusterID where geneID was the EntrezGene ID and clusterID was the cluster number from the peptide cluster analysis. The mRNA and genomic sequences for the four model organisms (chimpanzee, mouse, rat and dog) and chicken (outgroup) were extracted from GenBank (accessed September 2006).

2.2 ORTHOLOGUE CALLING PIPELINE

The primary method employed to find orthologues is to establish homology using reciprocal tBlastx searches (Altschul et al., 1990) as this approach is known to be conservative. The premise of using this method is if two sequences within the two genomes are orthologues, the BLAST hit between the two sequences should have a lower *p* value from BLAST than to an outparalogue. Although the reciprocal BLAST hit method to find orthologues is not a phylogenetic method, it is less computationally intensive and is suitable to find 1:1 orthologues. The tBlastx parameters used in the orthologue detection pipeline between the human and model organism sequence databases were B=50 V=50 W=3 E=0.01 topcomboN=2 wordmask=xnu+seg maskextra=6 Z=3000000000 progress=0 warnings cpus=1 matrix=blosum62 Q=30 R=2 and ctxfactor=36.0.

The sequence representing each human gene cluster was used as the seed query to account for alternatively spliced variants. Reciprocal best hits between the starting human gene and model organism gene were marked as the main orthologue pair for that human transcript query. Each cluster was Blast searches of human sequences against species databases that resulted in genomic sequence were processed further. Genewise (Birney et al., 2004) was used to identify a predicted cDNA using the human peptide as a template. The genomic sequences were masked for interspersed repeats and low complexity regions prior to analysis by Genewise, to increase specificity and improve gene predictions. The resulting cDNA sequence was then used as a query in the reciprocal tBlastX search against the human database. Highest scoring mRNA sequences were submitted to the reciprocal tBlastX search without modification.

Finding orthologues by reciprocal BLAST can result in erroneous pairs if genes have been modified by domain-shuffling and other forms of horizontal transfer. tBlastx versus genomic sequence is also susceptible to finding pseudogenes. To make orthology assignment more conservative, a human gene and a model organism gene were marked as the orthologue pair only if the negative log of the *p* value of the best hit of the human sequence against the model organism database was higher than 95% of the negative log of the *p* value of the best hit from the reciprocal step.

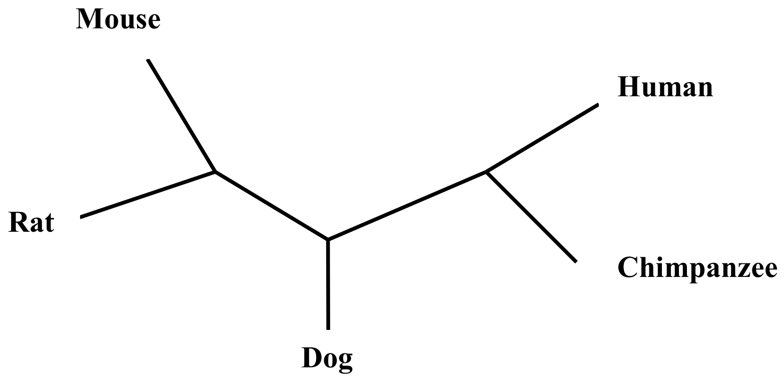
Incomplete genome sequencing will also contribute to error in orthologue calling. Reciprocal blasting is invalidated as a method for calling orthologues in these circumstances, as the absence of the true orthologue would cause a more divergent paralogue to be the top hit. To address this problem, a cut-off was added which required the *p* value of the putative orthologue for that species to be less than that of the chicken orthologue for that gene. The chicken was chosen for two reasons: the complete draft genome sequence was available (International Chicken Genome Sequencing Consortium, 2004) at sufficient coverage, and a non-mammalian species was required to serve as an outgroup for the mammalian orthologue calls. For the 392 human genes for which a chicken orthologue was not found, the model species orthologues were kept and labelled as lower confidence orthologues.

The last three nucleotides were removed for any sequences terminating in a stop codon. All putative orthologue sequences were assessed for the optimal ORF, using the human sequence as a template. The human sequence contained only the CDS and hence was translated from nucleotide one. Each orthologue sequence was translated in the three forward frames. From each translation, every peptide sequence between stop codons (or between stop codons and the sequence termini) was assessed for length. If the sequence was less than 50% of the length of the human protein, the peptide was excluded. Sequences meeting the length cut-off were aligned to the human protein using the EmBoss 'Needle' Needleman-Wunsch (Rice et al., 2000) algorithm. The peptide with the highest-scoring alignment was deemed to be the most appropriate ORF, and the encoding nucleotides were written to the output. If no peptides in any frame met the length criteria, the orthologue sequence was excluded from the output. All sequences resulting from the orthologue calling process were exported as fasta files labelled with the original accession number, the species name, and the human gene used as the starting query.

2.3 Alignment and PAML input files

The resulting nucleotide sequences were aligned using the SIM-based codon-centric algorithm implemented in SwissPDBviewer/Promod (Guex and Peitsch, 1997). A custom version was used that included support for trimming unaligned N- and C-termini, searching for the initial Methionine (within the first 60 residues and conserved in at least 50% of sequences), and alignment scoring based on a combined amino-acid matrix (Blosum70) plus codon identity penalty score, with gap=6 and gap_extension=4. Alignments were read out in PAML format. Unrooted tree files for each alignment were created using a standard mammalian species tree (Murphy et al, 2001) (Figure 2.1).

Figure 2.1 Unrooted five-species tree used in branch-site analyses



2.4 Branch-site analyses

2.4.1 Branch-site model

The branch-site model (Yang and Nielsen, 2002; Zhang et al., 2005) implemented in the codeml program from the PAML package (Yang, 1997) was used to test for positive selection. Each of the seven branches on the species phylogeny was tested, treating each in turn as the foreground branch, with all the other branches specified as background branches. Each branch-site model was run three times. Since the model can have local maxima, the requirement was that at least two of three replicate runs of each model should converge at or within 0.001 of the same log-likelihood value for convergence to be established. Runs that did not converge were indicative of problems with the data and were re-done until convergence was obtained or else reported as a convergence problem.

2.4.2 Optimisation of branch-site model parameters

To determine the values of branch lengths for the subsequent branch-site model run, the M0 model was run on all of the data sets. The M0 model assumes the same d_N/d_S ratio for all branches in the tree and among all codon sites in the gene (Goldman and Yang, 1994). To reduce complexity and computational time, branch lengths estimated by the M0 model were used as fixed values for the branch-site model (fix_blength = 2), as opposed to using them as initial values for the branch-site model (fix_blength = 1) (Yang, 2000). Two runs of the M0 model were performed on each alignment to check that values for log-likelihood, κ and branch lengths were consistent between the two runs. Runs that were not consistent were repeated until the values converged.

281 alignments were used to investigate the effects of fixing branch lengths on the other parameter estimates. The number of genes that had a signal for positive selection ($p < 0.05$) decreased when branch lengths and the transition/transversion ratio (κ) were fixed (Table 2.1). Other parameters such as proportions of sites under different selection categories, background-ω and

foreground-ω remained consistent between the two analyses. The overall run times reduced by two-thirds as fixing the branch lengths reduced the number of parameters to estimate from 12 to 5. Fixing branch lengths therefore makes the branch-site test more conservative and should improve convergence problems.

Table 2.1 Numbers of positively selected genes when branch lengths and *κ* were fixed and free to be estimated

	No of PSGs (free parameters)		**No of PSGs (fixed parameters)**	
	p<0.05	**p<0.01**	**p<0.05**	**p<0.01**
Human	6	3	5	2
Chimpanzee	106	106	107	106
Hominid	3	2	1	0
Mouse	18	15	15	11
Rat	37	30	28	23
Murid	21	10	6	2
Dog	40	33	26	20

2.4.3 Multiple hypothesis testing correction

Likelihood-ratio tests were performed with the Bonferroni correction for multiple testing (Anisimova and Yang, 2007). If *n* hypotheses are tested on a set of data, the Bonferroni correction raises the statistical significance level by 1/*n* times what the significance level would be if only one hypothesis was being tested. Prior to correction, the critical values at *p* values of 0.05, 0.01 and 0.001 and 1 degree of freedom for a chi-square distribution were 3.84, 6.63 and 10.83, respectively. After correction for seven tests, the critical values were raised to 6.63, 10.55 and 13.83.

Perl programs were written to analyse and perform calculations on the 3079 alignments in an automated manner. Descriptions of some of the major scripts can be found in Appendix 1.

2.5 UPDATING ALIGNMENTS

The initial analysis with the branch-site models resulted in a larger number of chimpanzee genes to be detected under positive selection than human (**result set A**) (see Section 3.2.1). Many of the positively selected chimpanzee genes were thought to be false positives. An investigation into the resulting branch lengths and manual inspection of alignments uncovered data problems with many sequences that had been designated as orthologues having single base deletions or substitutions when compared to the 'golden path' genomic sequence data in the UCSC Genome databases (http://genome.ucsc.edu/cgi-bin/hgGateway). Since the inception of this study, several updates to the NCBI genome databases had made the data used in this initial analysis out of date. In addition, a major update to the chimpanzee genome sequence increased the coverage from 4x (September 2005) to 6x (March 2006), subsequently increasing sequence quality as well.

To select which sequences were outdated and were to be replaced, for each gene, the sequences in the current alignments were aligned against the species-specific mRNA RefSeq sequences using fasta34 (fasta34 –q –H –d 1 –b 1 query library > results.txt', where -H = no histogram, -b = best scores, -d = number of best alignments and -q = quiet and library = protein file e.g. chimpRefSeqGenes.txt) (Pearson, 2004). RefSeq sequences were chosen as opposed to Ensembl or UCSC mRNA libraries as they were the most updated sequences at the time. mRNA and protein library sequences were downloaded from ftp://ftp:ncbi.nih.gov/genomes/H_sapiens/rna and

ftp://ftp:ncbi.nih.gov/genomes/H_sapiens/protein

The percentages of genes for each species that matched exactly to its RefSeq counterpart or had 90% identity are shown in Table 2.2.

Table 2.2 Percentages of sequences from each species that matched its RefSeq counterpart

Species	RefSeq library size (no. of sequences)	RefSeq library creation date	Percentage of sequences with 100% identity to RefSeq	Percentage of sequences with greater than 90% identity to RefSeq
human	34180	05/09/2006	82.83	98.96
chimpanzee	51947	25/09/2006	45.65	81.08
dog	33651	01/09/2006	60.37	87.92
mouse	46892	04/05/2006	80.82	98.48
rat	36496	10/07/2006	78.55	94.31

Sequences in the current alignments which had less than 100% identity and greater than 80% identity were replaced by their RefSeq equivalent and new protein alignments were created using Muscle (Edgar, 2004). Alignments that did not have any sequences replaced were also aligned with Muscle (muscle –in seqs.fasta –out seqs.fa.musc) to ensure consistency. Protein alignments were then converted to their corresponding DNA alignments and the ends trimmed to remove single sequence areas. Alignments were checked using a simple alignment scorer which assigns a mismatch score to each sequence in the alignment. The mismatch score is the number of bases in a sequence which do not match to any other base in the alignment slice, expressed as a percentage of the length of the sequence (not the length of the alignment). The higher the mismatch score, the more problematic the sequence is in the alignment. RefSeq substituted sequences that made the mismatch score worse were reverted back to the original sequence.

Chimpanzee and dog sequences which had a 100% match to a RefSeq sequence but still had a high percentage of mismatches in the alignment were updated with new gene predictions using updated versions of their respective genome sequences. 214 chimpanzee and 445 dog sequences were re-predicted by the same method described in Section 2.2.

2.6 FRAMESHIFT CORRECTION

It was noted that some model species sequences had frameshifts when compared to their human orthologue due to missing or additional bases possibly from sequencing errors. Nucleotide codon alignments were scanned for potential frameshift mutations using a frameshift correction script. All alignments were corrected for frameshifts in the sequences from the model organisms relative to human. Starting from the first position in the nucleotide alignment, an insert was placed at every position till the end of the sequence to create the new translated amino acid alignment. If sequence identity in the next 5 amino acid sites (relative to insert) improved over the current alignment, the modified nucleotide sequence replaced the original sequence in the alignment. 400 alignments were corrected by this method.

2.7 G-BLOCKS CORRECTION

The program G-blocks (Castresana, 2000) was used to examine alignments for regions of high divergence. This program masks regions in the alignment that are poorly aligned and returns the remaining alignment. Alignments were scanned and categorised into two classes:

Class A - returns more than ~70% of whole alignment

Class B - returns less than ~70% of whole alignment.

35 alignments were returned in class B. These were manually inspected and when alignment quality was high enough for these alignments to be deemed useable for the branch-site analysis. The branch-site models were re-run on the new alignments to generate **result set B**.

When result set B, the product of the procedures described above, was examined closely some alignments had areas of ambiguous alignment or areas where sequences did not appear orthologous. A number of positively selected genes in all the tested lineages had many consecutive sites being reported by the Bayes Empirical Bayes method as having a high probability of being under positive selection. Upon closer inspection, it seemed that these sites corresponded to regions that were either misaligned or were non-orthologous. Therefore the data were subjected to further manual corrections detailed below.

2.8 LOW SIMILARITY SEQUENCE MASKING

To correct for regions of low similarity, all alignments were scanned to mask out parts of a sequence where more than 3 consecutive codons in that sequence were different to the other sequences in the alignment and these codons were flanked by gaps on one or both sides. These regions were then masked by Ns and the branch-site analysis re-run on the entire dataset.

2.9 MANUAL CURATION OF ALIGNMENTS

After re-running PAML on the entire dataset, the alignments were manually examined for all significant results at $p < 0.05$. A result was discarded if the gene sequence belonging to the lineage that was identified as being under positive selection had a frameshift or was ambiguously aligned. We also manually examined alignments in which the chimpanzee branch length was higher than 1 substitution per site (codon). A result was discarded if the gene sequence belonging to the lineage that was identified as being under positive selection had a frameshift, had many gaps or was misaligned. The set of significant genes remaining was termed **result set C**. Genes under positive selection along the hominid and murid lineages were not manually curated as a positive result in the hominid lineages arises if the human and chimpanzee sequences are similar to each other and different to the other sequences. Similarly, a positive result in the murid lineage arises if the mouse and rat sequences are the same and different to the other sequences. If an alignment had some sections which had good homology and also had sites under positive selection and other that looked misaligned, then the result for this alignment was not discarded because the gene would still come up as being under positive selection even if the misaligned area was corrected.

2.10 PHRED QUALITY VALUES OF CHIMPANZEE POSITIVELY SELECTED GENES

As a further precaution, the quality values of the sequences for chimpanzee genes under positive selection were also checked. The 162 sequences in result set C were aligned against the dataset of chimpanzee chromosome sequences using BLAT to obtain their genome coordinates. Four of these (MEL640, MEL677, MEL9636 and MEL81099) had a few missing bases and did not produce a BLAT result with 100% identity. The alignments of these four genes were manually examined and the missing bases had resulted in incorrect amino acids which then caused significance to be inferred incorrectly. Of the sequences that generated a BLAT result with 100% identity, the quality values corresponding to the genome coordinates were obtained. Only 1 of these (MEL51368) had phred quality values less than 20 among the bases inferred to be under positive selection. The results for these five genes were discarded.

2.11 CALCULATING RATE DIFFERENCES

To calculate ω for each branch of the species phylogeny, the free-ratio model in the codeml program was run on each alignment. The median value was chosen as the representative value for that branch after exclusion of maximal values (999) of ω. The free-ratio model was also run on the concatenated set of alignments, after removal of sites with ambiguity gaps and alignment gaps, which left 81% of the concatenated alignment.

2.12 ANALYSIS OF INTERACTION DATA

2.12.1 Interaction Data

A network consisting of protein-protein interactions such as binding and phosphorylation, transcriptional control and post-translational modification was used to search if genes under positive selection from this study interact together significantly. Interaction data in the network was licensed from several commercial vendors including Ingenuity (www.ingenuity.com), Jubilant (www.jubilantbiosys.com), GeneGO (www.genego.com), NetPro (www.molecularconnections.com) and HPRD (www.hprd.org). Data from these products is obtained from literature-derived information describing interactions of various kinds (binding, regulation, metabolic etc) between pairs of genes. In addition, high-quality, automatically extracted interactions licensed from the PRIME database (Koike and Takagi, 2005) were also included in the network. Interactions associated with transcriptional regulation were obtained from experimentally validated protein-DNA binding relationships licensed from the TransFac (Matys et al., 2003) and TRRD (Kolchanov et al., 2002) databases. No distinction is made between DNA, RNA and protein for a particular gene, and all three are represented as a single node in the network. Integrating data from all these sources that describe direct interactions between genes results in a network with 98, 095 unique interactions (edges) among 14781 genes and metabolites (nodes), of which 1035 are metabolites and the rest are genes.

2.12.2 Biological clustering algorithm

Searches of gene lists that resulted in a biological sub-network were conducted and scored as in Rajagopalan et al. (2005). Gene lists of PSGs from each lineage and the associated *p* value (after Bonferroni correction) was used as the input dataset. A similarity metric of the square of the Pearson correlation coefficient was calculated for all pairs of genes in the input dataset. The list of all pairs of genes is then filtered using 0.81 as a cutoff to generate a set of significant pairs of genes (set *S*). Pairs of genes in set *S* that are not supported by the interaction network are removed from set *S*. If the pairs of genes being considered are neighbours on the interaction network or have one intermediate network node that is a metabolite, the pair is preserved.

A simple greedy search algorithm is then applied to set *S* to form clusters of genes. The process starts with the best remaining pairs of genes, as measured by the correlation coefficient. All genes that are connected to either of these first two genes via an edge in set *S* are added to the cluster. Next genes that are in set *S* that are connected to genes already in the cluster are added to the cluster. The process continues successively until no more genes can be added to the cluster. A new cluster is started based on the best remaining pair in *S.* Clusters are then merged together if two clusters are separated by a node on the interaction network that is not contained in any cluster but the node is adjacent to a node in each of the two clusters. The resulting clusters of genes are PSGs from the lineage being analysed and are also connected together by prior biological knowledge.

Chapter 3

Genome Scan Patterns of Positive Selection

3.1 INTRODUCTION

Direct comparisons of human genomic and transcriptomic information to that of other species reveal three major types of molecular genetic changes which have contributed to species differences. The most obvious mode is the presence or absence of genes in different species, including gene duplication and gene inactivation. Much attention has been paid to genes that are unique to humans or lost in the human lineage (Olson and Varki, 2003; CSAC, 2005; Varki and Altheide, 2005; Kehrer-Sawatzki and Cooper, 2007). However these probably represent the 'tip of the iceberg' of human genomic differences compared to other species. The second class of molecular genetic changes consists of nucleotide substitutions that may cause functional changes in both protein coding and non-coding RNAs. The third category of molecular changes consists of variations in the levels of gene expression between species and in the mechanisms regulating gene expression (Gilad et al., 2006; Kehrer-Sawatzki and Cooper, 2007).

In this study we investigate the second type of molecular differences and focus on coding changes in protein-coding orthologous genes. An estimated 70% to 80% of orthologous protein sequences are distinct between humans and chimpanzees (CSAC, 2005; Glazko et al., 2005). However, a substantial proportion of differences may have no functional impact on human-specific diseases. Positive selection analyses can determine which nucleotide changes contribute to biological differences between species. This follows from the premise that the action of positive selection pressure in orthologous genes during evolution is indicative of divergence of gene function between species (Yang, 2005). Determining such genes on the human lineage is thus a rational and promising way to reveal the molecular changes implicated in human-specific diseases.

Our initial dataset was aggressively filtered to eliminate paralogous alignments, spurious annotations, pseudogenes in one or more species, and poor exon prediction. Hence only quintets for which we could assign orthology with high confidence were used in our analysis for positive selection. Due to this strict screening it must be noted that our orthologue dataset may contain a bias towards orthologues of high levels of conservation, thereby underestimating the number of positively selected genes and underestimating the average levels of divergence. Results prior to multiple hypothesis correction should not be used for subsequent analysis as the family-wise error rate is unacceptably high (Anisimova and Yang, 2007). Here we report results following a Bonferroni correction for multiple testing which is known to be conservative and hence, prediction of positive selection is particularly robust. The corollary of such a strict approach is the potential generation of false negatives.

As demonstrated in the yeast protein interaction network, evolutionary rate is thought to be negatively correlated with protein connectivity (Fraser et al., 2002; Fraser et al., 2003; Fraser and Hirsh, 2004). Hence, genes under positive selection are generally believed to be less promiscuous, that is, they interact with fewer genes compared to genes under neutral evolution or purifying selection. This may be because promiscuous genes are subject to functional constraints due to their pivotal or multiple roles in biological pathways. However, others analysing the same data claim that the results are inconclusive (Bloom and Adami, 2003; Jordan et al., 2003). We investigate whether genes under adaptive evolution interact with fewer genes compared to genes not under positive selection but did not see a significant difference between the two gene groups.

We also investigated the hypothesis that a gene under adaptive evolution would drive complementary divergence of genes encoding interacting proteins. Many studies on the co-evolution of individual genes and the genes they interact with have been published. The most common examples are receptor-ligand couples that co-evolve to maintain or improve binding affinity and/or specificity. Examples of such genes include the prolactin (PRL) gene and its receptor (prolactin receptor, PRLR) in mammals (Li et al., 2005), primate killer cell immunoglobulin-like receptors (KIRs) that co-evolved with MHC class I molecules (Hao and Nei, 2005) and red and green visual pigment genes (Deeb et al., 1994). However, this phenomenon has never been investigated among mammalian genes on a genome-wide level. Here we present evidence that positively selected genes are significantly more likely to interact with other positively selected genes than genes evolving under neutral evolution or purifying selection.

3.2 RESULTS

3.2.1 Numbers of positively selected genes in result set A

The five species orthologue identification procedure and alignment pipeline resulted in 3079 orthologue alignments corresponding to 16% of the human genome. 2689 of the genes also had a chicken homologue (representing genes conserved at least in the chordata) and 390 genes did not have a chicken homologue (representing potential mammalian specific genes or genes that were missing in the incomplete chicken genome sequence).

The branch-site model in the codeml program of PAML (Zhang et al., 2005) was used to investigate the evolutionary rate of each gene on each branch of the species phylogeny. This set of genes under positive selection that resulted from the initial analysis was termed result set A. This set showed an unusually high number of positive genes in the chimpanzee lineage (Table 3.1). Table 3.1 includes the numbers of positively selected genes (PSGs) in result sets B and C which are discussed in turn below. Result set B is a subset of result set A and similarly, result set C is a subset of result sets A and B.

Table 3.1 Numbers of genes detected to be under positive selection by the branch-site model

	Human	Chimpanzee	Hominid	Mouse	Rat	Murid	Dog
Result set A (data from pipeline)							
$p < 0.05$	111	814	41	145	229	48	232
$p < 0.01$	76	756	18	110	191	16	177
Result set B (before data curation)							
$p < 0.05$	69	354	49	121	155	86	162
$p < 0.01$	46	325	24	94	126	41	127
Result set C (after data curation)							
$p < 0.05$	54	162	56	65	89	81	97
$p < 0.01$	32	137	56	47	64	81	62

A close examination of the alignments which resulted in significant results reveals many alignments with high branch lengths in the foreground lineage. The number of genes which had a branch length greater than 1 (measured as the number of nucleotide substitutions per codon) and were not significant for positive selection were compared to PSGs with branch lengths greater than 1 (Table 3.2). The branch length value of 1 substitution per codon was chosen arbitrarily as a gene having a branch length of 1 is considered quite high. There does not seem to be a correlation between the large number of positive genes in the chimpanzee lineage and high branch length as there were many genes which were not significant for positive selection that also had high branch lengths. Whilst long branch lengths may mean accelerated evolutionary rates, excessively long branches can also indicate alignment problems. These alignment problems could be the cause of the high number of significant results that ensued. Errors in alignment could have resulted from errors in sequencing or incorrect regions in gene predictions.

Table 3.2 Number of genes in each lineage with branch length greater than 1

	Human	Chimpanzee	Hominid	Mouse	Rat	Murid	Dog
No. of genes	1	41	5	24	62	27	69
No. of PSGs	1	22	0	3	16	5	23

The set of overall branch lengths was estimated by running the M0 model of the concatenated set of 3079 alignments. The resulting tree was:

((Human: 0.017784, Chimpanzee: 0.072954): 0.169006, (Mouse: 0.130561, Rat: 0.176092): 0.339659, Dog: 0.287177).

The total tree length was 1.19323. It can be seen that the chimpanzee branch is approximately four times longer than the human branch length, implying a four-fold acceleration of substitution rates in the chimpanzee lineage compared to the human lineage since they last shared a common ancestor. This is unusual as we would expect the distances to be somewhat similar for human and chimpanzee.

The M0 model was run again after removing alignments for which the chimpanzee branch length was greater than 1.0. The branch length of the chimpanzee branch decreased but only marginally. Hence the inclusion of more divergent chimpanzee genes does not contribute to overall branch length of the chimpanzee branch length.

Manual examination of some of the alignments showed that the current data problems were caused by sequencing errors of single base deletions or substitutions in many of the orthologous sequences when compared to the ‘golden path’ genomic sequence data in the UCSC Genome databases (http://genome.ucsc.edu/). Recent data updates to many of the genome sequences have improved the quality and coverage of the sequence, in particular the chimpanzee genome sequence, for which the coverage had increased from 4x to 6x. The low quality of some of the sequences in the current alignments could be the cause of false-positive results from the branch-site analysis.

3.2.2 Number of positively selected genes after data curation: result set B

Low quality sequences in the original alignments were replaced with NCBI RefSeq sequences and new chimpanzee and dog sequences from gene predictions on updated genome sequences. Automated frameshift correction was performed on the alignments and the branch-site analysis was performed again. The ensuing set of PSGs was named result set B. The analyses resulted in 69 PSGs ($p < 0.05$) along the human lineage (Table 3.1). The number of PSGs in the other lineages varied with the highest seen in the dog (162) and chimpanzee (354) lineages.

The consequence of employing frameshift correction for genes in result set B is that although frameshift errors due to alignment are corrected, frameshifts that exist in nature would have been corrected as well. To check if true frameshifts had been corrected we compared our set of PSGs in result set B with the set of genes that had undergone pseudogenisation along the human lineage (Wang et al., 2006). We did not find any overlaps, suggesting that our frameshift correction has only rectified sequence errors and not modified genuine species differences.

Alignments were again examined to look for areas of high divergence. Some alignments had areas of ambiguous alignment or areas where sequences did not appear orthologous. Areas of non-orthology could result from incomplete gene predictions due to gaps in the genomic sequence or absent or variant exons.

3.2.3 Number of genes under positive selection after manual curation: result set C

The alignments were subjected to masking of regions of low homology (see Section 2.8). A portion of the following numbers of sequences was masked in each species: chimpanzee 1209, dog 1474, rat 1294, mouse 1169 and human 835. The branch-site analyses were re-run and all alignments which resulted in a significant result in one or more lineages were manually inspected for alignment errors. Results were discarded if sequences were misaligned or if sequences did not appear orthologous. The resulting set of PSGs was named result set C.

In this set we found 1222 genes to have evolved under positive selection in at least one of the seven mammalian lineages (1707 genes from all lineages as some genes were significant in more than one lineage) but after data curation, we only considered 511 genes (604 genes from all lineages) for further analysis. Following Bonferroni corrections, all lineages tested showed significant ($p < 0.05$) evidence of genes evolving under positive selection varying from 54 genes along the human lineage to 162 along the chimpanzee lineage (Table 3.1). The rat lineage also showed a slightly higher number of positive genes (89) compared to the mouse lineage (65). In comparison with result set B, the number of PSGs in result set C in all lineages except in hominids decreased. The number of PSGs along the hominid lineage increased from 49 to 56 after data correction because data curation increased the identity of human-chimpanzee sequences, resulting in an increase in the number of PSGs in the hominid lineage. A complete list of PSGs that were detected in each lineage is available in Appendix 2.

Subsequent analyses were performed on both result sets B and C. We found that in all cases the findings for the two sets were similar although result set B may contain some false positives resulting from alignment errors.

3.2.4 Overall evolutionary rates

To obtain an overall perspective of the evolutionary rates of the genes in our dataset, the free-ratio model in the codeml program was run on each alignment (see Section 2.11). The median ω values for each lineage ranged from 0.14 in mouse and rat to 0.17 in human and 0.20 in chimpanzee (Figure 3.1).

Figure 3.1 Five species tree with branch-specific d_N/d_S ratios

The median ω value from free-ratio model estimates of evolutionary rates in 3079 genes for humans, chimpanzees, mouse, rat and dog. Branch lengths are proportional to absolute d_N values (Table 3.3).

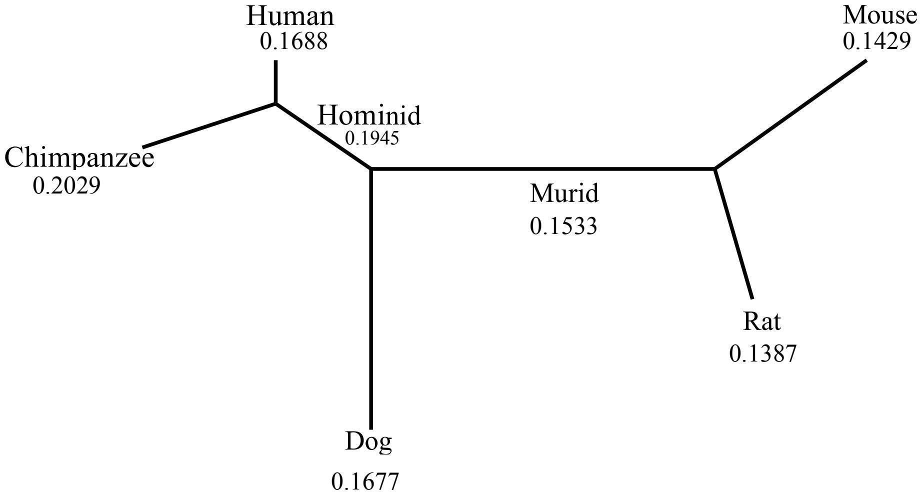


Table 3.3 Estimates of branch lengths, d_N, d_S (free-ratio model) on the concatenated alignment of 3079 genes

	Human	Chimpanzee	Hominid	Mouse	Rat	Murid	Dog
Branch	0.017 ±	0.060 ±	0.159 ±	0.125 ±	0.155 ±	0.346 ±	0.275 ±
length	0.00017	0.00027	0.00053	0.00048	0.00044	0.00073	0.00064
d_N	0.0038	0.0176	0.0249	0.0197	0.0249	0.0461	0.0444
d_S	0.0109	0.0268	0.1323	0.1032	0.1265	0.3095	0.2239

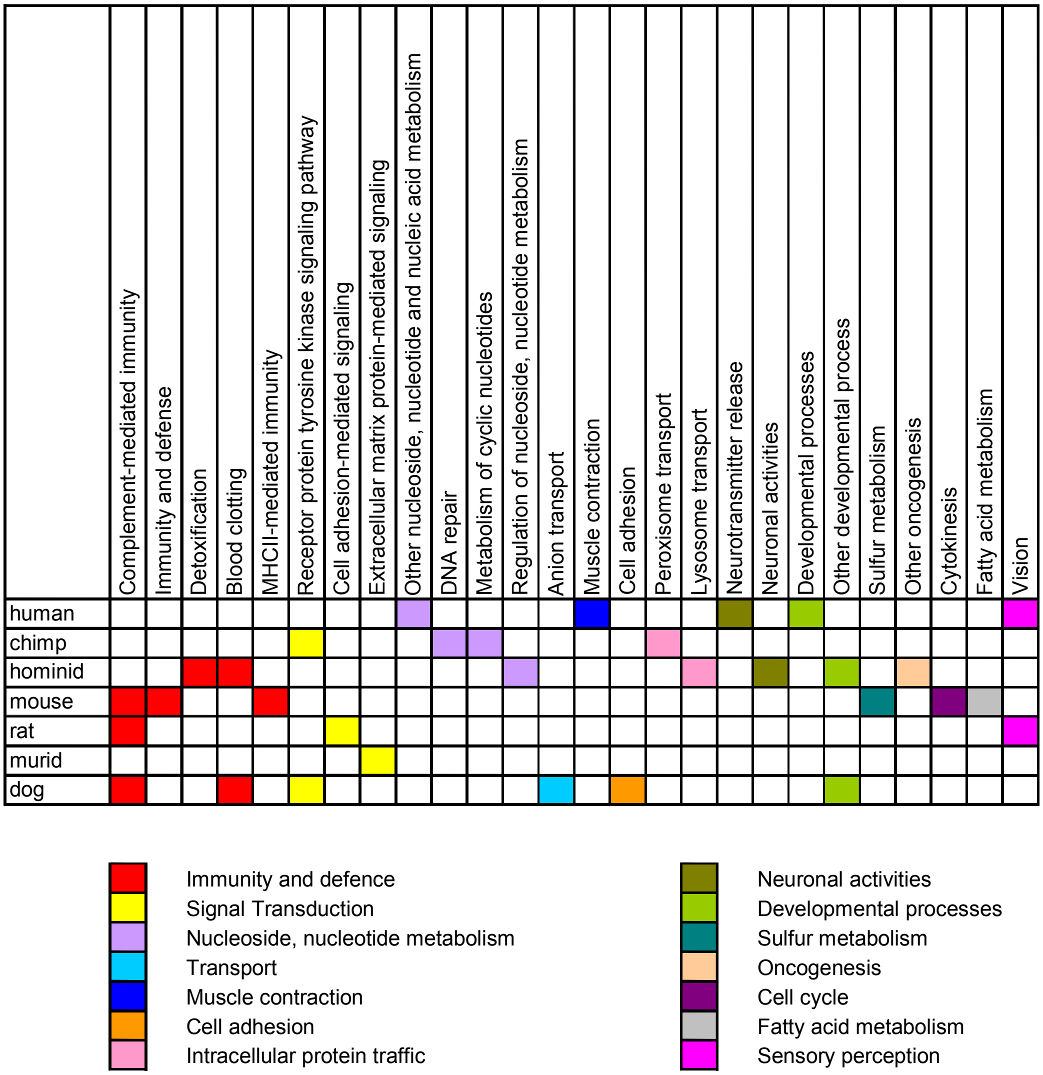
Note: Branch lengths are measured as the number of nucleotide substitutions per codon ± standard error calculated by the curvature method (Yang, 2007).

3.2.5 Functional processes affected by positive selection

A one-sided binomial test was used to test if the PSGs from each lineage were over-represented among the PANTHER Biological Process (BP) and Molecular Function (MF) ontology terms (Thomas et al., 2003). Each process was tested individually as a separate test. The overlaps between the PSGs from each lineage were very small so multiple testing corrections were not applied. None of the tests were expected to be significant so an FDR was not applied. The ontology terms that showed enrichment were then grouped by the BP family (Figure 3.2) and MF family (Figure 3.3) they belonged to, as defined by the PANTHER classification system (Thomas et al., 2003). Twenty-six BP ontology terms which belonged to fourteen BP families were enriched for PSGs ($p < 0.05$, binomial test). The ontologies that had the most representation by PSGs from the primate lineages were nucleic acid metabolism (*RBM16, RDM1, REPIN1, RKHD1* and *ZRSR2*) and transport (*CACNA1S, CNGA4, KCNK5, SLC5A9* and *SRL*). Categories of genes that can be associated with species-specific differences such as reproductive processes (*ARID2, INPP5B*), signal transduction (*CEACAM20, GPR111* and *GPRC6A, NR5A1, PDE6A, EMB, PIK3C2G, INPP5B, GIPC2*) and development (*IFRD2, MICALCL, MOV10, MYF5, ST8SIA3* and *TRIM67*) also showed enrichment. PSGs from the murid lineages showed over-representation mostly in the functional categories of immunity, defence and signal transduction. The same was done for each gene using the Molecular Function terms (Figure 3.3). Hydrolases and cell adhesion classes were also seen to have an excess of genes under positive selection across all species.

Figure 3.2 Biological Process ontologies over-represented by PSGs

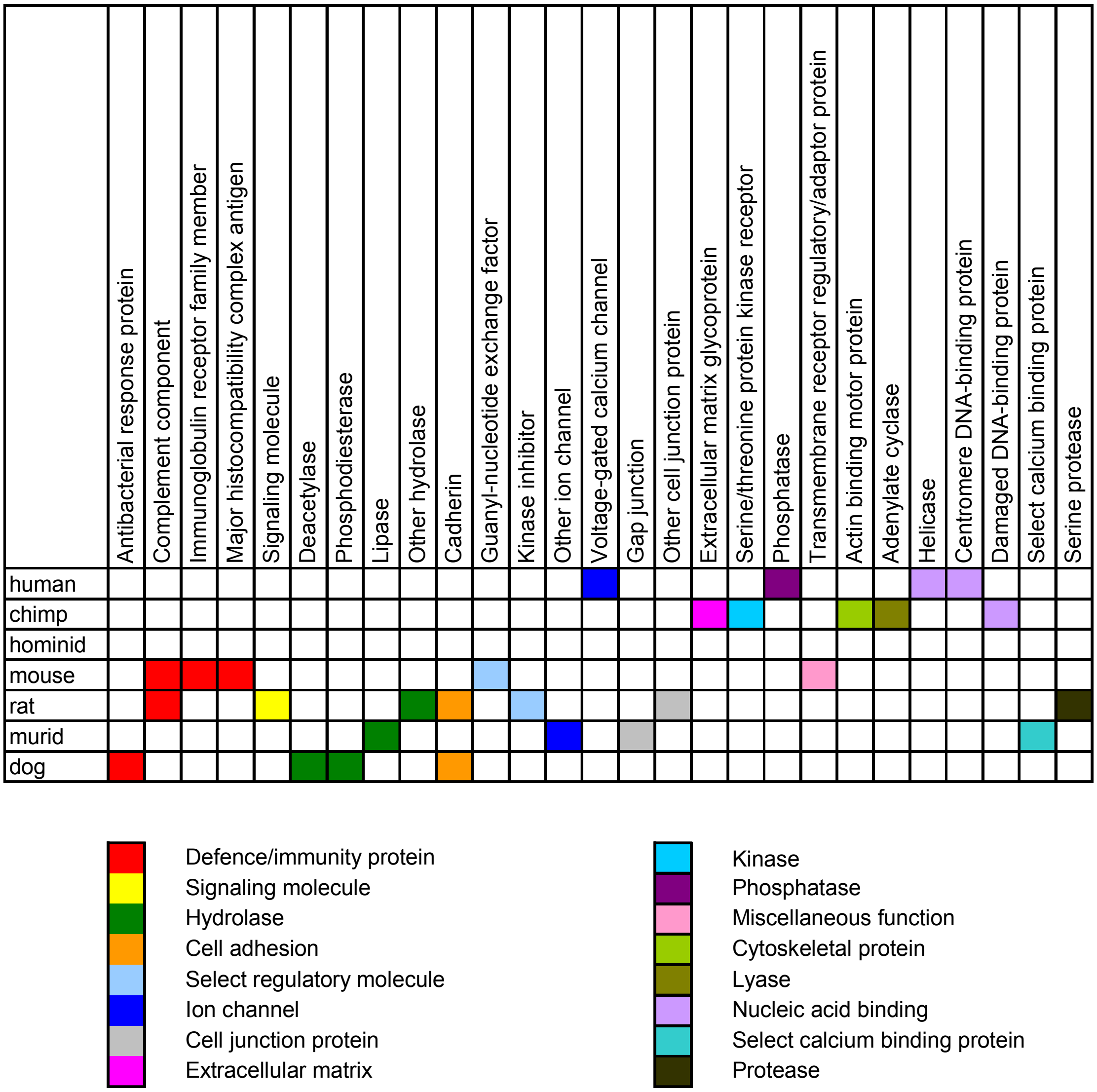
Biological Process ontology terms which had an over-representation of PSGs ($p < 0.05$). Ontology terms are grouped by functional protein PANTHER Biological Process families.



	Complement-mediated immunity	Immunity and defense	Detoxification	Blood clotting	MHCII-mediated immunity	Receptor protein tyrosine kinase signaling pathway	Cell adhesion-mediated signaling	Extracellular matrix protein-mediated signaling	Other nucleoside, nucleotide and nucleic acid metabolism	DNA repair	Metabolism of cyclic nucleotides	Regulation of nucleoside, nucleotide metabolism	Anion transport	Muscle contraction	Cell adhesion	Peroxisome transport	Lysosome transport	Neurotransmitter release	Neuronal activities	Developmental processes	Other developmental process	Sulfur metabolism	Other oncogenesis	Cytokinesis	Fatty acid metabolism	Vision
human									■					■				■		■						■
chimp						■				■	■					■										
hominid			■	■								■					■		■		■		■			
mouse	■	■			■																	■		■	■	
rat	■						■																			■
murid								■																		
dog	■			■		■							■		■						■					

Figure 3.3 Molecular Functions ontologies over-represented by PSGs

Molecular Function ontology terms which had an over-representation of PSGs ($p < 0.05$). Ontology terms are grouped by functional protein PANTHER Molecular Function families.



	Antibacterial response protein	Complement component	Immunoglobulin receptor family member	Major histocompatibility complex antigen	Signaling molecule	Deacetylase	Phosphodiesterase	Lipase	Other hydrolase	Cadherin	Guanyl-nucleotide exchange factor	Kinase inhibitor	Other ion channel	Voltage-gated calcium channel	Gap junction	Other cell junction protein	Extracellular matrix glycoprotein	Serine/threonine protein kinase receptor	Phosphatase	Transmembrane receptor regulatory/adaptor protein	Actin binding motor protein	Adenylate cyclase	Helicase	Centromere DNA-binding protein	Damaged DNA-binding protein	Select calcium binding protein	Serine protease
human														🟦					🟪				▣	▣			
chimp																	🟣	🔵			🟢	🟫			▣		
hominid																											
mouse		🟥	🟥	🟥							🔷									♦							
rat		🟥			🟨				🟩	🟧		🔷				▒											⬛
murid								🟩					🟦		▒											◈	
dog	🟥					🟩	🟩			🟧																	

🟥	Defence/immunity protein	🔵	Kinase
🟨	Signaling molecule	🟪	Phosphatase
🟩	Hydrolase	♦	Miscellaneous function
🟧	Cell adhesion	🟢	Cytoskeletal protein
🔷	Select regulatory molecule	🟫	Lyase
🟦	Ion channel	▣	Nucleic acid binding
▒	Cell junction protein	◈	Select calcium binding protein
🟣	Extracellular matrix	⬛	Protease

3.2.6 Positively selected genes on all lineages show evidence of co-evolution

To test if PSGs or proteins encoded by PSGs interact with fewer genes or proteins compared to genes that are not under positive selection, we queried a meta-database of biological interactions (see Section 2.12, (Rajagopalan and Agarwal, 2005)) with the list of all PSGs. The median number of interactors for 1) the genes subject to positive selection, and 2) the genes that were tested but for which no signal of positive selection was found, were calculated. For the 511 PSGs along all lineages, 155 (30%) did not have any annotated interactions with any other proteins and the median number of interactions was 5. For the 2568 genes in the test set with no evidence of positive selection, 783 (31%) did not have any interactors and the median number of interactors was also 5. Therefore it was concluded that PSGs do not have a lower median number of interactors than genes not under positive selection in the test set (p = 0.815; two-tailed Wilcoxon rank sum test), which suggests that the number of interactors is not a determinant for PSGs.

To determine if any of the PSGs interact with each other and form smaller clusters of sub-networks, we queried the same database with the lists of PSGs from each lineage. PSGs from all lineages except the human lineage formed clusters. For example, among the 162 chimpanzee PSGs, 9 clusters were found, consisting of 2 clusters of 3 genes and 7 clusters of 2 genes. We applied a permutation test to determine whether the number and size of the clusters formed was more than would be expected by chance. For example, for a random set of 162 genes (picked from the 3079 test set) would we expect the 9th cluster to be 2 nodes in size, given there are 8 other clusters of size 2 nodes or above? 1000 permutations were run. The overlaps between the PSGs from each lineage were very small so multiple testing corrections were not applied. For PSGs in both the chimpanzee and hominid lineages, the size of the smallest two clusters (chimpanzee clusters 8 (*PEX12, PEX19*) and 9 (*NRP1, MSI1*) and hominid clusters 3 (*DRD2, TH*) and 4 (*ITGAV, AZGP1*)) exceeded what would be expected by chance ($p < 0.05$) (Table 3.4) and in the dog lineage the third cluster (containing genes *SNTA1, DAG1* and *MUSK*) was significant. Therefore there is some evidence that PSGs are likely to interact and form sub networks. No interconnectivity was found between the genes positively selected within the human lineage and this statistic was not significant for the mouse, rat and murid lineages.

We also tested each cluster to determine whether the size of the cluster was more than that expected by chance given the number of interactors for each individual gene in the cluster. Again a permutation test was run to answer the question: would a random group of genes (from the 3079 test set) with the same number of interactors as those in the cluster, be expected to interact with each other? Multiple testing corrections were not applied as the PSGs from each lineage did not contain many overlapping genes. All 28 clusters were found to be significant ($p < 0.05$; permutation test) (Table 3.4). Therefore there is a highly significant phenomenon of PSGs interacting with other PSGs. To confirm this observation, further analysis was performed on the genes that interact with the beta 2 integrin gene (*ITGB2*) which showed evidence of positive selection along the rat ($p < 0.001$) and murid ($p < 0.05$) lineages. Three of its four known interacting alpha subunits (Ewan et al., 2005) showed positive selection either on the murid branch (*ITGAL*, $p < 0.01$; *ITGAX*, $p < 0.05$) or on the mouse branch (*ITGAD*, $p < 0.001$).

One of the sites (M165Q) under positive selection in the *ITGAL* gene was found to be in the metal ion-dependent adhesion site (MIDAS) motif. This motif functions to mediate ligand binding in a metal ion-dependent manner. Positively selected residues in the *ITGB2* gene included one in the linker region between the PSI (plexin/semaphorin/integrin) domain and the I-like domain, prior to the MIDAS motif in this protein.

Table 3.4 Interacting clusters formed between PSGs on each lineage

Cluster number	Genes in cluster	*p* value of cluster size given previous clusters	*p* value of cluster given number of interactions per gene**
Chimpanzee			
1	PCSK5, BMP4, PHOX2A	0.981	0.0013
2	LHB, OTX1, JUB	0.391	0.0001
3	XPC, RAD23A	0.519	0.0035
4	NUCB1, PTGS1	0.346	0.0046
5	ITGB6, ALOX12	0.227	0.0030
6	MYO18A, TRADD	0.131	0.0028
7	GSTP1, MAP2K4	0.075	0.0442
8	PEX12, PEX19	0.036*	0.0003
9	NRP1, MSI1	0.019*	0.0008
Dog			
1	CFP, TAL1, SERPINB1, MMP12, PRF1, BCL2, HRG, ITGA5, COMP	0.385	< 0.0001
2	CD79A, HCLS1, LCP2	0.209	0.0012
3	SNTA1, DAG1, MUSK	0.036*	0.0002
4	LRP5, SLC2A2	0.171	0.0026
5	ALB, MCAM	0.082	0.0123
Hominid			
1	CCL19, CD86, MADCAM1	0.335	0.0015
2	MRC2, COL4A4	0.186	0.0028
3	DRD2, TH	0.045*	0.0488
4	ITGAV, AZGP1	0.008*	0.0080
Mouse			
1	HLA-DRB1, HLA-DQA2	0.755	0.0123
2	C1R, C1QA	0.288	0.0030
Murid			
1	TLR5, CD86, PTGIR	0.678	0.0001
2	SCNN1G, SPTA1, HECW1	0.432	0.0021
3	CNR1, RAPGEF1	0.190	0.0110
4	F5, GP1BA	0.064	0.0032
Rat			
1	CDKN2D, TRIM21, CDKN1B, CAST, ICAM1, CFD, ITGB2, C3	0.360	< 0.0001
2	KCNA4, ACTN2, PIK3R5	0.526	0.0016
3	PIM1, RP9	0.280	0.0063
4	ASPH, HDAC4	0.118	0.0053

*$p < 0.05$

**All tests to investigate whether the size of the cluster would be more than that expected by chance, given the number of interactors for each individual gene in that cluster, were significant ($p < 0.05$).

3.2.7 Are malleable genes common targets of positive selection pressure?

There were several genes that showed signatures of selection in multiple lineages. We found that 17 genes overlap between the human and chimpanzee PSGs, 8 genes overlap between the mouse and rat PSGs and 8 genes intersect between the hominid and murid PSGs, all significantly greater than that expected by chance (2.8; $p < 7 \times 10^{-10}$, 1.9; $p < 5 \times 10^{-04}$, 1.5; $p < 8 \times 10^{-05}$, Fisher's exact test of proportions) (Table 3.5).

These genes suggest the presence of some common targets of positive selection in each of the pairs of species and may represent malleable genes that are involved in adapting to changing external environments, like sensory perception and dietary content. Some examples of such genes are the olfactory receptor *OR4F17*, a PSG in both the human and chimpanzee lineages, which illustrates what is known about both humans and chimpanzees showing extensive evidence of olfactory adaptation (Gilad et al., 2005). *DHDH*, a PSG in both the mouse and rat lineages, is involved in carbohydrate metabolism.

The hominid and murid lineages share PSGs involved in cell differentiation (*FZD2*) and reproduction (*TXNDC3*). The *TXDNC3* protein (Sptrx-2) participates in the final stages of sperm tail maturation in the testis and/or epididymis and is a structural component of the mature fibrous sheath of spermatozoa (Miranda-Vizuete et al., 2004). Proteins involved in reproduction tend to have evolved under positive selection (Wyckoff et al., 2000; Swanson and Vacquier, 2002).

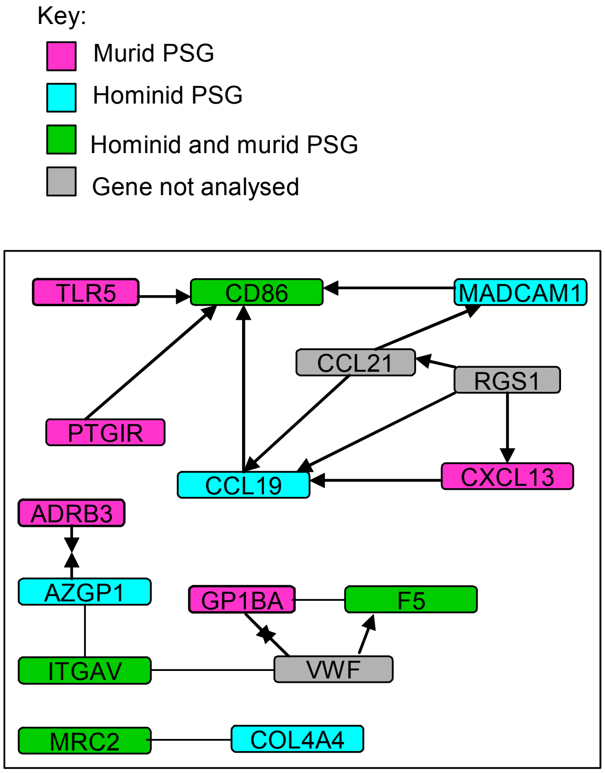
Table 3.5 Positively selected genes common in adjacent lineages

Lineages	*p* value	Gene Symbol
Human and Chimpanzee	$p < 7 \times 10^{-10}$	ABCF1, ALPPL2, CNGA4, PIK3C2G, ZRSR2, KIAA0372, C8ORF42, ANGEL1, MICALCL, OR4F17, ZNF324B, ANKRD35, GIPC2, RUFY4, RBM16, MGC50722, INPP5B
Mouse and Rat	$p < 5 \times 10^{-04}$	SYT4, STS, CA6, CDC14B, TARP, RRAGA, DHDH, C19ORF16
Hominid and Murid	$p < 8 \times 10^{-05}$	TXNDC3, ITGAV, MRC2, CLSTN2, ZNF665, CD86, FZD2, F5

Note: The *p* value is from Fisher's exact test of proportions to test for significance.

We combined some of the common murid and hominid PSGs with PSGs only in the hominid or murid lineage. We then used this dataset to query the previously described database of biological interactions to find significant networks and found three networks of genes involved in inflammation processes (Figure 3.4). All the functional processes concerned with inflammation are represented by the genes that appear in these networks: genes such as *F5*, *GP1BA*, *VWF*, *PTGIR* are involved in blood coagulation, cell-adhesion genes such as *MADCAM1*, *ITGAV*, genes that participate in inflammatory response (*TLR5, CXCL13, CCL19, CCL21*) and immune defence (*CD86, AZGP1*) as well as other related transport proteins. As the challenges to the immune system are constantly evolving and changing, we would expect immune system genes to be constantly under positive selection pressure to adapt to new incoming challenges. This can be seen with the MHC molecules (Bernatchez and Landry, 2003) and may also be true of the genes in our network.

Figure 3.4 Positively selected genes along the hominid and murid lineages cluster to form networks involved in inflammatory processes



3.3 Discussion

This comprehensive evolutionary study offers the first genome scan for the action of positive selection pressure influencing human genes, their orthologues in model organisms and the two ancestral lineages. Having generated a robust set of genes that have undergone positive selection in many closely-related species, we have the opportunity to ask a wide array of fascinating scientific questions on the relationships between these genes.

As seen in other studies as well (Arbiza et al., 2006; Bakewell et al., 2007), the number of PSGs in chimpanzee was much higher than in human. We considered the elevation in chimpanzee PSGs to be artefactual, perhaps caused by sequencing errors in the unfinished chimpanzee genome sequence. The rat lineage also shows a slightly higher number of PSGs (89) compared to the mouse lineage (65).

We noticed that the second release of the chimpanzee genome (6x, Pan_troglodytes-2.1) was a radical improvement to the 4x sequence (Pan_troglodytes-1.0 (CSAC, 2005)). We had used the 4x sequence, which was also used in previous genome scans (CSAC, 2005; Arbiza et al., 2006; Bakewell et al., 2007), in prior analyses (result set A) and encountered unusually large numbers of chimpanzee PSGs, and accordingly re-ran with the 6x sequence anticipating fewer PSGs.

Since use of the 6x genome sequence also resulted in large numbers of chimpanzee PSGs (result set B), we considered that perhaps errors were also present in the 6x genome sequence. However, one would expect the number of sequencing errors in the chimpanzee and dog genomes to be approximately the same as both genomes have comparable coverage (chimpanzee at 6x; dog at 7.6x) but at lower sequence coverage than the human genome sequence (CSAC, 2005). On this basis the number of PSGs in the dog lineage would be elevated as well, but perhaps we observe more chimpanzee PSGs because low sequence quality would have a bigger impact on lineages with short branches in the species phylogeny. The dog branch is much longer than the chimpanzee branch and thus the dog branch is able to accommodate more nucleotide substitutions, masking the detrimental impact of sequencing errors on selection results along the dog lineage.

Any analysis of positive selection must first ascertain complete confidence in the homology between genes used for inference of positive selection and the robust identification of their open reading frames and alignment of their sequences. Hence another source of error, particularly under automation, would be incorrect or incomplete gene predictions that may result in frameshifts in open reading frames or alignment errors. Ambiguous orthologue calls or misalignment may cause sequences to appear more divergent than they are and hence generate false positive results.

To correct for errors in genome sequence and errors in gene prediction and alignment, we applied conservative filters and complete manual checking to ensure that our results set was robust. We applied stringent cut-offs during the orthologue calling procedure to ensure we studied only truly orthologous sets and also controlled sequencing errors by masking out the divergent ends of partial sequences. We manually inspected alignments, discarding orthologous sequences which contained frameshifts relative to the human sequence or those that included regions of very low similarity (see Section 2.9). It is believed that any automated process of orthologue calling, open reading frame prediction and sequence alignment is prone to errors and it is suggested that manual examination and adjustment is the only way to prevent the possibility of false positives. The impact of this manual adjustment is indicated by the 392 positive selection results that were disregarded from results set B during our data curation steps. The high level of quality control is also the reason that we have identified comparatively fewer PSGs than some other studies (Clark et al., 2003; Arbiza et al., 2006), despite the increased power associated with the inclusion of more species.

The overall ω values that we obtained from the free-ratios model for each lineage (Figure 3.1) were comparable to the median ω values published by the Chimpanzee Sequencing and Analysis Consortium (CSAC, 2005) (mouse 0.142, rat 0.137, human 0.208, chimp 0.194) but were more similar to those from Rhesus Macaque Genome Sequencing and Analysis Consortium (Gibbs et al., 2007) (human 0.169, chimpanzee 0.175, mouse 0.104). This suggests that the strict criteria used to select our input gene set have not introduced a bias for genes with high ω values in humans and chimpanzees. The higher median values observed in the chimpanzee lineage suggest that overall nonsynonymous mutations were fixed much faster along the chimpanzee lineage than along the human lineage since the separation of the two species.

The functional categories enriched for PSGs in this study were found to closely correlate with those detected in previous genome scans (Holbrook and Sanseau, 2007). The consensus is compelling given the different techniques used in each study and the risk of false positives inherent in these large-scale studies. It is interesting to note that among the five species analysed, protein families with distinct functions could be identified as evolving under positive selection for each species. Techniques to connect positive selection with function are still in the early stages of development, but gradual progress is being made. As more data becomes available on the function of each individual amino acid, from structural or mutagenesis studies, it will become possible to connect function and positive selection. Data, such as that generated by this study, provides a preliminary starting point for experimental follow-up.

Positive selection pressure would be expected to act not just on one gene at a time but on pathways of genes, but evidence has been scant so far. We found that genes that were subjected to positive selection along the same lineage were significantly more likely to interact with each other than with genes not under selection, the first evidence for co-evolution of genes as a widespread phenomenon in mammals. We suggest that the high level of connectivity between PSGs is caused by compensatory change of a protein's interaction partners when a protein undergoes change in response to selection. This was exemplified by the evidence of positive selection pressure in *ITGB2* and its interacting alpha subunits, *ITGAL*, *ITGAX* and *ITGAD*. This suggests that the major participants of integrin-signalling have co-adaptively evolved in the rodent species.

Chapter 4

Evidence of Excessive Adaptation in Chimpanzees

4.1 INTRODUCTION

Comparative analyses of the genomes of mammalian model organisms can provide insight into human adaptation as the availability of high quality functional annotation allows prediction of the likely consequences of adaptive evolution in particular genes. Such analyses can also indicate the numbers of genes that have undergone positive selection in other species such as chimpanzees, which we commonly believe to be fewer in number than in human (Hawks et al., 2007).

In result set C the number of PSGs detected on the human lineage was 54 ($p < 0.05$) whilst the number of PSGs was still highest along the chimpanzee lineage (162, $p < 0.05$), having many more genes than any of the other lineages and approximately three times more than along the human lineage. This was surprising despite the findings of other reports which mention that a high number of genes underwent positive selection during chimpanzee evolution (Arbiza et al., 2006; Bakewell et al., 2007). Bakewell et al. (using a wholly different methodology to this study) identified 21 positive chimpanzee genes and 2 positive human genes from an initial data set of 13,888 genes. Elevated numbers of PSGs along the chimpanzee lineage were also found by Arbiza et al. (2006) who obtained 1.12% of genes under positive selection in the human genome and 5.96% in the chimpanzee genome, which is in close accordance with 1.75% (human) and 5.26% (chimpanzee) obtained here.

In the following discussion, evidence is presented to argue against the possibilities that this result is due to artefacts introduced by genome sequence coverage, gene sample selection or algorithmic sensitivity to errors in sequence data or alignments. Instead, it is concluded that the elevated number of chimpanzee positively selected genes is a true reflection of evolutionary history and is most likely due to positive selection being more effective in the large populations of chimpanzees in the past or possibly remarkable adaptation in the chimpanzee lineage.

4.2 Results and Discussion

4.2.1 Taxon sampling does not affect detection of positive selection

This study included five species exemplars whilst previous studies have been more restricted (Nielsen et al., 2005; Bakewell et al., 2007). However, the effect of taxon sampling on the detection of positive selection is largely unknown. To address this question we conducted permutation analyses of the original five-species alignments to determine if the exclusion of each non-human species in turn affects the results obtained.

Both alignments from results sets B and C were used in this analysis to ensure that the manual curation step performed to generate results set C did not skew our results. For the first permutation test, after the sequence was removed from the alignment, the remaining sequences were not re-aligned prior to analysis with the branch-site model. For the second permutation test, the sequences remaining after a sequence was excluded were re-aligned. The number of PSGs each analysis had in common with result sets B (Table 4.1) and C (Table 4.2) was calculated. The common set of PSGs in the five- and four-species analyses number includes genes that were identified as being under positive selection in both tests. The difference in *p* value for a particular gene to be under positive selection in the five- and four-species analyses is simply an outcome of the value of the test statistic and does not reflect a disparity in the significance of the result. This is also emphasised by the fact that although the *p* values might be different between the two analyses, in the majority of cases, the same residues were being reported as having high probabilities of being under positive selection (data not shown).

The number of genes reported in the permutation tests was very similar to the number of PSGs in result sets B and C. For example, in result set B, the mouse lineage had 65 PSGs and in the permutation test without re-alignment, there were 74 mouse PSGs when the chimpanzee sequence was removed and 67 PSGs when the dog sequence was removed. However, this does not represent a complete overlap as some genes that were not significant for positive selection in the test using the five-species alignment became non-significant in the test using the four species alignment, probably due to loss of power when fewer species were used in the analysis.

Counter-intuitively, many genes that were not significant in the five-species analyses became significant in the four-species analysis. It would be expected that as the number of species included in the analysis was decreased, the number of positive genes found would also decrease. This could be because if a divergent sequence caused ambiguity in the five-species alignment, then re-aligning the data after removal of the divergent sequence results in a more conservative alignment.

The numbers of PSGs in each lineage after sequence exclusion without re-alignment were almost the same as the numbers of PSGs after sequence exclusion and with re-alignment. For example, from results set C, the number of human PSGs when the dog sequence was removed was 54 in permutation test 1 and 50 in permutation test 2. The two tests also had a significant number of PSGs in common (Table 4.3), which indicated that re-aligning the data did not make a significant difference to the results for positive selection.

Table 4.1 Number of PSGs after sequence exclusion (result set B)

Taxon removed	Human	Chimpanzee	Hominid	Mouse	Rat	Murid	Dog
Number of PSGs after taxon exclusion and no re-alignment (permutation test 1)							
Chimpanzee				134 (94)	176 (96)	101 (85)	199 (96)
Dog	73 (71)	368 (94)		115 (74)	149 (79)		
Rat	74 (90)	364 (97)	63 (71)				175 (89)
Mouse	67 (80)	369 (98)	51 (61)				161 (85)
Number of PSGs after taxon exclusion followed by re-alignment (permutation test 2)							
Chimpanzee				133 (80)	183 (88)	99 (69)	200 (85)
Dog	63 (61)	372 (87)		100 (65)	155 (70)		
Rat	71 (77)	364 (90)	69 (55)				170 (77)
Mouse	71 (72)	361 (90)	39 (43)				153 (72)

Note: Numbers ($p < 0.05$) are only shown for lineages for which there were no changes in topology when the taxon in question was removed from the tree.
In parentheses are the numbers of common genes in the analyses with the sequence excluded and the five-species alignment as a percentage of the number of PSGs in the five-species alignment.

Table 4.2 Number of PSGs after sequence exclusion (result set C)

Taxon removed ↓	Human	Chimpanzee	Hominid	Mouse	Rat	Murid	Dog
Number of PSGs after taxon exclusion and no re-alignment (permutation test 1)							
Chimpanzee				74 (88)	87 (75)	80 (70)	113 (79)
Dog	54 (67)	174 (89)		67 (74)	77 (58)		
Rat	54 (83)	162 (92)	53 (57)				89 (72)
Mouse	59 (80)	163 (90)	50 (54)				79 (81)
Number of PSGs after taxon exclusion followed by re-alignment (permutation test 2)							
Chimpanzee				72 (77)	100 (70)	72 (56)	123 (75)
Dog	50 (50)	200 (86)		66 (65)	87 (54)		
Rat	48 (57)	188 (86)	58 (52)				98 (68)
Mouse	62 (67)	189 (86)	46 (46)				92 (64)

Note: Numbers ($p < 0.05$) are only shown for lineages for which there were no changes in topology when the taxon in question was removed from the tree.
In parentheses are the number of PSGs that were common to the four-species analyses and the original five-species alignment (result set C) as a percentage of the number of PSGs in the five-species alignment.

Table 4.3 Number of PSGs common to permutation tests 1 and 2

Taxon removed ↓	Human	Chimpanzee	Hominid	Mouse	Rat	Murid	Dog
Chimpanzee							
Result set B				108	154	80	167
Result set C				60	74	61	101
Dog							
Result set B	49	324		88	124		
Result set C	38	164		55	64		
Rat							
Result set B	57	328	45				139
Result set C	36	152	40				77
Mouse							
Result set B	54	325	30				128
Result set C	46	149	37				74

It should be noted that the percentages of PSGs the analysis with five species had in common with the permutation tests using four species were lower for result set C compared to result set B (Tables 4.1 and 4.2). This is due to the smaller overall size of the result set C; the raw numbers of genes in common were similar.

Comparison of the same branches in both the re-aligned and non-re-aligned analyses shows that the effect of taxon elimination on the number of PSGs was most pronounced on the ancestral lineages, resulting in the most severe loss in the number of detected PSGs. This could be because there are no direct observations to obtain data for the internal branches; instead the sequence is inferred by ancestral reconstruction. If more species were used in the analysis, sequence reconstruction can be performed more accurately. Hence, in this case, the removal of one sequence influences reconstruction and can considerably affect the number of positive genes detected.

Among the extant lineages, removal of the dog sequence seems to have had the most severe effect (percentages of overlapping genes range from 50-89%). This is probably because dog is an outgroup in the species phylogeny and hence removal of this branch substantially decreases the power to detect positive selection. The exclusion of a sequence from the remaining three taxa has slightly lesser effects (percentages range from 57% to 92%) but an assessment of the number of genes that were positive in all three analyses (see Table 4.4 for comparison with result set B and Table 4.5 for comparison with result set C) shows that detection of positive selection in a substantial portion of genes is robust to all manipulations and species exclusions.

Table 4.4 Number of PSGs common to both permutation tests and the five-species analysis (result set B)

Taxon removed	Human	Chimpanzee	Hominid	Mouse	Rat	Murid	Dog
Chimpanzee				96 (79)	143 (92)	68 (79)	143 (88)
Dog	40 (58)	302 (85)		77 (64)	110 (71)		
Rat	51 (74)	315 (89)	28 (57)				128 (79)
Mouse	45 (65)	315 (89)	22 (45)				121 (75)

Note: In parentheses are the numbers of genes that the three analyses had in common, shown as a percentage of the number of PSGs in the five-species alignment.

Table 4.5 Number of PSGs common to both permutation tests and the five-species analysis (result set C)

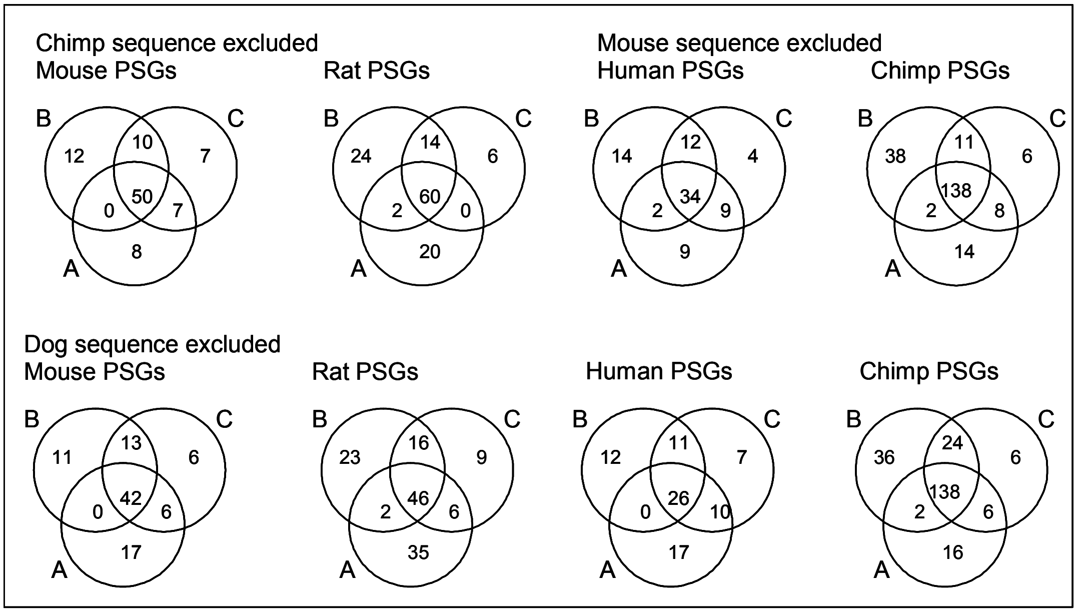
Taxon removed	Human	Chimpanzee	Hominid	Mouse	Rat	Murid	Dog
Chimpanzee				50 (77)	60 (67)	45 (56)	73 (75)
Dog	26 (48)	138 (85)		42 (65)	46 (52)		
Rat	30 (56)	140 (86)	28 (50)				64 (66)
Mouse	34 (63)	138 (85)	25 (45)				59 (61)

Note: In parentheses are the numbers of genes that the three analyses had in common, shown as a percentage of the number of PSGs in the five-species alignment.

When we compare our two most divergent species (Figure 4.1), it can be seen that the numbers of mouse and rat genes that were under positive selection when the chimpanzee sequence was excluded were approximately the same as when the dog sequence was excluded. Similarly the number of human and chimpanzee genes under selection were quite similar when the mouse sequence was excluded and when the dog sequence was excluded.

Figure 4.1 Summary of results from taxon exclusion studies

Circle A: Five-species alignment
Circle B: Species exclusion and no re-alignment
Circle C: Species exclusion and re-alignment
The exclusion of the chimpanzee sequence and the mouse sequence are compared to the exclusion of the dog sequence. The dog lineage, as the outgroup, had the most impact on the number of PSGs on the other lineages.



The high numbers seen are a reflection of the stability of the results regardless of changes in the number of taxa used, changes in tree topology and also changes in the alignment. We conclude that once rigorous orthologues are established the results are fairly consistent regardless of the species being removed. We can conclude that PAML is robust to the effects of taxon sampling and the determination of PSGs reported in this study is accurate and not an effect of taxon sampling. In particular, the inclusion of chimpanzee sequences in our study did not affect the inference of positive in the other four species, nor in the ancestral lineages.

4.2.2 Chimpanzee PSGs are lineage specific

To determine whether the PSGs seen in the chimpanzee lineage were also under positive selection in other non-human primates, we performed a pilot study with other primate sequences, whose draft assemblies were available (macaque, orangutan and marmoset) at the same coverage (5-6x) as the chimpanzee genome. Marmoset (6x coverage), orangutan (6x coverage) and macaque (5.2x coverage) supercontigs were downloaded from the Washington University Genome Sequencing Center (http://genome.wustl.edu/).

Eleven of the 162 chimpanzee PSGs were selected as they had the most number of residues predicted to be under positive selection. For these genes, orthologous sequences in the three primate genomes were obtained by gene prediction using GeneWise (Birney et al., 2004). The protein sequences for these orthologous sequences were added to the original alignment of the five species used in the genome scan analysis.

We then performed positive selection analyses under the branch-site model on the resulting new alignments. All primate branches and branches leading to primates were tested as the foreground lineage in turn. The tree topology used for all the analyses was (((((Human, Chimp), Orangutan), Macaque), Marmoset), (Mouse, Rat), Dog). The addition of these three primate sequences to the original 5-species alignment did not change the length of the alignment as these genes were approximately the same length in all these mammalian species.

Table 4.6 Test statistic (2x Δ*ln*L) from chimpanzee lineage branch-site analyses with and without orangutan, marmoset and macaque sequences

Gene	Chimpanzee (original analysis)	Chimpanzee (with primate sequences)	Orangutan	Macaque	Marmoset	Ancestor to HCOM
AQP2	87.08**	77.41**				
EEF1G	40.88**	36.23**			131.94**	8.54*
ELF4	78.34**	64.28**		166.93**	28.11**	
HCRTR1	49.61**	53.64**				
TKTL1	25.85**	24.84**			19.74**	
DYRK2	78.86**	85.44**			8.74*	
PIGV	127.93**	134.6**				
PSD2	150.79**	139.6**	216.47**		66.84**	
CCDC97	90.45**	100.5**				
CXorf38	52.64**	57.87**				
GIYD1	47.88**	33.88**				

* $p < 0.05$; ** $p < 0.001$

All eleven genes remained subject to positive selection along the chimpanzee lineage (Table 4.6). The sites predicted to be under selection in the chimpanzee lineage were also the same in the analysis before and after addition of the primate sequences. Some of the genes were under positive selection along the other primate lineages, with five of the genes under positive selection in the marmoset lineage (a New World monkey). The amino acid differences observed in the eleven chimpanzee sequences are specific to the chimpanzee, with the other primate sequences having the same state as the human sequence. This suggests that the human state is the ancestral state and the chimpanzee state is the derived change with the adaptation observed in the chimpanzee being lineage specific.

4.2.3 Functional analysis of chimpanzee PSGs

It is also important to identify as far as possible the functional significance or biological grouping of these chimpanzee genes under positive selection. Classification of PSGs using the PANTHER Biological Process (BP) and

Molecular Function (MF) ontology terms (Thomas et al., 2003) showed that biological processes that were over-represented by chimpanzee PSGs included DNA repair, metabolism of cyclic nucleotides, peroxisome transport and the serine/threonine kinase signalling pathway. Each BP and MF ontology term was tested separately so multiple testing correction was not applied.

Interestingly, approximately a third (52 out of 162) of the chimpanzee PSGs are orthologues of human genes that are of unknown biological function. This proportion of 52 genes is significantly high ($p < 0.036$; Fisher's exact test of proportions) compared to the number of genes with unknown function among the human PSGs ($p < 0.053$). Of these, 13 genes were identified to contain the IMP dehydrogenase/GMP reductase domains. This family is involved in the biosynthesis of guanosine nucleotide. IMP dehydrogenase catalyses the rate-limiting reaction of de novo GTP biosynthesis, the NAD-dependent reduction of IMP into XMP. GMP reductase converts nucleobase, nucleoside and nucleotide derivatives of G to A nucleotides, and maintains the intracellular balance of A and G nucleotides. IMP dehydrogenase is associated with cell proliferation and genes that contain this domain are considered to be possible targets for cancer chemotherapy. Nielsen et al. (2005) also found many genes with unknown biological functions in both their set of chimpanzee and human PSGs but showed sequence similarity to known transcription factors.

There is the possibility that these genes of unknown function might contain incorrectly predicted open-reading frames and hence might be falsely detected to be under positive selection when compared to the other mammalian species in our analyses. Current gene sets are mostly built by gene-prediction software of which the error rate can only be determined by manual annotation of genes and their alternately-spliced variants.

4.2.4 No correlation with genes under selection in human populations

The number of PSGs from result set B was compared with genes shown to be under positive selection pressure within human populations (Voight et al., 2006; Tang et al., 2007). We did not see any evidence of a relationship between a gene being positively selected within human populations and in our mammalian species. In fact, there seems to be a trend that suggests that genes are less likely to have been selected along the hominid branch if they were under selection in recent human history. This is evident in the lower proportion of genes that were both under recent positive selection and positively selected along the hominid branch (0.0011) compared to the proportion of genes under positive selection along the hominid branch alone (0.0149).

4.2.5 Hypotheses to explain the high number of PSGs on the chimpanzee lineage

Our results after data curation and from investigating the effects of taxon sampling exclude the possibility that taxon sampling have affected the results. However there is still the possibility that sequence errors in the genes from a species might affect the number of PSGs and could perhaps cause the relatively high numbers of PSGs in the chimpanzee lineage. Other likely explanations for the elevation in PSGs along the chimpanzee lineage include:

1. High chimpanzee polymorphism: the individual chimpanzee sequence has been reported to have many high quality, single-nucleotide polymorphisms (SNPs) with a heterozygosity rate of 9.5 x 10^{-4}. This rate is slightly higher than what was seen among West African chimpanzees (8.0 x 10^{-4}) (CSAC, 2005) which have similar diversity levels to human populations (Sachidanandam et al., 2001).
2. Population structure: one of the factors affecting the strength of selection is population size. Since selection must be greater than 1/4Ne (where Ne is the effective population size) to reach fixation, it follows that genes are under stronger purifying selection in larger populations. Chimpanzees are noted to be more polymorphic indicating that they had larger (22,400 to 27,900) (Won and Hey, 2005) long-term populations than humans (10,000). Comparison of human sequences with great ape sequences revealed humans have reduced nucleotide diversity and a signal of population expansion (Kaessmann et al., 2001). Thus positive selection may have reduced efficacy in humans than in chimpanzees which may explain some of the elevation in chimpanzee PSGs.
3. Positive selection may have acted on the human and chimpanzees lineages during different periods of evolutionary time. In humans many selective events are relatively recent and are thought to be a result of adaptation to migration and domesticity (Tishkoff and Verrelli, 2003; Voight et al., 2006; Balaresque et al., 2007; Tang et al., 2007). This recent positive selection pressure is not amenable to detection by PAML and is more likely to be discerned by study of human population data (e.g. (Tang et al., 2007)). Comparison of our results with data from Tang et al. (2007) did not show a

large overlap of genes. This suggests that different sets of genes were subject to positive selection after the separation of the human and chimpanzee lineages than those sets of genes subject to recent positive selection within human populations.

We observe many chimpanzee genes and possibly genes in other primates have been subject to positive selection during the evolution of their anthropoid ancestor. Since medical research and the vast majority of biological research have focussed on discovering more about human biology, we know a lot less about chimpanzee-specific characteristics. The number of PSGs on the chimpanzee lineage that are not false positives due to sequence errors or false gene predictions, would suggest that these chimpanzee adaptations are at least as striking as our much-vaunted human-specificities.

4.2.6 Summary

The comparative study of mammals offers many fascinating questions for researchers. In this study molecular evolutionary signals were used to predict how mammalian species have evolved. The approach taken in this study has confirmed an uncharacterised set of positively selected genes in the chimpanzee lineage. This adaptation could be in previously unrecognised aspects of chimpanzee biology, even for instance, in sensory or cognitive systems. Many of the chimpanzee PSGs have unknown functions which might suggest they belong to pathways that researchers have not focussed on as being relevant to human biology. It is anticipated that as more primate genome sequences become available, we will be able to determine whether other primates also have high numbers of genes under positive selection, as seen in the chimpanzee lineage. It is also hoped that the study of these groups will yield answers to the questions of what is "humanness" and "chimpness" at a genetic level.

Chapter 5

Positively Selected Genes and Associations with Human Diseases

5.1 INTRODUCTION

Much scientific and medical progress has depended on experimental findings in model organisms being extrapolated to humans because it is very often the case that diseases shared between humans and other mammals mirror each other in molecular processes. However, even species that are evolutionarily close to each other, such as humans and chimpanzees, often experience the same medical condition with varying symptomatology, as seen in the cases of Alzheimer's disease or AIDS. In addition, many diseases are far more prevalent in humans than in other primates (Olson and Varki, 2003; Varki and Altheide, 2005).

Comparison of disease prevalence and symptomatology across species is complicated by the fact that modern human lifestyles, very far from the conditions of early human evolution, may reveal susceptibilities to disease that were not evident in the early history of the human species (Young et al., 2005). However, there are biomedical differences between humans and other animals that cannot be wholly explained by lifestyle (Olson and Varki, 2003; Varki and Altheide, 2005).

Genetic disease can occur as a by-product of an adaptation which confers a large selective advantage (Nesse and Williams, 1995). For instance, the seemingly human-specific disease of schizophrenia (Crespi et al., 2007) and the greater human susceptibility to Alzheimer's disease compared with primates (Gearing et al., 1994) may be a by-product of the human specialisation for higher cognitive function (Keller and Miller, 2006). Genes that have evolved different functions since the divergence of humans from other primates may be involved in this adaptation and therefore in diseases that affect the adaptation.

Besides Alzheimer's disease and schizophrenia, many other diseases also differ in frequency and symptomatology between humans and other mammals. Olson and Varki (2003) and Varki and Altheide (2005) list some of these diseases with the emphasis on differences between humans and non-human primates, indicating that for these diseases chimpanzees are not good models despite their close evolutionary relationship with humans. Comparative evolutionary genomics may offer insights into these disease mechanisms as correlations between molecular differences that arose during species evolution and phenotypic differences in diseases between species may throw light on disease-causative genes and pathways. Consistent with this rationale, genes included in the Online Mendelian Inheritance in Man (OMIM) database (http://www.ncbi.nlm.nih.gov/Omim) have been found to be enriched for signals of positive selection pressure ((Clark et al., 2003; Smith and Eyre-Walker, 2003; Huang et al., 2004; Bustamante et al., 2005; Bakewell et al., 2007); but see (CSAC, 2005)).

5.1.1 Gene products that are common drug targets

Some genes that were reported to have demonstrated evidence of positive selection in the genome scan described in this thesis are currently being pursued as drug targets in the pharmaceutical industry. However these genes are not represented in the OMIM database as they are involved with substance abuse and addiction.

One example is the cannabinoid receptor, *CNR1*, a G-protein coupled receptor. The cannabinoid receptor binds cannabinoids, the psychoactive ingredients of marijuana, mainly delta-9-tetrahydrocannabinol, as well as other synthetic analogs. This gene is the target for marketed drugs such as Nabilone and Marinol with indications for anorexia and emesis. Although the endogenous ligand that binds to the cannabinoid receptor has not yet been characterised, the well known psychoactive effects and other CNS actions caused by the binding of marijuana such as hallucinations, memory deficits, altered time and space perception, CNS depression and appetite stimulation, have been extensively studied. The cannabinoid receptors are also believed to play a role in neurogenesis during development (Julian et al., 2003). In comparisons between primates and rodents, it appears that the distribution in the forebrain of the cannabinoid receptor has altered during evolution (Harkany et al., 2005), but so far there have been no reports of positive selection pressure. The genome scan in this thesis found positive selection pressure in the murid lineage suggesting divergence between human and rodent species.

Another drug target is the dopamine receptor 2, *DRD2,* which has been associated with alcohol and tobacco dependence, substance abuse and myoclonic dystonia (Klein et al., 2000). D2 receptors are also known targets of antipsychotic drugs, such as Levodopa, Haloperidol and Promazine. These drugs are used to treat many neuropsychiatric disorders including schizophrenia, bipolar disorder, Parkinson's disease, anxiety and Tourette's disorder. Signalling through dopamine D2 receptors governs physiologic functions related to locomotion, hormone production, and drug abuse. *DRD2* is also associated with the melanin pathway and shows large differences in SNP occurrences within the European, Asian and African populations, having undergone recent positive selection for skin pigmentation (Wang et al., 2004; Lao et al., 2007) but has not been reported as being under positive selection along the human or hominid lineages. In our analysis, the *DRD2* gene displayed a signal for positive selection along the hominid branch ($p < 0.05$) suggesting divergence between human and animal models, with an ω of 8.9 for 0.8% of the sites along the protein.

The serotonin receptor 1D, encoded by the gene *HTR1D*, is also a target for many marketed drugs such as Rizatriptan, Tegaserod, Almotriptan and Naratriptan among others, with indications for migraine, schizophrenia and inflammatory bowel disorder. Serotonin is also one of the neurotransmitters involved in the aetiology of attention-deficit hyperactivity disorder and autism, with the release of serotonin in the brain being regulated by serotonin 1D autoreceptors (Coutinho et al., 2007). Positive selection detected along the hominid lineage in the *HTR1D* gene maybe linked with increased cognition in primates compared to rodents.

In this chapter, I relate genes predicted to have changed function during mammalian evolution from the genome scan to the diseases known to show biomedical differences between humans and model organisms. These genes may be causative of the phenotypic disease differences between species and are promising targets for therapeutic intervention. I also attempt to confirm and pinpoint the lineage in which positive selection signal occurred in the three known drug targets, *CNR1*, *DRD2* and *HTR1D*. These genes were investigated further in order to identify residues under positive selection that play a role in ligand-binding or activity modulation.

5.2 METHODS

5.2.1 Comparison with OMIM genes

Data from the Online Mendelian Inheritance in Man (OMIM) database (http://www.ncbi.nlm.nih.gov/omim), a catalogue of genetic diseases, was used to investigate whether genes under positive selection were over-represented among genes associated with disease. The OMIM Morbid map, an alphabetical list of disorders and their cytogenetic locations, was downloaded on 2nd May 2007. Gene names were mapped to Entrez Gene IDs using ENSEMBL BioMart (http://www.ensembl.org/biomart/) and batch searching of NCBI (http://www.ncbi.nlm.nih.gov/).

5.2.2 Comparison with disease categories

To investigate if genes from my dataset of PSGs were enriched among genes of a particular disease category, an ontology of major disease groups was created. Each gene in OMIM and its related medical term were placed in one of 19 specific groups based on the anatomical system affected by that disease. OMIM disease terms were mapped to MeSH terms (Medical Subject Headings), with identical terms being used whenever possible. When an identical term was not available a higher MeSH term which superseded the OMIM term was used. The MeSH terms were then mapped to the MeSH ontology to find a higher MeSH term. These were finally collapsed into one of 19 broad categories (Cardiovascular Diseases; Congenital, Hereditary, and Neonatal Diseases and Abnormalities; Digestive System Diseases; Endocrine System Diseases; Hemic and Lymphatic Diseases; Immune System Diseases; Metabolic Diseases; Muscular Diseases; Neoplasms; Nervous System Diseases; Psychiatric Disease;

Reproductive and Sexual Disease; Respiratory Tract Diseases; Skeletal Diseases; Sensory Disease; Skin Diseases; Tooth Diseases; Trait; Urologic Diseases). When a MeSH term was not available for a particular OMIM disease term, the ontological term was assigned by manual judgement after inspection of the OMIM record or another online resource (Pubmed, eMedicine, Encyclopaedia of Genetic Diseases or The Office of Rare Diseases). For an OMIM term, if abnormalities were seen in multiple systems (e.g. Down syndrome), the higher term, Congenital, was used. When a term did not seem to be a disease or disorder (e.g. blood group, hair colour) the higher term Trait was used. Terms describing oncology were mapped to neoplasms. When several classifications were available the one likely to be affected by selection pressure was used. For example, a failure of metabolism that had a neurological clinical presentation was categorised as a neurological disorder.

5.2.3 Site model analyses

To confirm and isolate the location of positive selection in the genes that are commonly employed as drugs targets, *CNR1*, *DRD2* and *HTR1D*, two likelihood ratio tests were performed using the site models described in Section 1.5.2. In the first test, the null model M1a (nearly neutral) was compared with the M2a model (selection). The second test compares the model M7 (beta) with the model M8 (beta & ω) (Yang and Swanson, 2002). As described for the branch-site model analyses (see Section 2.4.2) for both tests values for branch-lengths and kappa are estimated from the M0 model. The test statistics obtained for each gene were compared to a chi square distribution with 2 degrees of freedom using critical values of 5.99, 9.21 and 13.82 at *p* values of 0.05, 0.01 and 0.001, respectively.

Orthologous sequences from other species were downloaded from GenBank (http://www.ncbi.nlm.nih.gov/Genbank/) for each gene. The accession numbers of the *CNR1* sequences used were AAG37763.1 (crested gibbon), AAG37762.1 (macaque), AAG37761.1 (spider monkey), AAG37760.1 (Western tarsier), AAG37764.1 (Goeldi's marmoset) and AAG37759.1 (lemur). These sequences were added to the *CNR1* alignment and an unrooted tree files were created using a standard mammalian species tree (Murphy et al., 2001) (((((Human, Chimpanzee), Gibbon), Macaque), (((Spider monkey, Western tarsier), Marmoset), Lemur)), (Mouse, Rat), Dog).

The accession numbers for the *DRD2* sequences used were XP_001085449.1 (macaque), ABA62305.1 (gorilla), ABA62309.1 (bonobo), ABA62306.1 (orangutan) and ABA62304.1 (gibbon). The tree topology based on a standard mammalian species tree was (((((((Human, Chimpanzee), Bonobo), Gorilla), Orangutan), Gibbon), Macaque), (Mouse, Rat), Dog).

The accession numbers for the *HTR1D* sequences used were XM_001102386.1 (macaque) and NM_214158.1 (pig) and the tree used with the additional two species was (((Human, Chimpanzee), Macaque), (Mouse, Rat), Dog, Pig).

5.3 RESULTS

5.3.1 OMIM is enriched for positively selected genes

In order to determine if our dataset of PSGs was significantly enriched for disease genes, we examined genes that caused or were linked to human diseases as defined by OMIM. Of the 3079 genes used in our analysis, 469 genes were associated with a disease term in OMIM. Of the 511 PSGs from result set C from all seven lineages, 99 genes were coupled with a disease term in OMIM (Table 5.1). The numbers were slightly higher for result set B as 146 genes were associated with a disease term in OMIM. A test based on the binomial distribution showed that there was a significant link between PSGs and disease (p = 0.0067). Multiple testing corrections were not applied as the PSGs from each lineage showed little data overlap. While PSGs along the murid lineage were significantly over-represented in OMIM (p = 0.0087), PSGs along the human, chimp or hominid lineages did not display any over-representation (p < 0.05). The enrichment could be due to how the 3079 were chosen was we have only selected genes that have 1:1 orthology to human in the other four species.

Table 5.1 Correlation of positively selected genes with genes in OMIM

Result set B				
	No. of PSGs	**No. of PSGs in OMIM**	**Expected**	***p* value**
Human	69	12	11	0.3572
Chimp	354	61	54	0.1650
Hominid	49	10	8	0.2043
Mouse	121	16	18	-
Rat	155	33	27	**0.0272***
Murid	86	24	13	**0.0019***
Dog	162	30	25	0.1462
All	775	146	118	**0.0037***

Result set C				
	No. of PSGs	**No. of PSGs in OMIM**	**Expected**	***p* value**
Human	54	8	8	0.5919
Chimp	162	26	25	0.4190
Hominid	56	13	9	0.0753
Mouse	65	11	10	0.4032
Rat	89	18	14	0.1242
Murid	81	21	12	**0.0087***
Dog	97	21	15	0.0577
All	511	99	78	**0.0067***

Note: The *p* value is from a binomial test to look for over-representation of PSGs within OMIM, using a probability of a gene being in OMIM as 469/3079. The mouse lineage in result set B was not analysed as the observed value was less than the expected value.
*, $p < 0.01$

5.3.2 Comparison of PSGs with major disease categories

Since the OMIM database contains information for a large variety of diseases, each disease was placed in one of 19 groups based on the anatomical system affected (see Methods). 408 genes were associated with one or more higher disease terms, giving 515 gene-disease terms. Table 5.2 presents the numbers of PSGs from result set C from each lineage found in each disease category. PSGs in each lineage had very few genes in common, nor did genes in each disease category show much overlap so multiple testing correction was not carried out. In the mouse, rat, murid and dog lineages, PSGs showed enrichment in some of the disease categories (Figure 5.1). For example, genes that cause or are involved in skin diseases were over-represented in both the mouse and murid lineages. Moreover, mouse PSGs and PSGs from all seven lineages were augmented among genes associated with reproductive/sexual diseases (p = 0.026 and 0.0357, respectively). The individual genes found in each of the over-represented disease categories are described in Table 5.3. In contrast, PSGs along the chimpanzee and human lineages were not enriched in any of the specific disease categories.

Figure 5.1 Correlations of PSGs with Disease Ontologies

Tree depicting lineages which had an excess of PSGs in the 19 major disease categories. Only genes from non-primate species were over-represented in specific disease classes.

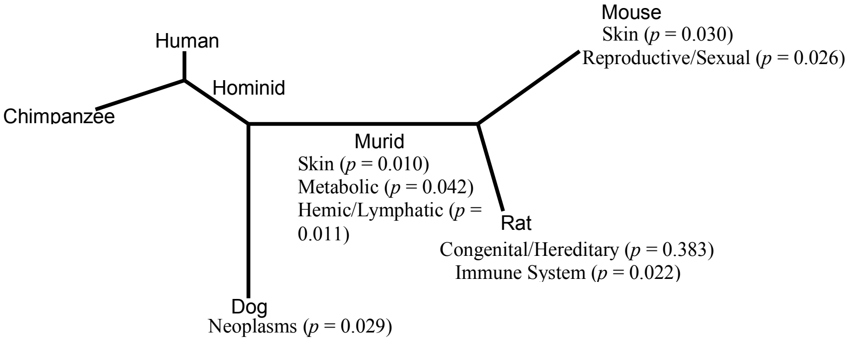

Mouse
Skin (p = 0.030)
Reproductive/Sexual (p = 0.026)
Human
Hominid
Chimpanzee
Murid
Skin (p = 0.010)
Metabolic (p = 0.042)
Hemic/Lymphatic (p = 0.011)
Rat
Congenital/Hereditary (p = 0.383)
Immune System (p = 0.022)
Dog
Neoplasms (p = 0.029)

Table 5.2 ***p* values from binomial test to look for over-representation of positively selected genes in 19 major disease categories**

	Muscular Diseases	Skin Diseases	Tooth Diseases	Congenital, Hereditary, and Neonatal Diseases and Abnormalities	Neoplasms	Sensory Disease	Urologic Diseases	Digestive System Diseases	none	Reproductive and Sexual Disease	Respiratory Tract Diseases	Skeletal Diseases	Metabolic Diseases	Nervous System Diseases	Psychiatric Disease	Cardiovascular Diseases	Immune System Diseases	Endocrine System Diseases	Hemic and Lymphatic Diseases	Trait
n	16	32	4	26	37	47	10	10	38	12	6	44	50	47	8	18	37	28	33	12
Human																				
m	1	1	0	1	1	1	0	1	2	1	0	2	0	1	0	0	0	1	0	1
p	0.25	0.43		0.36	0.5	0.56		0.16	0.14	0.19		0.18		0.56				0.40		0.19
Chimp																				
m	1	2	0	2	3	2	0	0	2	1	0	3	2	2	0	1	2	2	2	1
p	0.58	0.50		0.40	0.3 1				0.60	0.48		0.41				0.62	0.59	0.44	0.52	0.48
Hominid																				
m	0	0	0	0	1	1	1	0	1	0	1	2	1	1	0	1	2	1	1	1
p					0.5	0.58	0.17		0.50		0.10	0.19	0.60	0.58		0.28	0.15	0.40	0.45	0.20
Mouse																				
m	0	3	1	0	0	1	0	0	0	2	0	0	2	1	0	0	1	1	0	0
p		0.02	0.08			0.63				0.03			0.29	0.63			0.55	0.45		
Rat																				
m	0	2	0	3	1	2	0	0	1	1	0	3	0	1	1	1	4	0	1	0
p		0.23		0.04		0.40				0.30		0.13			0.21	0.41	0.02		0.62	
Murid																				
m	0	4	1	0	0	1	1	0	1	1	1	1	4	3	0	0	1	0	4	0
p		0.01	0.10			0.71	0.23		0.64	0.27	0.15	0.70	0.04	0.12			0.63		0.01	
Dog																				
m	2	0	1	0	4	2	1	0	2	0	0	3	3	0	0	1	1	2	3	0
p	0.09		0.12		0.1	0.44	0.27		0.34			0.16	0.21			0.44		0.22	0.09	

Note: Number of genes in 3079 genes in disease category (n), number of PSGs from each lineage in disease category (m) and *p* values from a binomial test to look for over-representation of genes under selection in these disease categories.

Table 5.3 Disease categories over-represented by PSGs

Lineage	Gene Name	Chromosomal Location	Disease	Biological Process
MOUSE				
	Skin Diseases			
	STS	Xp22.32	Ichthyosis, X-linked	Phospholipid metabolism; Sulfur metabolism
	HLA-DRB1	6p21.3	Pemphigoid, susceptibility to	MHCII-mediated immunity
	KRT2	12q11-q13	Ichthyosis bullosa of Siemens	Ectoderm development; Cell structure
	Reproductive Diseases			
	STS	Xp22.32	Placental steroid sulfatase deficiency	Phospholipid metabolism; Sulfur metabolism
	SYCP3	12q	Azoospermia due to perturbations of meiosis	Meiosis
RAT				
	Congenital, Hereditary, and Neonatal Diseases and Abnormalities			
	ATRX	Xq13.1-q21.1	Smith-Fineman-Myers syndrome	mRNA transcription regulation
	PRSS1	7q32-qter	Trypsinogen deficiency	Proteolysis
	UBR1	15q13	Johanson-Blizzard syndrome	Proteolysis
	Immune System Diseases			
	C3	19p13.3-p13.2	C3 deficiency	Complement-mediated immunity
	C8B	1p32	C8 deficiency, type II	Complement-mediated immunity
	CFD	19p13.3	Complement factor D deficiency	Proteolysis; Complement-mediated immunity
	ITGB2	21q22.3	Leukocyte adhesion deficiency	Cell adhesion-mediated signalling; Extracellular matrix protein-mediated signalling; Cell adhesion
MURID				
	Skin Diseases			
	HLA-C	6p21.3	Psoriasis, early onset, susceptibility to	MHCI-mediated immunity
	KRT2	12q11-q13	Ichthyosis bullosa of Siemens	Ectoderm development; Cell structure
	LAMC2	1q25-q31	Epidermolysis bullosa, generalized atrophic benign; Epidermolysis bullosa, Herlitz junctional	Cell adhesion-mediated signalling; Extracellular matrix protein-mediated signalling; Cell adhesion
	ENAM	4q13.3	Hypoplastic enamel pitting, localized	Developmental processes
	Metabolic Diseases			
	LYZ	12q15	Amyloidosis, renal	Carbohydrate metabolism; Stress response
	PNLIP	10q26.1	Pancreatic lipase deficiency	Lipid metabolism
	SCNN1G	16p12	Pseudohypoaldosteronism, type I	Cation transport; Taste; Regulation of vasoconstriction, dilation
	SLC34A3	9q34	Hypophosphatemic rickets with hypercalciuria	Phosphate transport; Cation transport; Other homeostasis activities
	Hemic and Lymphatic Diseases			
	F5	1q23	Hemorrhagic diathesis due to factor V deficiency	Proteolysis; Signal transduction; Blood clotting
	GP1BA	17pter-p12	von Willebrand disease, platelet-type	Other receptor mediated signalling pathway; Developmental processes
	HBD	11p15.5	Thalassemia due to Hb Lepore	Transport; Blood circulation and gas exchange
	SPTA1	1q21	Elliptocytosis	Biological process unclassified
DOG				
	Neoplasms			
	BCL3	19q13.1-q13.2	Leukemia/lymphoma, B-cell	mRNA transcription regulation

SLC22A18	11p15.5	Rhabdomyosarcoma	Small molecule transport; Other transport
PRF1	10q22	Lymphoma, non-Hodgkin	Immunity and defence
TAL1	1p32	Leukemia, T-cell acute lymphocytic	mRNA transcription regulation; Oncogene

5.3.3 Confirmation of selection in genes involved in drug discovery

Cannabinoid receptor 1

To further confirm the presence of positive selection along the rodent lineage and identify residues in the protein inferred to be under positive selection pressure, six primate sequences (crested gibbon, macaque, spider monkey, Western tarsier, Goeldi’s marmoset, and lemur) were added to the alignment (see Section 5.2.3). An 11-species tree was created and the branch-site model analysis rerun. Even with the addition of more primate sequences, the test of the murid branch resulted in a strong signal for positive selection. The Bayes empirical Bayes analysis inferred the residues 76 and 77, both of which code for threonine, to be under positive selection with posterior probabilities greater than 0.95.

A site model analysis was also performed to consider if selection pressure had occurred over the entire evolutionary history of the gene. Of the two likelihood-ratio tests which were performed, only the test of M8 versus M7 was significant ($p < 0.05$). The Bayes empirical Bayes procedure again inferred residues 76 and 77 to be under positive selection. The residues 76 and 77 are glutamine and valine in the primate lineages. Glutamine is a hydrophilic, polar amino acid and valine is a hydrophobic, aliphatic residue whereas threonine is a hydrophobic polar residue.

Dopamine receptor 2

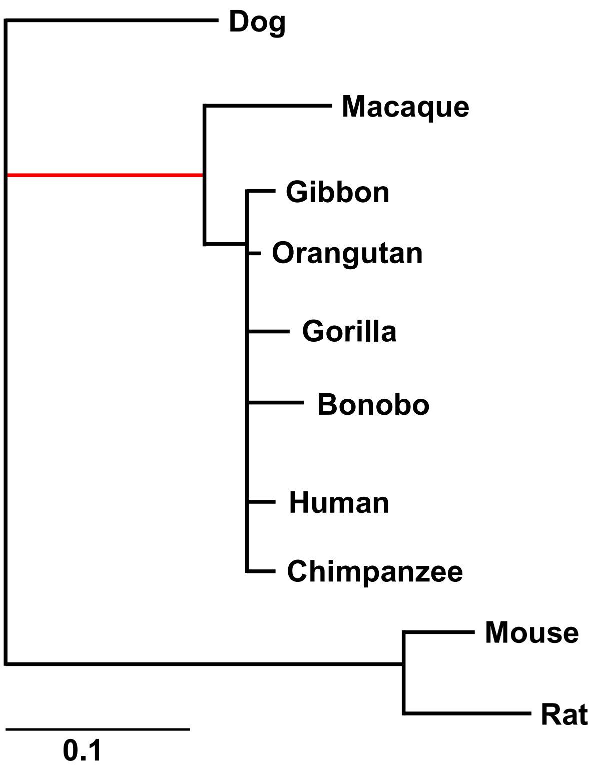
In order to identify the exact split in the hominid lineage which had been subject to positive selection in the evolution of the *DRD2* gene and also to ascertain positively selected sites within the open reading frame, *DRD2* sequences from five other primates (see Section 5.2.3 for accession numbers for macaque, gibbon, orangutan, gorilla and bonobo sequences) were added to the existing alignment and the branch-site analysis repeated.

All the lineages leading to the primates and primate ancestors were tested in turn to determine which branches were under positive selection. The branch representing the root of the primate lineage (Figure 5.2) was significant at $p < 0.05$, with an ω of 13.8. Residue 46 in the alignment was significant with a posterior probability of 0.892. Two likelihood-ratio tests using the site models, M2a-M1a and M8-M7 did not result in any significant signals for positive selection.

Serotonin receptor 1D

The *HTR1D* gene was also found to be under positive selection along the hominid branch suggesting divergence between human and animal models. To investigate the selection signal from the gene *HTR1D*, orthologues from macaque, marmoset, orangutan and pig sequences were added to the existing alignment. However, the branch-site analysis with the additional sequences resulted in the alternate model failing to converge. Several permutations of the run parameters including fix_blength = 1 and method = 0 and numerous repetitions of the branch-site run did not facilitate the alternate model to converge. Use of the gene tree as opposed to the standard species tree resulted in the runs converging but the test for positive selection of the *HTR1D* gene along the hominid lineage was no longer significant. The gene tree differed from the species tree in that in the gene tree the dog and pig lineages were in the same clade whereas in the species tree they were in separate clades. The gene tree is incorrect in having the pig and dog lineages in the same clade: the pig and dog belong to different orders in mammalian phylogeny. The site model analysis did not generate any significant results either.

Figure 5.2 Tree showing selection along primate ancestor in *DRD2*



5.4 DISCUSSION

Overall, I observed that PSGs from all lineages were over-represented among genes found in OMIM. Yet in contrast to the findings of Clark et al. (2003) along the human lineage, PSGs were not seen to display any over-representation in

OMIM, nor in any of the disease categories ($p < 0.05$). However the set of genes they found to be significant were the results of a branch test (model 2) and not all of these genes had $\omega > 1$. A branch test tends to average ω values across the gene and across the lineage, so genes that show episodic changes in ω at particular sites along the gene will not be found to be significant. The initial set of 7645 genes tested by Clark et al. and our primary data set of 3079 genes share only 851 genes and since the type of analyses done by both groups are different, the results cannot be directly compared.

My findings, however, were consistent with other recent studies that found no significant associations (CSAC, 2005) or only marginal associations (Bakewell et al., 2007) between PSGs and human diseases. Note that the OMIM database only contains genes exhibiting direct Mendelian disease inheritance but not the genes involved in the much more common, polygenic human disorders.

PSGs implicated in diseases with biomedical differences between mammals

Initial examination of the individual PSGs along the human and hominid lineages (Table 5.4) that had disease associations in OMIM did not reveal any patterns or major disease implications. However a closer look at specific diseases that show biomedical differences between mammals revealed positively selected genes that are implicated in such diseases (Table 5.5). These are diseases that show differences in severity and frequency between humans and great apes, as experienced by primate centres and zoos over the last century.

Table 5.4 Human and hominid PSGs associated with disease in OMIM

Gene Name	Chromosomal Location	Disease	Disease Type	Biological Process
Human				
CACNA1A	19p13.2-p13.1	Cerebellar ataxia	Nervous System	Cation transport; Neurotransmitter release; Muscle contraction
CACNA1S	1q32	Thyrotoxic periodic paralysis	Muscular	Cation transport; Neurotransmitter release; Muscle contraction
		Malignant hyperthermia	none	
COL11A1	1p21	Marshall syndrome	Skeletal	Mesoderm development
EDNRB	13q22	ABCD syndrome	Congenital, Hereditary, and Neonatal	G-protein mediated signalling; Calcium mediated signalling; Muscle contraction;
		Hirschsprung disease	Digestive System	
NR5A1	9q33	Adrenocortical insufficiency without ovarian defect	Endocrine System	mRNA transcription regulation; Signal transduction; Developmental processes
		Sex reversal, XY, with adrenal failure	Reproductive and Sexual	
MC1R	16q24.3	Melanoma	Neoplasms	G-protein mediated signalling; Vision
		Analgesia from kappa-opioid receptor agonist, female-specific	none	
		Oculocutaneous albinism, type II, modifier of	Sensory	
		UV-induced skin damage,	Skin	
		Blond/light brown hair and/or fair skin	Trait	
UMPS	3q13	Orotica aciduria	Skeletal	Biological process unclassified
Hominid				
ADRB2	5q31-q32	Asthma, nocturnal	Immune System	G-protein mediated signalling; Calcium mediated signalling;
		Beta-adrenoreceptor agonist	none	
APOE	19q13.2	Myocardial infarction	Cardiovascular	Lipid and fatty acid transport; Transport
		Sea-blue histiocyte disease	Metabolic	
		Macular degeneration, age-related	Sensory	
C1QA	1p36.3-p34.1	C1q deficiency, type A	Immune System	Complement-mediated immunity
COL4A4	2q35-q37	Alport syndrome	Urologic	Biological process unclassified
COL11A1	1p21	Marshall syndrome	Skeletal	Mesoderm development
COMP	19p13.1	Pseudoachondroplasia	Endocrine System	Blood clotting; Other developmental process
		Epiphyseal dysplasia	Skeletal	
F5	1q23	Hemorrhagic diathesis	Hemic and Lymphatic	Proteolysis; Signal transduction; Blood clotting
MSH2	2p22-p21	Cafe-au-lait spots, multiple, with leukaemia	Neoplasms	DNA repair; Meiosis
TH	11p15.5	Segawa syndrome	Nervous System	Other amino acid metabolism; Signal transduction
TXNDC3	7p14.1	Ciliary dyskinesia	Respiratory Tract	Pyrimidine metabolism
ABCC11	16q12.1	Earwax, wet/dry	Trait	Small molecule transport; Extracellular transport and import; Detoxification

Table 5.5 Differences between humans and apes in incidence or severity of medical conditions and PSGs associated with them

Medical condition	Humans	Great apes	PSG
Definite			
HIV progression to AIDS	Common	Very rare	*HIVEP3*
Hepatitis B/C late complications	Moderate to severe	Mild	*NR5A1*
P. falciparum malaria	Susceptible	Resistant	
Myocardial infarction	Common	Very rare	
Endemic infectious retroviruses	Rare	Common	
Influenza A symptomatology	Moderate to severe	Mild	
Probable			
Menopause	Universal	Rare?	*NR5A1*
Alzheimer's disease pathology	Complete	No neurofibrillary tangles	*APOE*
Epithelial cancers	Common	Rare?	*MC1R, EDNRB, ALPPL2, GIPC2, MSH2, ABCC11 TFPT, ZNF384*
Atherosclerotic strokes	Common	Rare?	
Hydatiform molar pregnancy	Common	Rare?	
Possible			
Rheumatoid arthritis	Common	Rare?	
Endometriosis	Common	Rare?	*NR5A1*
Toxemia of pregnancy	Common	Rare?	
Early fetal wastage (aneuploidy)	Common	Rare?	*UMPS*
Bronchial asthma	Common	Rare?	*ADRB2, TXNDC3*
Autoimmune diseases	Common	Rare?	*CENP-B*
Major psychoses	Common	Rare?	*PIK3C2G, XRCC1, GFRalpha3*

Note: Adapted from Varki and Altheide, 2005

Below we illustrate how some of the human and hominid PSGs identified in our study are linked to medical conditions described as being more prevalent or having increasing severity in humans compared to apes (Olson and Varki, 2003; Varki and Altheide, 2005).

Epithelial cancers

Human cancers are thought to be the cause of over 20% of deaths in modern human populations whereas among non-human primates, the rates are as low as 2-4% (Varki, 2000). Although this may be partly attributed to carcinogenic factors in the lifestyles of modern humans and differences in life expectancy, there are many intriguing lines of evidence to suggest that another overwhelming factor is the presence of susceptibility genes in human (McClure, 1973; Seibold and Wolf, 1973; Beniashvili, 1989; Crespi and Summers, 2006; Puente et al., 2006; Coggins, 2007; Kehrer-Sawatzki and Cooper, 2007). Among the human lineage PSGs, a number of genes have been implicated in the development of epithelial cancers:

- **MC1R** (melanocortin-1 receptor) modulates the quantity and type of melanin synthesised in melanocytes by acting as the receptor protein for the melanocyte-stimulating hormone (MSH). With mutations in this gene associated with melanomas (Valverde et al., 1996), this receptor is a major determining factor in sun sensitivity. In other species, mutations in this gene have been associated with coat colour variation, with changes being driven by positive selection (Mundy and Kelly, 2003). Residue 186 (threonine in human, valine in other species), which showed a high posterior probability for being under positive selection, is part of an extracellular loop in the transmembrane protein so it could potentially be involved in ligand recognition or binding.

- The G-protein coupled receptor **EDNRB** (endothelin type-B receptor) and its physiological ligand, endothelin 3, are thought to play key roles in the development of melanocytes and other neural crest lineages (McCallion and Chakravarti, 2001). EDNRB promotes early expansion and migration of melanocyte precursors and delays their differentiation. EDNRB is greatly enhanced during the transformation of normal melanocytes to melanoma cells where it is thought to play a role in the associated loss of differentiation seen in melanoma cells (Lahav, 2005).
- The presence of the ***ALPPL2*** gene product, an alkaline phosphatase isoenzyme, has been shown to increase the potential of premeiotic male germ cells to malignant transformation. Increased promoter activity of this gene was seen in the process of tumour progression. *ALPPL2* has now been confirmed as a marker for testicular germ cell tumours (Tascou et al., 2001).
- ***GIPC2*** mRNAs is expressed in cells derived from a diffuse-type of gastric cancer, and also shows increased expression in several cases of primary gastric cancer (Katoh, 2002). The PDZ domain of the GIPC2 protein interacts with several genes that are involved in modulation of growth factor signalling and cell adhesion (e.g. *FZD3*, *IGF-1* and NTRK1). Thus GIPC2 may play key roles in carcinogenesis and embryogenesis.

In the hominid lineage, several PSGs have also been implicated in epithelial cancer development suggesting differences in cancer disease processes between hominids and other mammals:

- ***MSH2*** is a DNA mismatch-repair gene that was identified as a common locus in which germline mutations cause hereditary nonpolyposis colon cancer (HNPCC) (Yoon et al., 2008). As deficiencies in any DNA repair gene would potentially increase cancer risk, the whole group is of interest in investigation of species differences in cancer prevalence. I found that genes that are involved in DNA repair and nucleotide metabolism were over-represented for PSGs along the chimpanzee and human lineages respectively (Figure 3.2). Enrichment of PSGs within the nucleotide metabolism category has also been reported previously (Holbrook and Sanseau, 2007).
- The ***ABCC11*** (ABC-binding cassette, subfamily C, member 11) gene product is highly expressed in breast cancer compared to normal tissue. *ABCC11* is regulated by ERα, which mediates the tumour promoting effects of estrogens in breast cancer (Laganiere et al., 2005). The allele with Arg-184 is responsible for the dry earwax phenotype in some human populations. This gene participates in physiological processes involving bile acids, conjugated steroids and cyclic nucleotides and enhances the cellular extrusion of cAMP and cGMP.
- ***TFPT*** (TCF3 (E2A) fusion partner (in childhood Leukaemia)) and ***ZNF384*** (zinc finger protein 384) are listed in Futreal et al. (2004) as genes that are mutated in cancer and causally implicated in oncogenesis.

Alzheimer’s disease

A gene ***APOE***, implicated in Alzheimer’s disease (Hanlon and Rubinsztein, 1995; Mahley and Huang, 1999), was under positive selection along the hominid lineage. Selection for functional changes of the *APOE* gene in the hominid lineage could be related to either its role in neurological development or in lipid metabolism. Of the eight amino acids found to be under positive selection in this study, four are present in the lipid-binding carboxyl terminus.

Suggestion that there are species differences in Alzheimer’s disease between humans and other mammalian species comes from the observation that the complete pathological lesions including the neurofibrillary tangles associated with human Alzheimer’s disease have never been observed in the brains of elderly chimpanzees or rhesus monkeys (Gearing et al., 1994; Gearing et al., 1996) or elephants (Cole and Neal, 1990). Also, transgenic mouse models of Alzheimer's disease that presented β-amyloid neuropathology do not exhibit the cognitive decline at the first appearance of amyloid plaques seen in humans (Howlett et al., 2004). Finally and intriguingly, mammals other than humans seem to have just one allelic form of *APOE*, the E4 allele, which in humans predisposes carriers to a much higher risk of Alzheimer’s disease (Strittmatter et al., 1993). In humans, polymorphisms at two sites within the *APOE* gene result in three isoforms: E2, E3, and E4. The most common isoform, E3, has a cysteine at position 112 and arginine at position 158; isoform E2 has cysteines at both sites, whereas E4 has arginines at both sites (Hanlon and Rubinsztein, 1995). The APOE4 allele is highly associated with late-onset Alzheimer’s disease (Strittmatter et al., 1993) and also with relatively elevated LDL-cholesterol levels compared to other genotypes (Mahley, 1988).

Schizophrenia

Neurological studies have shown that brain areas differentially dysregulated in schizophrenia are also subject to the most evolutionary change in the human lineage (Brune, 2004). A number of PSGs along the human lineage are associated with schizophrenia:

- SNPs in the gene ***PIK3C2G*** (phosphoinositide-3-kinase) have recently been shown to be associated with schizophrenia (Jungerius et al., 2007). This gene is related to the phosphatidylinositol signalling pathway, and thus is a probable candidate for schizophrenia and bipolar disorder (Stopkova et al., 2003).
- Another candidate for chronic schizophrenia is the Q399 allele of the **XRCC1** protein, which plays a role in base excision repair (Saadat et al., 2008). The pathophysiology of schizophrenia is associated with an increased susceptibility to apoptosis. Mutations in *XRCC1* may cause DNA damage, which, if detected, cause apoptosis regulators to arrest cell cycle progression.

Other cognitive disorders

Equally detected under positive selection pressure along the human lineage was the gene ***GFRalpha3***, a receptor for artemin. Artemin is a member of the glial cell line-derived neurotrophic factor (GDNF) family of ligands. This gene acts as a signalling factor regulating the development and maintenance of many sympathetic neuronal populations (Wang et al., 2006). In particular, along with other GDNF family members, artemin plays a role in synaptic plasticity, a mechanism thought to be central to memory (Kim and Linden, 2007).

Autoimmune diseases

Autoimmune diseases are rare in non-human primates whereas they are relatively common in humans (Varki, 2000). **CENP-B** is one of three centromere DNA binding proteins that are present in centromere heterochromatin throughout the cell cycle. Autoantibodies to these proteins are often seen in patients with autoimmune diseases, such as limited systemic sclerosis, systemic lupus erythematosus, and rheumatoid arthritis (Russo et al., 2000). The positive selection pressure acting on this gene during human evolution is consistent with experimental results that antigenic specificity in the C-terminus of this protein is species-specific (Sugimoto et al., 1992).

Recurrent Miscarriage

Varki (2000) postulates that the high rate of early foetal wastage seen in humans could be caused by fertilisation of deteriorating eggs which contain gross chromosomal and genetic abnormalities. This occurs when fertilisation takes place at sub-optimal times as a result of the absence of external signs of ovulation in human females. It has long been known that in cattle a deficiency in uridine monophosphate synthetase (DUMPS) is a mono-genic autosomal recessive disorder that results in early embryogenic death of homozygous offspring (Schwenger et al., 1993; Ghanem et al., 2006). In humans, recessive mutations in the orthologous gene, ***UMPS***, are known to cause a rare disorder called orotic aciduria which is linked to congenital abnormality in progeny (Harden and Robinson, 1987; Bensen et al., 1991). This gene has undergone selection only in the human lineage and further experimental evidence needs to be obtained to ascertain whether this can be related to embryonic death in humans and hence contribute to a higher rate of human foetal wastage in comparison with non-human primates.

Bronchial asthma

Two genes that were positively selected along the hominid lineage are associated with respiratory diseases. These genes were:

- The β2-adrenergic receptor gene, ***ADRB2*** mediates bronchodilatation in response to exogenous and endogenous beta-adrenoceptor agonists. Point mutations and various polymorphic forms of this gene have been linked to nocturnal asthma, obesity and type 2 diabetes as well as individual differences to therapeutic drug responses. Residue 92 of the protein (alanine in hominids, serine in other species) is part of the second transmembrane subunit of the protein and was identified as being under positive selection.
- The other gene ***TXNDC3*** encodes a thioredoxin–nucleoside diphosphate kinase and is implicated in primary ciliary dyskinesia (PCD), a genetic condition characterized by chronic respiratory tract infections, left–right asymmetry randomization, and male infertility. This gene was positively selected on both the hominid and murid lineages.

Ataxia and Migraine

The calcium channel gene of type P/Q, ***CACNA1A***, was found to be under positive selection along the human lineage. *CACNA1A* is predominantly expressed in neuronal tissue, with Purkinje neurons in the cerebellum containing predominantly P-type voltage-sensitive calcium channels and the Q-type being the prominent calcium current in cerebellar granule cells. Within this gene, residue 27 (serine in human, alanine in other species) was shown to be under positive selection, and is in the cytoplasmic region, N-terminal to the first transmembrane domain. In humans, mutations in *CACNA1A* are associated with channelopathies, such as spinocerebellar ataxia 6 and episodic ataxia type 2 (Jen et al., 2007) as well as with more prevalent conditions such as familial hemiplegic migraine, dystonia, epilepsy, myasthenia and even intermittent coma (Jouvenceau et al., 2001). The benefits of enhanced CNS excitability may outweigh the risk of severe headache and disability, the symptoms of migraines (Loder, 2002). It could also be an artefact of design constraints in the brain resulting from imperfect interconnections between older and more recently evolved brain structures (Nesse and Williams, 1995).

Positive selection in genes affecting the regulation of other genes

Selection events on coding sequences may also have effects on gene expression regulation. One such gene is ***NR5A1***, the transcriptional regulator SF1 (steroidogenic factor 1), which had evolved under positive selection along the human lineage. The implications of functional divergence in human *NR5A1* are considerable as SF1 is an orphan nuclear receptor that plays an essential role in the development of the adrenal gland, testis, ovary, pituitary gonadotropes, and hypothalamus (Luo et al., 1999). Examination of gene interaction data suggests that this gene is involved in regulation of about 900 genes in the human genome (Kolchanov et al., 2002). Some examples of genes regulated by SF1 and implicated in diseases with biomedical differences between species are:

- Aromatase P450 is a gene which is expressed at very high levels in endometriosis (Yang et al., 2002), an estrogen-dependent disease affecting females of reproductive age. Endometriosis seems to be less common in non-human primates than in humans (Varki and Altheide, 2005). The regulation of aromatase P450 also has implications for menopause, a condition that is only seen in humans but has not been observed in long-lived captive non-human primates (Bellino and Wise, 2003).

- *SCARB1* (scavenger receptor B1) is associated with entry and progression of hepatitis B and C viruses (Grove et al., 2007). In humans, both hepatitis B and C trigger complications that are not seen as frequently in experimentally induced viral infections in chimpanzees (Makuwa et al., 2006).

Another transcription factor that showed signs of positive selection along the human lineage was ***HIVEP3*** (immunodeficiency virus type I enhancer binding protein 3). This gene belongs to a family of zinc-finger proteins whose functions include activating HIV gene expression by binding to the NF-kappaB motif of the HIV-1 long terminal repeat (Seeler et al., 1994). It is commonly known that HIV infection in chimpanzees does not progress to the level of medical complexity that is seen in human AIDS (Varki, 2000), where the virus proceeds to infect and destroy helper T-cells. In chimpanzees however, the virus lives in a benign relationship within the immune system.

Some regulatory elements of gene expression also showed evidence of positive selection along the human lineage. One is the ***MOV10*** gene (Moloney leukaemia virus 10, homolog), an RNA helicase contained in a multiprotein complex along with proteins of the 60S ribosome subunit. *MOV10* is associated with human RISC (RNA-induced silencing complex) (Chendrimada et al., 2007). RNA silencing or interference (RNAi) has been recently described as an important therapeutic application for modulating gene expression at the transcript level or for silencing disease-causing genes (Barnes et al., 2007; Federici and Boulis, 2007). Any functional changes in the *MOV10* gene due to selection may affect transcriptional control of multiple genes.

Genes associated with major disease categories

The total number of PSGs from all lineages was over-represented in the diseases involved in reproductive processes. The enhancement in this category might be expected because alleles of genes involved in reproduction are likely to be positively selected during speciation events and hence cause reproductive isolation between the two new species. PSGs along the human lineage do not show any over-representation in any individual category possibly because they are evenly distributed across all categories of diseases.

PSGs on all the non-primate lineages show an association with one or more disease categories perhaps because genes that have undergone positive selection in other lineages are prone to disease in humans (since we are only looking at human diseases). The correlation between PSGs along the murid lineage and disease genes has implications for rodent models in drug discovery. If genes have undergone positive selection along the murid lineage, and in the process have acquired new, different or additional functions, then their use as drug targets in animal models might not accurately predict drug responses in humans. An illustration of this is seen in the *KLF11* gene which codes for a pancreas-enriched Sp1-like transcription factor, which in my study had undergone positive selection in the mouse lineage. This gene does not cause any disease in mouse when mutated (Eppig et al., 2005) but in humans mutations in this gene causes maturity-onset diabetes of the young, type VII. The reverse hypothesis is that genes that have selective advantages in humans might be linked to murine diseases but this is harder to prove as data on the subject of naturally occurring murine diseases are scarce.

I found that PSGs from the mouse, rat and murid lineages were associated with human diseases (Figure 5.1, Table 5.3) but genes that had undergone positive selection in the primate lineages were not associated with disease. The skin disease genes which were enhanced among PSGs from the mouse and murid lineages are linked to several types of skin disorders. The skin is known for its functional integrity (Wehrli et al., 2000) since skin cells are generally more tolerant of mutations. Mutations which might be lethal in other organs cause severe skin diseases which are tolerated in humans.

Examples of such diseases include X-linked ichthyosis (continual and widespread scaling of the skin) which is associated with the gene *STS* (steroid sulfatase (microsomal), isozyme S). *STS*, which underwent adaptive evolution in the mouse lineage, is known to be pseudoautosomal in mouse (Keitges et al., 1985) and X-linked in man. The *STS* locus is in the pseudoautosomal segment of the X and Y mouse chromosomes but not in human. (Yen, 1998) suggested that a pericentric inversion of the Y chromosome occurred during primate evolution, disrupting the former pseudoautosomal arrangement of these genes. Several attempts to clone the mouse homolog of the *STS* gene have failed, suggesting a substantial divergence between these genes (Salido et al., 1996). This is evident in the positive selection result I observed for the *STS* gene along the mouse and murid lineages.

Another example of a serious but tolerable skin disease is ichthyosis bullosa of Siemens which is caused by mutations in the gene *KRT2* (keratin 1) (Kremer et al., 1994). Ichthyosis bullosa of Siemens is characterised by generalized reddening of the skin and widespread blistering. *KRT2*, which underwent positive selection along the mouse and murid lineages, is expressed in terminally differentiated epidermis and it is known that even conserved substitutions in the keratin gene affect the structure of the protein (Schweizer et al., 2006). Another murid PSG is *HLA-C (*major histocompatibility complex, class I, C) which has been implicated with psoriasis, a T cell-mediated autoimmune disorder, which affects a large proportion (2%) of the human population (Tiilikainen, 1980).

PSGs along the rat lineage also showed a strong link to human disease genes including several genes (*C3*, *C8B*, *CFD* and *ITGB2*) linked to immune diseases. A deficiency of the C8 protein causes recurrent neisserial infections, predominantly with meningococcus infection of rare serotypes, which suggests that this protein plays a role in immune system signalling (Stark et al., 1980). The protein product of the complement factor D gene (*CFD*) is part of the alternative complement pathway and converts complement factor B to its end product. Moreover it is also secreted by adipocytes into the bloodstream as a serine protease, adipsin, and has been found to be deficient in several animal models of obesity. The deficiency in adipsin was not seen in mouse models which were obese due to overfeeding (Flier et al., 1987).

Genes products utilised as common drug targets

Analysis of the gene *CNR1* revealed two residues, 76 and 77, to have evolved under positive selection pressure. These two residues are not known to be involved in ligand binding. However, the adjacent residue 78 is understood to be a potential attachment site for oligosaccharide N-acetylglucosamine (Andersson et al., 2003) and could, therefore, be an important target for post-translational modification. The two amino acids lie in the N-terminal region preceding the first transmembrane domain that begins at residue 118. It can be hypothesized that the substitution of residue 76 from a glutamine in primates, to a threonine (in murids), might have caused the protein to lose some binding power as glutamines are frequently involved in protein active or binding sites (Betts and Russell, 2003). Glutamines contain a polar side-chain which is suited for interactions with other polar or charged atoms as well as a hydroxyl group which can form hydrogen bonds with a variety of polar substrates. Additionally, the N-terminal region of the protein CB1 plays a key role in the efficiency at which the N-tail is translocated across the endoplasmic reticulum membrane and hence contributes to the stability of the protein in the cytosol (Andersson et al., 2003). This is turn would affect the amount of tissue and cell distribution of the CB1 protein.

The *DRD2* gene showed significant evidence of positive selection along the branch representing the root of the primate lineage (Figure 5.2). Residue 46 (alanine in primates, phenylalanine in non-primate species) is present in the first of seven transmembrane domains (http://www.expasy.org/uniprot/P14416; PDB id = 1I15) but is not known to be a ligand binding site. Interestingly, the gene that regulates the synthesis of dopamine, tyrosine hydroxylase (*TH*), was also found to have undergone positive selection in hominids in this study. When dopamine D2 receptors are activated, dopamine release is inhibited through regulation by *TH* (Lindgren et al., 2001). Pathway databases show that *DRD2* and *TH* interact together, confirming our previous finding (see Section 3.2.6) that positively selected genes have a tendency to interact together, as functional changes in one gene will drive adaptation in the other.

Furthermore, CB_1 cannabinoid (CB_1) and D_2 dopamine (D_2) receptors are known to couple to the G protein $G\alpha_{i/o}$. The concurrent activation of D_2 receptors and CB_1 receptors promotes functional CB_1 receptor coupling to $G\alpha_s$ resulting in elevation of intracellular cyclic AMP levels, presumably through $G\alpha_s$ (Jarrahian et al., 2004). The co-expression of the D_2 receptor with the CB_1 receptor is sufficient to switch the coupling of the CB_1 receptors from $G\alpha_{i/o}$ to $G\alpha_s$, thus providing another example of positively selected genes interacting together.

Summary: The utility of comparative methods in studies of human disease

I conclude that comparative evolutionary genomics has an important contribution to make to the study of mammalian disease, enabling identification of candidate genes for further *in vivo* investigation. Researchers traditionally see the biomedical differences between humans and model organisms as an obstacle to progress. However, I have shown that these differences also provide an opportunity when studied at the codon level. To take advantage of this opportunity, we need powerful computational evolutionary algorithms (such as those used in this study) and a robust approach to utilise the ever-expanding genomic sequence data. However, it is also necessary to obtain detailed accounts of the physiological differences in disease occurrence and symptomatology between species. Such data are currently sparse and thus it is important to collect observations on biomedical differences between species.

Understanding the evolutionary history of disease genes can also significantly impact the choice of pre-clinical animal models in the drug discovery process (Searls, 2003). The success rate in pharmaceutical pipelines remains low, one reason being the difficulties in successfully translating safety and efficacy studies from animal models to humans. Pre-clinical studies assume that drug targets in the experimental species and in humans are functionally equivalent, which is not always the case (Holbrook and Sanseau, 2007). In particular, animal models of neurodegenerative diseases have been shown not to have predictive validity in humans (Heemskerk et al., 2002). Studies of selection pressure during gene evolution can provide valuable information for the choice of animal models for drug target validation. Our results of PSGs in the five mammalian species serve as an informative resource that can be consulted prior to selecting appropriate animal models during drug target validation in the pharmaceutical industry.

Chapter 6

Analysis of Positive Selection in Nuclear Receptors

6.1 MOTIVATION AND OBJECTIVES

The previous chapter discussed a number of genes that regulate gene expression being under positive selection. Genes that are involved in transcription regulation and which respond to a wide variety of ligands are particularly interesting to study as functional changes in them would significantly impact the phenotype of the organism. One such superfamily is the nuclear receptors (NRs), which in humans consists of 48 genes with various roles in metabolic homeostasis, embryonic development, cell differentiation and detoxification (Laudet and Gronemeyer, 2002).

The genome scan described in Chapters 2 to 5 identified nuclear receptor genes subject to positive selection in the five species studied. These included the nuclear receptor, *NR5A1*, which was under positive selection in both the human and chimpanzee lineages. *NR5A1* regulates the expression of many genes that are involved in diseases with biomedical differences between species (see Section 5.4). Another NR of particular interest and one that warrants this further investigation is the pregnane X receptor (*NR1I2*) which was found to be under selection along the murid lineage in this work. Nuclear receptors, *NR2F2* and *NR2F6* were also found to be under positive selection along the chimpanzee and dog lineages, respectively. Other genes, *NR2F1*, *NR0B1* and *NR0B2* were also analysed but were not found to be under positive selection.

This chapter extends the analysis of variation in selection pressure in functionally distinct regions of transcription factors by an in-depth analysis of the 48 human nuclear receptors and their mammalian orthologues. The site and branch-site models were used to detect positive selection in the conserved DNA-binding domain and a variable ligand-binding domain present in these nuclear receptors.

6.2 BACKGROUND TO NUCLEAR RECEPTORS

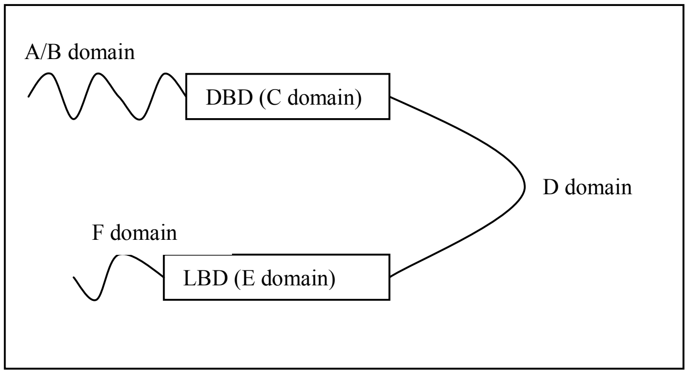
The nuclear receptor family consists of hormone receptors for thyroid hormones, retinoic acids, sex steroids (estrogen, progesterone and androgen), glucocorticoids, mineralcorticoids, vitamin D3, leukotrienes, prostaglandins (Escriva et al., 2000) and ‘orphan’ nuclear receptors, for which ligands have not yet been identified (Giguere, 1999). They are present in varying numbers in arthropods, vertebrates and nematodes as a result of periods of gene duplication and lineage-specific gene loss (Bertrand et al., 2004). Some examples of nuclear receptors are the peroxisome proliferator-activated receptors (PPARs) which exert direct effects on fat and carbohydrate metabolism and are major targets for therapeutic agents in diseases such as cholesterol disorders, diabetes mellitus and hyperlipidaemia. The peroxisome proliferator-activated receptor gamma (*PPARG*) receptor plays a role in adipogenesis and this receptor targets genes that mediate insulin sensitisation (Rosen and Spiegelman, 2001). The liver X receptor (*LXR*) functions as a cholesterol sensor and its close relative, the farnesoid X receptor (*FXR*), acts as a bile acid sensor and regulates an array of genes involved in bile acid metabolism.

6.2.1 Structure

All nuclear receptors share a common structure (Figure 6.1), consisting of a variable N-terminal A/B domain which contains at least one constitutively active transactivation region (AF-1). The A/B domain is followed by a well conserved

DNA-binding region (DBD, C-domain) (Olefsky, 2001). The DBD contains the P-box, a motif that binds DNA sequences containing the AGGTCA motif (Robinson-Rechavi et al., 2003), and is involved in the dimerisation of nuclear receptors. Adjacent to the DBD is a non-conserved hinge-like D domain which contains the nuclear localisation signal that sits posterior to the E-domain or ligand-binding domain (LBD). The LBD recognises specific hormonal and non-hormonal ligands. It has a conserved secondary structure of 12 alpha-helices and also contains another transactivation region (AF-2). A variable length F-domain whose function is not known lies at the C-terminus. The F-domain is absent in some NRs.

Figure 6.1 Schematic view of the functional domains in a nuclear receptor
Adapted from Olefsky, 2001.



6.2.2 Function

Nuclear receptors can act as homodimers and/or heterodimers and are thought to act in three steps (Figure 6.2) (Robinson-Rechavi et al., 2003):

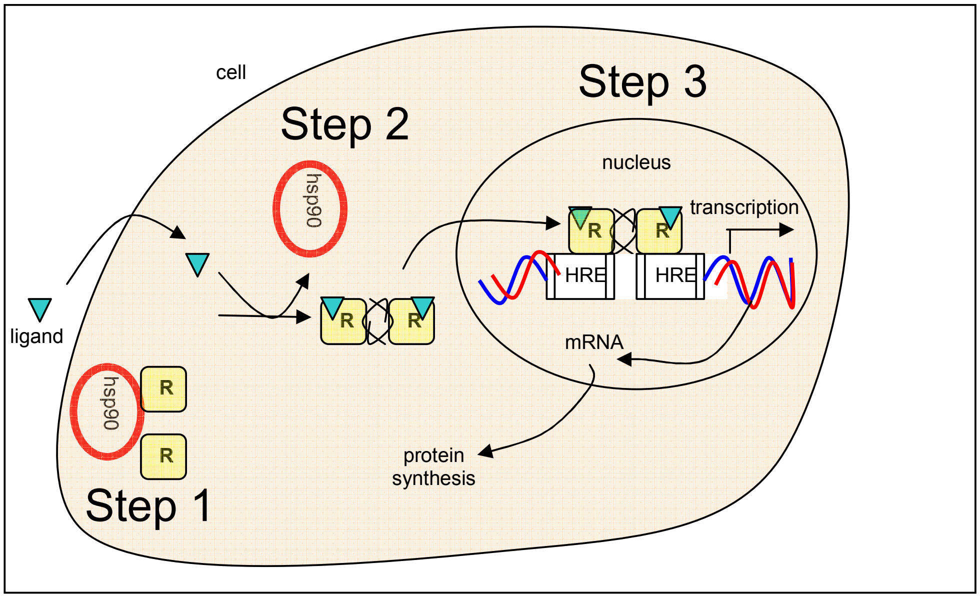
1) Repression – apo (unliganded)-nuclear receptor recruits corepressor complex with histone deacetylase activity (HDAC).

2) Depression – occurs after ligand binding which dissociates this complex and recruits a first coactivator complex with histone acetyltransferase (HAT) activity, which results in chromatin decondensation. The receptor translocates into the nucleus.

3) Transcription activation – the HAT complex dissociates and a second coactivator complex is assembled which establishes contact with the basal transcription machinery and results in transcription activation of the target gene (Moras and Gronemeyer, 1998).

The precise order of events is still debated and the mechanism varies among receptors.

Figure 6.2 Mechanism by which nuclear receptors bind to their co-activators



6.2.3 Evolutionary history

It is thought that the first NR was an ‘orphan’ receptor and the ability to bind to ligands was a function that was acquired during evolution (Escriva et al., 1997). Phylogenetic studies have shown that the first steroid receptor was an estrogen receptor, followed by a progesterone receptor (Thornton, 2001). The full complement of mammalian nuclear receptors evolved from these ancestral receptors by two large-scale genome duplications, one before the emergence of jawed vertebrates and one after (Escriva Garcia et al., 2003). Novel receptor-hormone pairs are created by gene duplication of receptors. The duplicated gene may gain a new function (neo-functionalisation) during a time of relaxed selection (Hurles, 2004). New receptors have been shown to evolve affinity for intermediates in a biosynthetic pathway in which the terminal ligand was the ligand of the parent receptor (Escriva et al., 1997). The human genome has been found to contain 48 nuclear receptors (chosen for analysis in this study), 47 nuclear receptors have been identified in the rat genome and 49 in mouse (Figure 6.3). It is also known that genome duplication events at the origin of ray-finned fishes gave rise to the expanded family (68 genes) found in the pufferfish genome.

6.2.4 Nomenclature

Nuclear receptors fall into 7 subfamilies (NR0-6) (Nuclear Receptors Nomenclature Committee, 1999) (Table 6.1).The two members of the NR0 family, *DAX-1* (NR0B1; dosage-sensitive sex and AHC critical region on the X-chromosome) and *SHP* (NR0B2; small heterodimer partner) lack a DBD.

Figure 6.3 Known members of the nuclear receptor superfamily

Adapted from (Mangelsdorf et al., 1995).

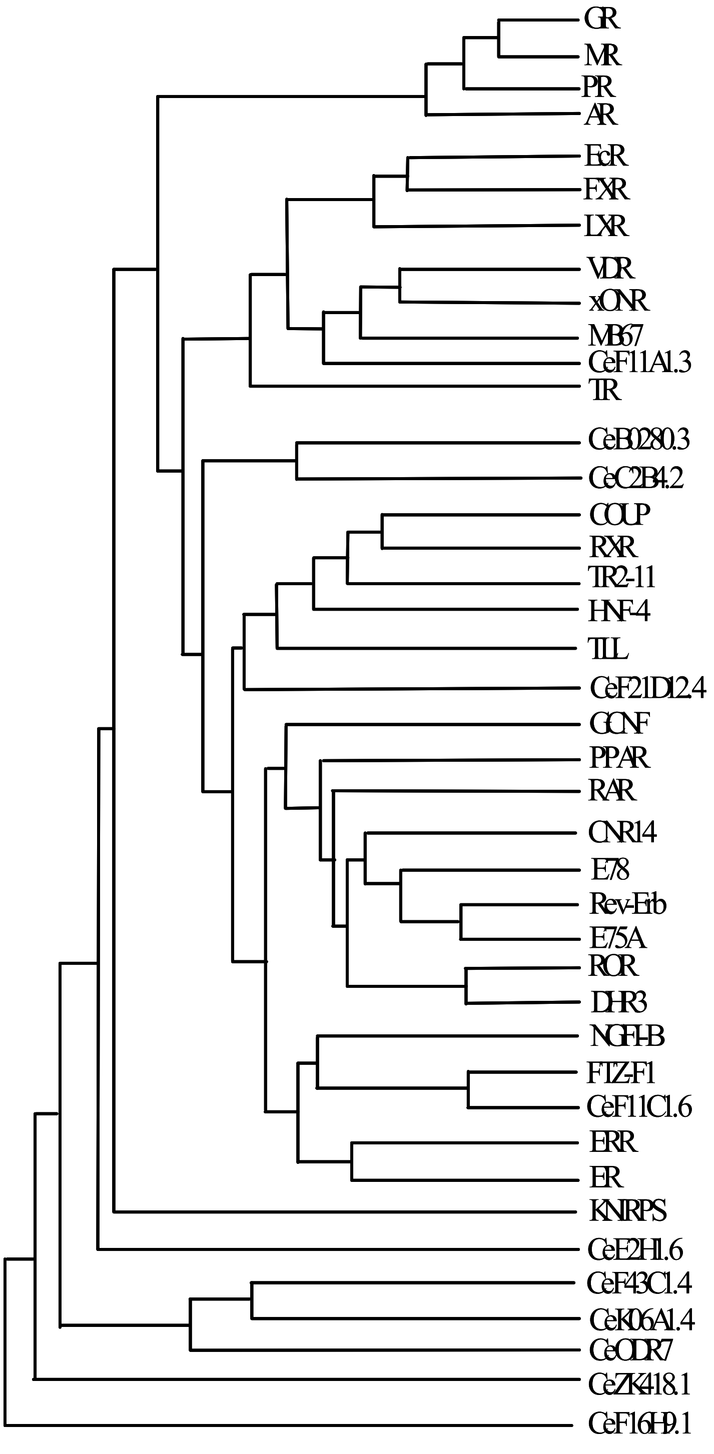


Table 6.1 Human nuclear receptors, nomenclature and family name

External Gene ID	HGNC symbol	Common Name	Project ID	Ensembl Gene ID	Entrez Gene ID	Known ligands?
NR0B1	NR0B1	DAX1	9	ENSG00000169297	190	orphan
NR0B2	NR0B2	SHP	10	ENSG00000131910	8431	orphan
NR1A1	THRA	THRA	46	ENSG00000126351	7067	yes
NR1A2	THRB	THRB	47	ENSG00000151090	7068	yes
NR1B1	RARA	RARA	37	ENSG00000131759	5914	yes
NR1B2	RARB	RARB	38	ENSG00000077092	5915	yes
NR1B3	RARG	RARG	39	ENSG00000172819	5916	yes
NR1C1	PPARA	PPARA	34	ENSG00000186951	5465	orphan
NR1C2	PPARD	PPARD	35	ENSG00000112033	5467	orphan
NR1C3	PPARG	PPARG	36	ENSG00000132170	5468	orphan
NR1D1	NR1D1	REVERBA	11	ENSG00000126368	9572	orphan
NR1D2	NR1D2	REVERBB	12	ENSG00000174738	9975	orphan
NR1F1	RORA	RORA	40	ENSG00000069667	6095	orphan
NR1F2	RORB	RORB	41	ENSG00000198963	6096	orphan
NR1F3	RORC	RORC	42	ENSG00000143365	6097	orphan
NR1H2	NR1H2	LXRB	13	ENSG00000131408	7376	orphan
NR1H3	NR1H3	LXRA	14	ENSG00000025434	10062	orphan
NR1H4	NR1H4	FXR	15	ENSG00000012504	9971	orphan
NR1I1	VDR	VDR	48	ENSG00000111424	7421	yes
NR1I2	NR1I2	PXR	16	ENSG00000144852	8856	orphan
NR1I3	NR1I3	CAR	17	ENSG00000143257	9970	orphan
NR2A1	HNF4A	HNF4A	7	ENSG00000101076	3172	orphan
NR2A2	HNF4G	HNF4G	8	ENSG00000164749	3174	orphan
NR2B1	RXRA	RXRA	43	ENSG00000186350	6256	orphan
NR2B2	RXRB	RXRB	44	ENSG00000206218	6257	orphan
NR2B3	RXRG	RXRG	45	ENSG00000143171	6258	orphan
NR2C1	NR2C1	TR2	18	ENSG00000120798	7181	orphan
NR2C2	NR2C2	TR4	19	ENSG00000177463	7182	orphan
NR2E1	NR2E1	TLL	20	ENSG00000112333	7101	orphan
NR2E3	NR2E3	Tailless	21	ENSG00000031544	10002	yes
NR2F1	NR2F1	SVP40	22	ENSG00000175745	7025	orphan
NR2F2	NR2F2	SVP44	23	ENSG00000185551	7026	orphan
NR2F6	NR2F6	EAR2	24	ENSG00000160113	2063	orphan
NR3A1	ESR1	ESR1	2	ENSG00000091831	2099	yes
NR3A2	ESR2	ESR2	3	ENSG00000140009	2100	yes
NR3B1	ESRRA	ESRRA	4	ENSG00000173153	2101	orphan
NR3B2	ESRRB	ESRRB	5	ENSG00000119715	2103	orphan
NR3B3	ESRRG	ESRRG	6	ENSG00000196482	2104	orphan
NR3C1	NR3C1	GR	25	ENSG00000113580	2908	yes
NR3C2	NR3C2	MR	26	ENSG00000151623	4306	yes
NR3C3	PGR	PGR (PR)	33	ENSG00000082175	5241	yes
NR3C4	AR	AR	1	ENSG00000169083	367	yes
NR4A1	NR4A1	HMR	27	ENSG00000123358	3164	orphan
NR4A2	NR4A2	NURR1	28	ENSG00000153234	4929	orphan
NR4A3	NR4A3	NOR-1	29	ENSG00000119508	8013	orphan
NR5A1	NR5A1	SF1	30	ENSG00000136931	2516	orphan
NR5A2	NR5A2	LRH1	31	ENSG00000116833	2494	orphan
NR6A1	NR6A1	RTR	32	ENSG00000148200	2649	orphan

6.2.5 Previous studies of positive selection

Following the completion of the rat genome sequence, Zhang et al. (2004) performed a phylogenetic analysis on the entire set of NRs in human, mouse and rat and found that by comparison of d_N/d_S ratios of the LBDs, NRs were subject to strong purifying selection. However, they found the pairwise d_N/d_S ratios of the *PXR* (pregnane X receptor) and *CAR* (constitutive androstane receptor) genes to be 4-6 times higher than the average. The biological functions of the *PXR* and *CAR* genes could form a basis to explain the positive selection result. The orphan genes, *CAR* and *PXR,* are found only in mammals and mediate transcription of a variety of detoxifying enzymes that are members of cytochrome P450 molecules in response to xenobiotic compounds. The cytochrome P450 molecules have been extensively studied and are known to be subject to diversifying evolution by gene duplication in mammals (Waterston et al., 2002; Emes et al., 2003; Gibbs et al., 2004). It is also known that *PXR* orthologues in mouse and human display differential sensitivity to various xenobiotic agents, providing a basis for species specificity of xenobiotic responses (Xie and Evans, 2001).

A more in-depth analysis was carried out by Krasowski et al. (2005), who analysed the entire set of NRs in vertebrates. Their analysis used the PAML site models (Yang and Swanson, 2002) in two likelihood-ratio tests (LRTs). The first LRT, which compared the model M0 with the more complex model M3, found that 41 of 48 nuclear receptors had an ω greater than 0.5 in the LBD and only 1 (*PPARG*) had an ω greater than 0.5 in the DBD. The second LRT, which compared model M7 against model M8, resulted in 10 of 132 receptors being identified as being significant for positive selection across all vertebrates, but only 3 of them were present in mammals. None of the M8 full-length analyses identified any residues to have ω exceeding 1. Only the LBD of the *PXR* gene had a sub-population of codons (5%) with an inferred value of ω to be greater than 1. They concluded that nuclear receptors were subject to strong purifying selection, especially within the DBD.

The nuclear receptors play a wide role in the aetiology of many human diseases (cancer, diabetes or hormone-resistant syndromes) and are important as therapeutic agents in the pharmaceutical industry. Thus, a comprehensive understanding of their evolution is required to aid in the development of new drug treatments (Chen, 2008). These receptors also control various metabolites and detoxifying enzymes, including the cytochrome P450s. The cytochrome P450 family is one of most duplicated gene families in mammalian genomes with at least 58 members in human and 102 in mouse (Nelson et al., 2004).

Earlier studies to investigate positive selection in nuclear receptors used site models that were later found to be inaccurate in inferring positive selection. The work described in this chapter will also use more sensitive methods than previous studies which have the potential to detect positive selection that affect only a few residues in genes in which most residues are under purifying selection. In addition, with the availability of more mammalian genomes, the use of new orthologous sequences can greatly improve the power of maximum likelihood techniques to detect positive selection.

6.3 METHODS

Coding DNA sequences and their corresponding protein sequences of the 48 nuclear receptors in human and their available orthologues in 28 mammalian genomes were downloaded from the Ensembl database using Ensembl Biomart (http://www.ensembl.org/biomart/martview/). The accession numbers of the human genes used are given in Table 6.1.

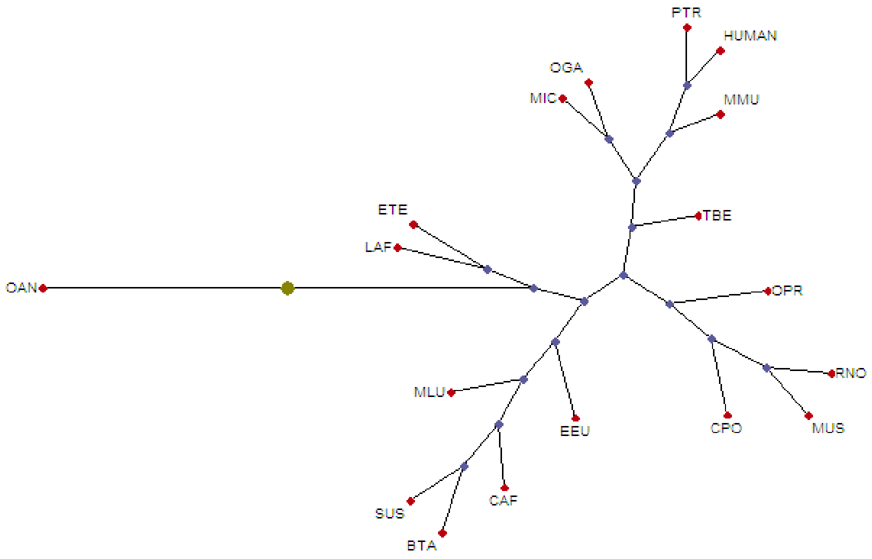
The protein sequences were aligned with DiALIGN (Morgenstern, 2007), Clustaw (Larkin et al., 2007) and Muscle (Edgar, 2004). The alignments generated by DiALIGN were chosen as manual curation based on the number of gaps and alignment length showed that it produced the most accurate local alignments (data not shown). Since nuclear receptors vary in the number of insertions and deletions between species, it is important that the homologous regions are aligned correctly. A cDNA alignment based on the corresponding protein alignment was generated by revtrans (Wernersson and Pedersen, 2003) and the files were converted to PAML (Yang, 1997) format.

Two trees were produced for each alignment, one a species tree based on the standard mammalian species tree (Murphy et al., 2001) (Figure 6.4a) and the other a gene tree (Figure 6.4b). The standard mammalian species tree was trimmed to only include the species in the alignment using prunetree (Ziheng Yang, unpublished). The gene trees were created from the nucleotide alignment using phyml (Guindon et al., 2005) with model TN93 (Tamura and Nei, 1993) and 4 rate categories in the discrete gamma models of rate variation among sites (Yang, 1994).

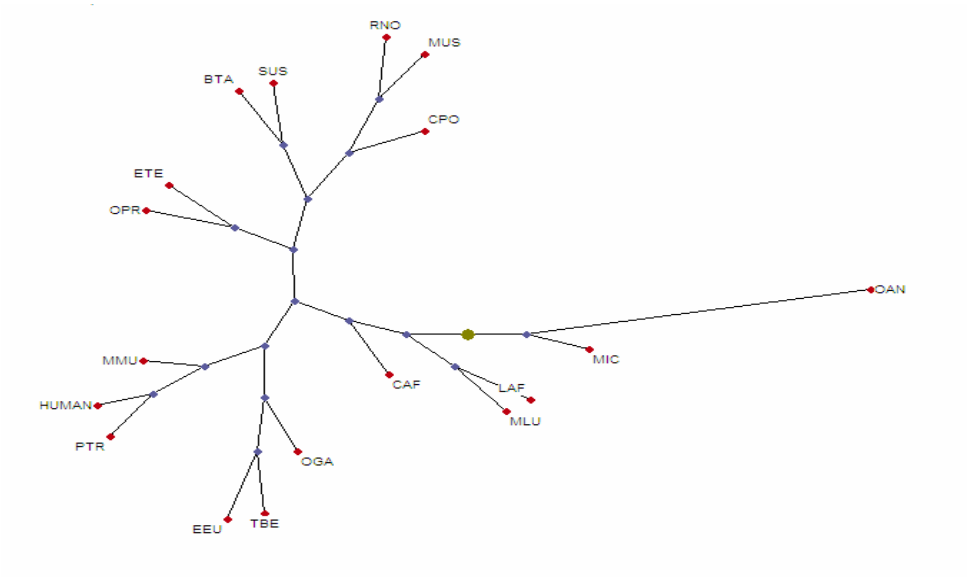
Figure 6.4 Example of a species tree (A) and a gene tree (B) based on the VDR gene

Species Abbreviations: LAF: Elephant, ETE: Tenrec, MIC: Mouse Lemur, OGA: Bushbaby, HUM: Human, PTR: Chimpanzee, MMU: Macaque, RNO: Rat, MUS: Mouse, CPO: Guinea Pig, STO: Squirrel, OPR: Pika, OCU: Rabbit, TBE: Tree Shrew, SAR: Common Shrew, EEU: Hedgehog, FCA: Cat, CAF: Dog, EQC: Horse, BTA: Cow, SUS: Pig, MLU: Bat, DNO: Armadillo, MOD: Opossum, OAN: Platypus

Tree A



Tree B



The branch-site model analysis was run on the 48 alignments as described in Section 2.4 except that only external branches were tested. The analysis was performed first with the species tree and then repeated with the gene tree. A site model analysis was also carried out with two likelihood ratio tests in the same manner as in Section 5.2.3.

To test if different domains were subject to differing selective pressures, the sequences were partitioned into functional domains. To determine the location of each domain in the alignment, a local version of the Pfam HMM library (Pfam_ls - downloaded March 2008) (Bateman et al., 2002) was used to search all sequences using the HMMER software. The resulting co-ordinates were then used to split the alignments into sub-alignments constituting of the A-B, C, D, E and F domains. The search against the Pfam HMM library found that the F domain was absent in 29 nuclear receptors. The DBD (C domain) was missing in the two genes of the NR0 family, as expected. The AB domain of NR2C2 and the D domain of the NR1H4 gene did not pass the length cut-off of 100 codons to be analysed so these were omitted from the analysis.

6.4 RESULTS

6.4.1 Comparisons of analyses using species trees and gene trees

Analyses were performed using both species trees and gene trees to observe the effect of tree topology on the inference of the positive selection. The site model analyses using gene trees resulted in 8 genes under positive selection ($p < 0.05$) in the M1a-M2a comparison and 27 genes in the M7-M8 comparison (Table 6.3). Eight genes were significant by both tests. When the species trees were used in the analysis, 12 genes were under positive selection ($p < 0.05$) with the M1a-M2 model, 28 genes were significant under the M7-M8 model and 8 genes were significant under both models (Tables 6.2 and 6.3). When the test resulted in a significant result, the sites inferred to be under positive selection by the Bayes empirical Bayes procedure were identical in the analyses between the two tree topologies. The branch-site model analysis using gene trees was less conservative. Eleven of the genes analysed using gene trees resulted in more lineages under positive selection compared to the same analysis performed using species trees. Eight of the genes analysed had more lineages under positive selection with species trees compared to the same analysis performed using gene trees.

In 41 of the 48 genes, the site model analyses gave the same results using species trees as those obtained with gene trees. The differences in the results of the remaining 7 genes were caused by differences in tree topology between the species tree and the gene tree. Such differences often arise after a gene duplication event from each gene copy having its own history (Hahn, 2007). Since use of the species trees resulted in more conservative results with the branch-site models and the mammalian species tree has been well studied, to allow comparisons between analyses, the results presented below are from analyses performed using the species tree.

Table 6.2 Number of genes in each species that were under positive selection (p < 0.05) using site models and the species tree

	Full sequence	AB	C (DBD)	D	E (LBD)	F
M2a vs M1a	12	4	2	2	0	3
M8 vs M7	28	12	3	6	5	3

Table 6.3 Nuclear receptors under positive selection using site models

Key: M2a – significant in the M1a-M2a comparison; M8 – significant in the M7-M8 comparison. The domain specific results are from analyses with species trees.

Gene Name	Full sequence (Species Tree)	AB domain	C (DBD) domain	D domain	E (LBD) domain	F domain	Full sequence (Gene tree)
AR	M2a/M8	M2a/M8					M2a/M8
ESR1		M8					
ESR2					M8	absent	
ESRRA	M8						
ESRRB	M2a/M8	M8					M2a/M8
ESRRG	M8			M2a/M8		absent	M8
HNF4A	M2a/M8						M2a/M8
HNF4G		M8					
NR0B1	M8	absent	absent	M8	M8	absent	M8
NR0B2		absent	absent			absent	
NR1D1	M8				M8	absent	M8
NR1D2						absent	
NR1H2	M8			M8	M8	absent	M8
NR1H3							
NR1H4	M8	M8		absent		absent	M8
NR1I2	M2a/M8	M2a/M8			M8		M2a/M8
NR1I3	M8		M8	M8		absent	M8
NR2C1	M8						M8
NR2C2		absent				absent	
NR2E1	M2a/M8						M2a/M8
NR2E3						absent	
NR2F1						absent	
NR2F2							
NR2F6	M8					absent	M8
NR3C1	M8	M8	M8			absent	M8
NR3C2	M8	M8					M8
NR4A1	M8	M2a/M8				absent	M8
NR4A2						absent	
NR4A3						absent	
NR5A1	M8						M8
NR5A2							
NR6A1							
PGR						absent	
PPARA	M8			M8		absent	M8
PPARD	M8					absent	M8
PPARG	M8					absent	M8
RARA	M2a/M8					M2a/M8	M2a/M8
RARB	M2a/M8					M2a/M8	M8
RARG	M2a/M8	M8				absent	M8
RORA	M2a/M8		M2a/M8			absent	M2a/M8
RORB		M2a/M8				absent	M8
RORC						absent	
RXRA						absent	
RXRB			M2a				
RXRG						absent	
THRA	M2a/M8					M2a/M8	M2a/M8
THRB	M2a/M8					absent	
VDR	M2a/M8	M8		M2a/M8			M8

6.4.2 Results from site analyses

The site model analysis conducted in this study inferred 12 of the 48 genes to be under positive selection in both the M1a-M2a and M7-M8 analyses ($p < 0.05$). An additional 16 genes were significant under the M7-M8 analysis alone. The number of genes expected to under positive selection by chance cannot be truly known in analyses involving maximum likelihood techniques, however at significance level 5%, there should be less than 5% significant tests if no genes were under positive selection.

The genes in the NR superfamily generally show nucleotide variation across species consistent with strong purifying selection, particularly in the DBDs. The LBD domain might be more partial to positive selection pressure particularly in gene products that detect endogenous and xenobiotic compounds, that likely differ between species. The investigation into the constituent domains of each nuclear receptor under the M1a-M2a models also found 1 or more domains in 25 genes to be under positive selection (Table 6.3) including the DBD domain, which is thought to be highly conserved. Most genes showed signals of positive selection in the AB domain and the D (hinge) domain with site class patterns as seen in Figure 6.5. The genes *PXR* (gene 16) and *CAR* (gene 17) were both found to be under positive selection and also surprisingly, their close relative *VDR* (Figure 6.5), which had not been found to be under positive selection previously.

The *PXR* gene also had many residues that were predicted to be under positive selection by the Bayes empirical Bayes method by both the M1a-M2a and M7-M8 models:

M2a: 6* 10* 17P 69* 101D 108L** 338V 380G

M8: 2E 6* 10** 17P 19M 66L 69* 70* 73* 75* 101D 106L 108L** 115S 253* 329* 334G 338V 380G 459* 510G (sites marked with one asterisk had a posterior probability of 0.95 and those reaching 0.99 are marked with a double asterisk). The protein structure of the ligand-binding domain of *PXR* (1M13.pdb) was used to map the 5 positively selected sites within this region (Figure 6.6).

Figure 6.6 Crystal structure of the *PXR* ligand-binding domain
Positively-selected sites (329, 334, 338, 380 and 459) are marked in red.

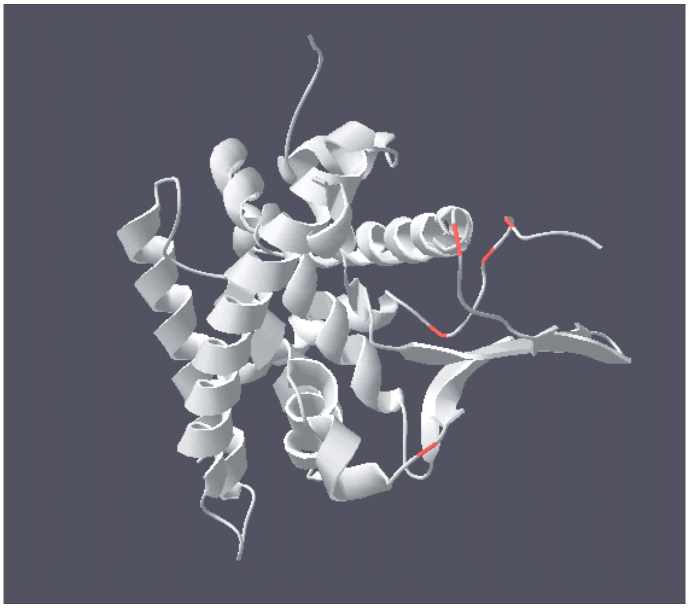
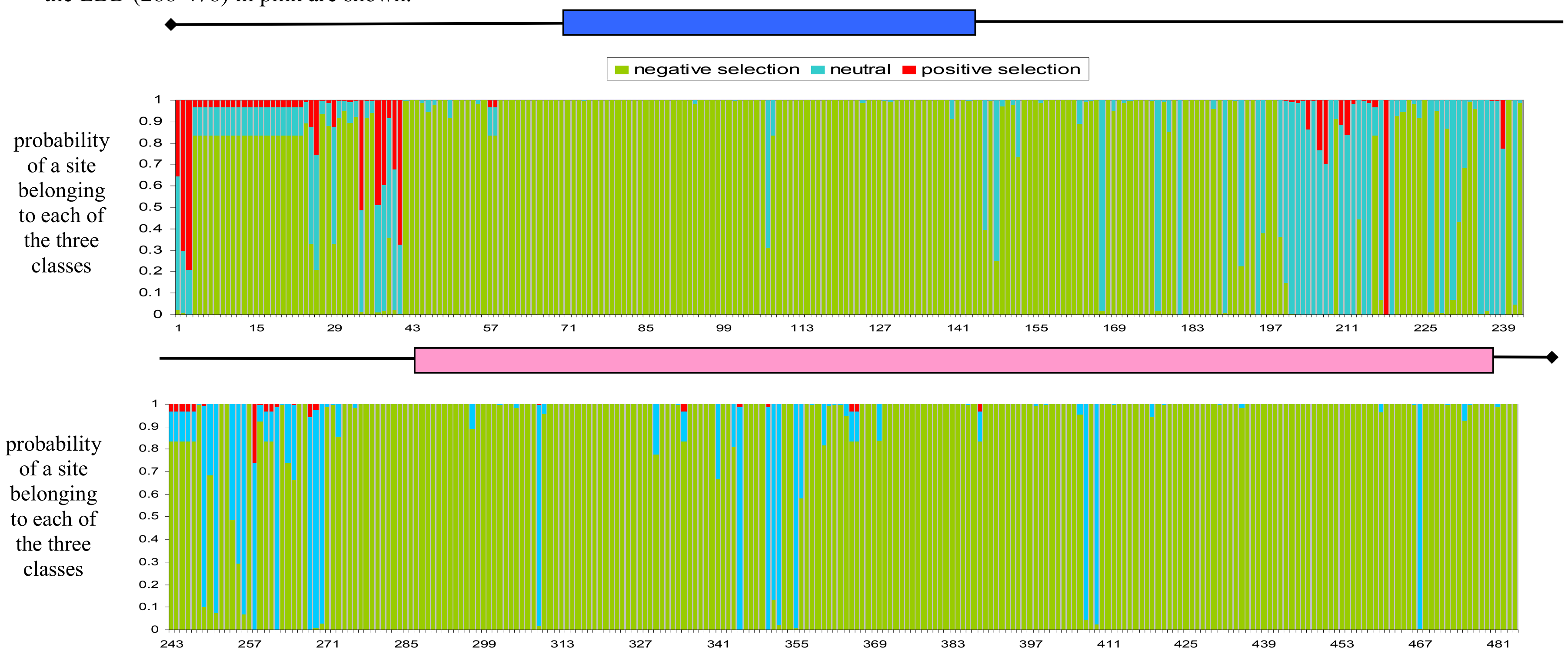


Figure 6.5 Posterior probabilities for sites in the *VDR* gene being under positive (red), purifying (green) or under relaxed functional constraints (blue) in the positive selection model M2a

The schematic diagram indicates the beginning and end of each domain within the gene. The residues encompassing the DBD (66-141) in blue and the LBD (288-478) in pink are shown.



6.4.3 Results under branch-site models

The branch-site analyses detected 3 genes under positive selection along the human lineage (Table 6.4). The numbers in the other species ranged from 2 in mouse, chimpanzee, pig and pika to 18 and 21 in the bushbaby and platypus, respectively, which are novel findings. The number of species that were under positive selection for each gene when the full length sequence was used is listed in Table 6.5. The results from the analysis of individual domains can be found in Appendix 3. The LBD was under positive selection in 43 of the 48 genes in one or more species. The number of genes under positive selection for the other domains varied from 32 for the AB domain, 22 for the DBD domain and 31 for the D domain. Once again, the platypus had several genes and domains that had faster rates of evolution compared with other species.

Table 6.4 Number of genes in each species that were under positive selection under the branch-site models ($p < 0.05$)

Ensembl code	Species Name	Full length	AB	C (DBD)	D	E (LBD)	F
Afrotheria							
LAF	Elephant	7	2	5	0	3	1
ETE	Tenrec	7	6	0	2	1	0
Primates							
MIC	Mouse Lemur	9	6	1	2	5	2
OGA	Bushbaby	18	6	1	5	10	1
HUM	Human	3	2	0	1	0	0
PTR	Chimpanzee	2	4	0	1	1	0
MMU	Macaque	8	3	4	1	5	0
Rodents							
RNO	Rat	3	0	0	0	4	2
MUS	Mouse	2	3	0	0	1	0
CPO	Guinea Pig	3	1	1	3	3	0
STO	Squirrel	11	5	4	4	6	2
Lagomorpha							
OPR	Pika	2	5	0	2	1	1
OCU	Rabbit	8	6	2	3	2	0
Scandentia							
TBE	Tree Shrew	7	4	2	2	3	1
Laurasiatheria							
SAR	Common Shrew	6	3	1	2	2	1
EEU	Hedgehog	6	4	4	0	2	1
FCA	Cat	5	1	0	3	4	0
CAF	Dog	13	8	3	5	6	0
EQC	Horse	6	3	0	0	1	1
BTA	Cow	7	4	1	0	5	1
SUS	Pig	2	1	0	0	0	0
MLU	Bat	7	5	2	4	2	0
Xenarthra							
DNO	Armadillo	7	2	2	1	4	1
Monotremes and Marsupials							
MOD	Opossum	12	5	0	4	6	1
OAN	Platypus	21	4	1	5	12	1

Table 6.5 Results of branch-site analyses by species full length sequences

Gene Name	No. of species analysed	No. of species under selection									
AR	22	4	MOD	STO	HUM	ETE					
ESR1	23	6	OAN	MLU	BTA	STO	RNO	OGA			
ESR2	26	2	EQC	LAF							
ESRRA	18	3	CAF	OCU	TBE						
ESRRB	21	5	MOD	STO	MMU	ETE	LAF				
ESRRG	23	3	OAN	OCU	OGA						
HNF4A	19	5	DNO	EEU	STO	CPO	PTR				
HNF4G	21	3	OAN	EEU	RNO						
NR0B1	19	1	OCU								
NR0B2	20	2	OAN	MIC							
NR1D1	22	3	MOD	SAR	MMU						
NR1D2	20	6	OAN	MOD	MLU	CAF	STO	CPO			
NR1H2	19	3	MLU	BTA	STO						
NR1H3	24	3	OAN	MOD	FCA						
NR1H4	24	1	SAR								
NR1I2	20	1	SUS								
NR1I3	22	4	MLU	CAF	MUS	OGA					
NR2C1	22	4	DNO	MLU	OGA	MIC					
NR2C2	22	6	OAN	FCA	OCU	MMU	OGA	LAF			
NR2E1	24	4	OAN	STO	OCU	MIC					
NR2E3	17	3	OAN	MOD	OGA						
NR2F1	19	4	BTA	EEU	SAR	TBE					
NR2F2	13	1	OGA								
NR2F6	15	1	CAF								
NR3C1	24	6	DNO	FCA	STO	MMU	OGA	MIC			
NR3C2	22	9	MOD	EQC	STO	RNO	OCU	TBE	MMU	MIC	ETE
NR4A1	21	1	OCU								
NR4A2	23	5	DNO	MLU	SAR	OCU	LAF				
NR4A3	21	6	OAN	MOD	BTA	CAF	EEU	MIC			
NR5A1	18	3	OAN	SUS	STO						
NR5A2	23	4	EQC	TBE	OGA	MIC					
NR6A1	21	5	MOD	DNO	EQC	CAF	HUM				
PGR	18	6	BTA	CAF	OPR	TBE	OGA	MIC			
PPARA	22	2	OAN	OGA							
PPARD	22	4	CAF	TBE	MMU	OGA					
PPARG	22	3	OAN	CAF	SAR						
RARA	22	8	OAN	DNO	MLU	EEU	SAR	STO	TBE	OGA	
RARB	23	8	OAN	EQC	FCA	MUS	MMU	MIC	ETE	LAF	
RARG	19	1	OAN								
RORA	21	3	OAN	DNO	OGA						
RORB	22	5	OAN	MOD	CAF	OPR	LAF				
RORC	9	3	OAN	MOD	HUM						
RXRA	9	3	CAF	MMU	OGA						
RXRB	21	6	MOD	EQC	CAF	OGA	ETE	LAF			
RXRG	20	2	FCA	OGA							
THRA	20	6	OAN	BTA	CAF	EEU	OGA	ETE			
THRB	23	2	BTA	ETE							
VDR	18	3	OAN	CPO	PTR						

The genes detected to be under positive selection in the human lineage were:

AR (androgen receptor)

Analysis of the full sequence of the androgen receptor resulted in a significant result for the test for positive selection along the human lineage. The AB domain of the *AR* gene was also under positive selection when each domain was analysed separately. In both the analyses of full length sequence and the AB domain, one residue, 233 (asparagine in human, serine in other mammals) was inferred to be under positive selection by the Bayes empirical Bayes method (Yang et al., 2005). This residue is within the AF-1 region and adjacent to conserved hydrophobic residues that are important for receptor-dependent gene transcription (Betney and McEwan, 2003). The AB domain of the *AR* gene was also under positive selection in opossum, rabbit, mouse lemur and the tenrec but the positively selected sites in these four species were different from that in human.

NR6A1 (retinoid-related, testis-associated receptor (RTR))

Within the *RTR* gene, the residue 241, which is glutamine, a polar molecule in humans and proline (non-polar) in other mammals, was reported as having had faster evolution rates. This residue is within the DBD of this protein. The structure of the *RTR* protein is not yet available.

NR1H2 (liver X receptor B)

The *NR1H2* gene did not give a significant result when the full length sequence was used but the D domain (hinge region) when tested independently, gave a significant result along the human lineage ($p < 0.001$). The sites under positive selection were consecutive from, residues 177 to 181 (ESQSQ).

6.5 DISCUSSION

Analysis of the full complement of nuclear receptors in the human genome and their mammalian orthologues clearly shows that nuclear receptors are under more positive selective pressure than previously thought. The most recent study by Krasowski et al. (2005) found only 3 NRs to be under positive selection under the M7-M8 models with none of the M8 full length analyses identifying residues to have ω exceeding 1. Only the LBD of the *PXR* gene had a sub-population of residues with an inferred value of ω to be greater than 1. Along with the *CAR* and *PXR* genes, the same analysis using more species in this study found 26 other genes to have exhibited adaptive evolution throughout the history of the gene. I also found the DBDs of 4 genes and the LBDs of 5 genes to contain codons under positive selection. The site models used in previous studies (Zhang et al., 2004; Krasowski et al., 2005) tend to average rates over time and hence lack power, compared to the branch-site models (Zhang et al., 2005). Moreover, the M3 model used in the Krasowski study has been shown to be not as precise in distinguishing positive selection as the M8 or M2a models. When there is a large fraction of neutral sites, the M3 model can yield a large number of incorrectly predicted sites and hence is no longer supported (Yang, 2007). This study using only mammalian sequences also narrows the length of evolutionary time under analysis. Hence short bursts of positive selection would be easier to detect under the branch-site model.

Many of the genomes used in this study were sequenced at low coverage (2-4x) and will inevitably exhibit increased levels of error in base calls, genome assemblies, orthologue identification (due to short contigs), and alignment, which can all lead to spurious signals for positive selection. The genes predicted to be under positive selection from this study in the low-coverage genomes provide an indication into the number and type of nuclear receptors that could potentially be under positive selection. These results will need to be confirmed when further releases of the low-coverage genome sequences are provided with improved data quality and sequence coverage.

I found that the androgen receptor gene had gone through positive selection in the AB domain with residue 233 having a high posterior probability of being under selection. This region also contains the activation domain, AF-1. The androgen receptor is unusual among nuclear receptors in that most, if not all, of its activity is mediated via the constitutive activation function in the N terminus (Bevan et al., 1999). This is in contrast to what occurs with the closely related estrogen receptor (*ER*), in which AF-2 is the major activation domain and AF-1 has little independent activity. Hence, the site of positive selection could possibly affect the function of the androgen receptor in humans. The most potent ligand of the androgen receptor is a metabolite of testosterone, 5α-dihydrotestosterone, among other molecules with agonist and antagonistic activity (Laudet and Gronemeyer, 2002). However, ligand binding in receptors varies among species and the ability to bind to new substrates can evolve from gene duplication (Emes et al., 2004). Frequently duplicated genes are often associated with adaptation and neo-functionalisation (Ohno, 1999; Blomme et al., 2006).

The *RTR* (retinoid-related, testis-associated receptor) gene was also positively selected along the human lineage with a positively selected site present in the DBD region. The *RTR* participates in the regulation of neurogenesis and reproductive functions (Greschik and Schule, 1998). No ligand or activator has yet been described for *RTR*. However, it is known that the *RTR* gene represses transcription via its DBD and the DBD has been shown to be essential for the function of *RTR* during early embryogenesis (Lan et al., 2002). Positive selection of the *RTR* gene along the human lineage may change the function of the DBD in humans.

The unusually large number of positively selected genes in the platypus lineage is perhaps an indication of its multifaceted reproductive and lactative systems, characteristics of both reptiles and mammals. The genome of this fascinating monotreme is still in the early stages of analysis (Warren et al., 2008). Much work still needs to be done before we can determine the contribution nuclear receptors make to the platypus' biology.

The patterns of evolution within the nuclear receptor family are complex, with many events of duplication, sometimes followed by pseudogenisation. This analysis demonstrates that domains within the nuclear receptor genes evolve independently of each other, which perhaps gives rise to new members of a family within some species. Furthermore, this study has shown that all domains, not just the LBD, are under positive selection in one or more species, which indicates that each and every one of the regions of these genes have an important function.

Conclusions

Genomic scans to detect the action of positive selection pressure can provide great insights into the underlying factors that contribute to biological differences between species. I identified genes that underwent positive selection during the evolution of humans and four mammals which are often used to model human diseases (mouse, rat, chimpanzee and dog) using maximum likelihood methods. This is the largest number of species investigated to date. Inclusion of more species increases the sensitivity of the methods and provides information about gene evolution in important animal models of human disease. Sources of error in genome scans such as sequencing errors, orthologue identification and alignment were rigorously addressed and the results subjected to an unprecedented level of quality control.

I show that genes that have been subject to positive selection pressure during human evolution are implicated in diseases such as epithelial cancers, schizophrenia, autoimmune diseases and Alzheimer's disease. This is one of the primary analyses trying to connect positive selection and phenotypic evidence from literature. These genes may be causative of the phenotypic disease differences between species and are promising targets for therapeutic intervention. This approach is of interest to drug development as detection of positive selection in a drug target or members of a disease pathway may cause animal models to be nonpredictive of human biology and explain some observed biomedical differences between species (Vamathevan et al., 2007). The dataset I present, of PSGs in five species, serves as an informative resource that can be consulted prior to selecting appropriate animal models during drug target validation.

The chimpanzee lineage was found to have many more genes under positive selection than any of the other lineages and three times more than the number of genes in the human lineage. I present evidence to argue against the possibility that this result is due to artefacts introduced by genome sequence coverage, gene sample selection or algorithmic sensitivity to errors in sequence data or alignments. Instead, we conclude that the elevated number of chimpanzee PSGs is a true reflection of evolutionary history and is most likely due to positive selection being more effective in the large population sizes chimpanzees have had in the past or possibly remarkable adaptation in the chimpanzee lineage. The extravagant adaptation seen in the chimpanzee lineage is interesting. Whether this pattern is specific to the chimpanzee can only be realised when more Old World monkeys are sequenced and analysed.

From these sets of genes, evidence was found to support the hypothesis that PSGs are significantly more likely to interact with other PSGs than genes evolving under neutral evolution or purifying selection, presumably because the functional divergence in one gene drives selection in its functional partners. This is the first such evidence to be detected widely among mammalian genes and is exemplified by evidence of co-evolution in integrin genes. It is suggested that the high level of connectivity between PSGs is caused by compensatory change of a protein's interaction partners when a protein undergoes change in response to selection.

One of the outcomes from a large-scale genome scan is the identification of potentially interesting genes or gene families that can then be analysed further. One gene family that arose from this genome scan was the nuclear receptors, which previously had been thought to be under strong purifying selection. The extended study of nuclear receptors, which have a well-studied, conserved structure that has been maintained throughout evolution from flies to humans, found that several domains were under positive selection. The co-evolution experiment demonstrated that positively selected genes do not have less interactors than genes under negative selection or neutral evolution. Even genes with key roles such as transcriptional regulation and those which interact with many other genes and play roles in vital functions can be under adaptive evolution. Regions of genes that were previously thought to be unimportant, if under strong selection pressure, must have key functions which we are yet to uncover. To further extend the co-evolution hypothesis, a wide study of nuclear receptors and their co-repressors and co-activators can be performed to investigate if these molecules, which are known to interact together, show signals for positive selection in the same species. The positive selection signals detected can indicate subtle variations in the functions of nuclear receptors among species.

This study encompasses several avenues of exploration into the fascinating area of molecular evolution. As Dobzhansky's famous quote "Nothing in biology makes sense except in the light of evolution" implies, an evolutionary point of view can shed light on almost all aspects of biology. It is becoming clear that the marriage of disciplines such as molecular evolution and genomics is going to bring about great advances and we are now only at the threshold.

In the years to come, as more genomes are sequenced using next-generation sequencing technology (Gupta, 2008), the potential to uncover the hidden truths of our past only becomes more tantalising. Current projects such as the Tree of Life project (http://www.tigr.org/tol/), the 1000 Genomes Project (http://www.1000genomes.org/) and the Encode project (Birney et al., 2007) are already making headway to improve the understanding of the complex genetic variation that exists. The Encode project, which I am currently a member of, has already changed the traditional definition of a gene and has uncovered complex patterns of dispersed regulation and abundant transcriptional landscape produced by the human genome. Understanding the patterns of variation in the genome that reveal where unusual selective forces have been acting is important in understanding the mechanisms underlying human disease.

References

Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. 1990. Basic local alignment search tool. J Mol Biol **215**:403-410.

Andersson, H., A. M. D'Antona, D. A. Kendall, G. Von Heijne, and C. N. Chin. 2003. Membrane assembly of the cannabinoid receptor 1: impact of a long N-terminal tail. Mol Pharmacol **64**:570-577.

Anisimova, M., J. P. Bielawski, and Z. Yang. 2001. Accuracy and Power of the Likelihood Ratio Test in Detecting Adaptive Molecular Evolution. Mol Biol Evol **18**:1585-1592.

Anisimova, M., and Z. Yang. 2007. Multiple Hypothesis Testing to Detect Lineages under Positive Selection that Affects Only a Few Sites. Mol Biol Evol **24**:1219-1228.

Arbiza, L., J. Dopazo, and H. Dopazo. 2006. Positive selection, relaxation, and acceleration in the evolution of the human and chimp genome. PLoS Comput Biol **2**:e38.

Bakewell, M. A., P. Shi, and J. Zhang. 2007. More genes underwent positive selection in chimpanzee evolution than in human evolution. PNAS **104**:7489-7494.

Balaresque, P. L., S. J. Ballereau, and M. A. Jobling. 2007. Challenges in human genetic diversity: demographic history and adaptation. Hum Mol Genet **16 Spec No. 2**:R134-139.

Barnes, M. R., S. Deharo, R. J. Grocock, J. R. Brown, and P. Sanseau. 2007. The micro RNA target paradigm: a fundamental and polymorphic control layer of cellular expression. Expert Opin Biol Ther **7**:1387-1399.

Bateman, A., E. Birney, L. Cerruti, R. Durbin, L. Etwiller, S. R. Eddy, S. Griffiths-Jones, K. L. Howe, M. Marshall, and E. L. Sonnhammer. 2002. The Pfam protein families database. Nucleic Acids Res **30**:276-280.

Bellino, F. L., and P. M. Wise. 2003. Nonhuman primate models of menopause workshop. Biol Reprod **68**:10-18.

Beniashvili, D. S. 1989. An overview of the world literature on spontaneous tumors in nonhuman primates. J Med Primatol **18**:423-437.

Bensen, J. T., L. H. Nelson, M. J. Pettenati, S. M. Block, S. W. Brusilow, L. R. Livingstone, and B. K. Burton. 1991. First report of management and outcome of pregnancies associated with hereditary orotic aciduria. Am J Med Genet **41**:426-431.

Bernatchez, L., and C. Landry. 2003. MHC studies in nonmodel vertebrates: what have we learned about natural selection in 15 years? J Evol Biol **16**:363-377.

Bertrand, S., F. G. Brunet, H. Escriva, G. Parmentier, V. Laudet, and M. Robinson-Rechavi. 2004. Evolutionary genomics of nuclear receptors: from twenty-five ancestral genes to derived endocrine systems. Mol Biol Evol **21**:1923-1937.

Betney, R., and I. J. McEwan. 2003. Role of conserved hydrophobic amino acids in androgen receptor AF-1 function. J Mol Endocrinol **31**:427-439.

Betts, M. J., and R. B. Russell. 2003. Amino acid properties and consequences of subsitutions. *in* M. R. Barnes, and I. C. Gray, eds. Bioinformatics for Geneticists. Wiley.

Bevan, C. L., S. Hoare, F. Claessens, D. M. Heery, and M. G. Parker. 1999. The AF1 and AF2 domains of the androgen receptor interact with distinct regions of SRC1. Mol Cell Biol **19**:8383-8392.

Birney, E., M. Clamp, and R. Durbin. 2004. GeneWise and Genomewise. Genome Res **14**:988-995.

Birney, E.J. A. Stamatoyannopoulos, A. Dutta, R. Guigo, T. R. Gingeras, E. H. Margulies, Z. Weng, M. Snyder, E. T. Dermitzakis, R. E. Thurman et al. 2007. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature **447**:799-816.

Blomme, T., K. Vandepoele, S. De Bodt, C. Simillion, S. Maere, and Y. Van de Peer. 2006. The gain and loss of genes during 600 million years of vertebrate evolution. Genome Biol **7**:R43.

Bloom, J. D., and C. Adami. 2003. Apparent dependence of protein evolutionary rate on number of interactions is linked to biases in protein-protein interactions data sets. BMC Evol Biol **3**:21.

Brune, M. 2004. Schizophrenia-an evolutionary enigma? Neurosci Biobehav Rev **28**:41-53.

Bustamante, C. D., A. Fledel-Alon, S. Williamson, R. Nielsen, M. T. Hubisz, S. Glanowski, D. M. Tanenbaum, T. J. White, J. J. Sninsky, R. D. Hernandez et al. 2005. Natural selection on protein-coding genes in the human genome. Nature **437**:1153-1157.

Castresana, J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol Biol Evol **17**:540-552.

Chen, T. 2008. Nuclear receptor drug discovery. Curr Opin Chem Biol **12**:418-426.

Chendrimada, T. P., K. J. Finn, X. Ji, D. Baillat, R. I. Gregory, S. A. Liebhaber, A. E. Pasquinelli, and R. Shiekhattar. 2007. MicroRNA silencing through RISC recruitment of eIF6. Nature **447**:823-828.

Chimpanzee Sequencing and Analysis Consortium (CSAC). 2005. Initial sequence of the chimpanzee genome and comparison with the human genome. Nature **437**:69-87.

Clark, A. G., S. Glanowski, R. Nielsen, P. D. Thomas, A. Kejariwal, M. A. Todd, D. M. Tanenbaum, D. Civello, F. Lu, B. Murphy et al. 2003. Inferring nonneutral evolution from human-chimp-mouse orthologous gene trios. Science **302**:1960-1963.

Coggins, C. R. 2007. An updated review of inhalation studies with cigarette smoke in laboratory animals. Int J Toxicol **26**:331-338.

Cole, G., and J. W. Neal. 1990. The brain in aged elephants. J Neuropathol Exp Neurol **49**:190-192.

Coutinho, A. M., I. Sousa, M. Martins, C. Correia, T. Morgadinho, C. Bento, C. Marques, A. Ataide, T. S. Miguel, J. H. Moore et al. 2007. Evidence for epistasis between SLC6A4 and ITGB3 in autism etiology and in the determination of platelet serotonin levels. Hum Genet **121**:243-256.

Crespi, B., K. Summers, and S. Dorus. 2007. Adaptive evolution of genes underlying schizophrenia. Proc Biol Sci **274**:2801-2810.

Crespi, B. J., and K. Summers. 2006. Positive selection in the evolution of cancer. Biol Rev Camb Philos Soc **81**:407-424.

Darwin, C (1859) The Origin of Species. Penguin Books, London.

Deeb, S. S., A. L. Jorgensen, L. Battisti, L. Iwasaki, and A. G. Motulsky. 1994. Sequence divergence of the red and green visual pigments in great apes and humans. Proc Natl Acad Sci U S A **91**:7262-7266.

Dorus, S., E. J. Vallender, P. D. Evans, J. R. Anderson, S. L. Gilbert, M. Mahowald, G. J. Wyckoff, C. M. Malcom, and B. T. Lahn. 2004. Accelerated evolution of nervous system genes in the origin of Homo sapiens. Cell **119**:1027-1040.

Edgar, R. C. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res **32**:1792-1797.

Emes, R. D., L. Goodstadt, E. E. Winter, and C. P. Ponting. 2003. Comparison of the genomes of human and mouse lays the foundation of genome zoology. Hum Mol Genet **12**:701-709.

Emes, R. D., M. C. Riley, C. M. Laukaitis, L. Goodstadt, R. C. Karn, and C. P. Ponting. 2004. Comparative evolutionary genomics of androgen-binding protein genes. Genome Res **14**:1516-1529.

Endo, T., K. Ikeo, and T. Gojobori. 1996. Large-scale search for genes on which positive selection may operate. Mol Biol Evol **13**:685-690.

Eppig, J. T., C. J. Bult, J. A. Kadin, J. E. Richardson, J. A. Blake, A. Anagnostopoulos, R. M. Baldarelli, M. Baya, J. S. Beal, S. M. Bello et al. 2005. The Mouse Genome Database (MGD): from genes to mice--a community resource for mouse biology. Nucleic Acids Res **33**:D471-475.

Escriva Garcia, H., V. Laudet, and M. Robinson-Rechavi. 2003. Nuclear receptors are markers of animal genome evolution. J Struct Funct Genomics **3**:177-184.

Escriva, H., F. Delaunay, and V. Laudet. 2000. Ligand binding and nuclear receptor evolution. Bioessays **22**:717-727.

Escriva, H., R. Safi, C. Hanni, M. C. Langlois, P. Saumitou-Laprade, D. Stehelin, A. Capron, R. Pierce, and V. Laudet. 1997. Ligand binding was acquired during evolution of nuclear receptors. Proc Natl Acad Sci U S A **94**:6803-6808.

Ewan, R., J. Huxley-Jones, A. P. Mould, M. J. Humphries, D. L. Robertson, and R. P. Boot-Handford. 2005. The integrins of the urochordate Ciona intestinalis provide novel insights into the molecular evolution of the vertebrate integrin family. BMC Evol Biol **5**:31.

Federici, T., and N. M. Boulis. 2007. Ribonucleic acid interference for neurological disorders: candidate diseases, potential targets, and current approaches. Neurosurgery **60**:3-15; discussion 15-16.

Felsenstein, J. 1981. Evolutionary trees from DNA sequences: A maximum likelihood approach. J Mol Evol **17**:368 - 376.

Fitch, W. M. 2000. Homology a personal view on some of the problems. Trends Genet **16**:227-231.

Fitch, W. M. 1971. Towards defining the course of evolution: Minimum change for a specific tree topology. Systematic Zoology **20**:406-416.

Flier, J. S., K. S. Cook, P. Usher, and B. M. Spiegelman. 1987. Severely impaired adipsin expression in genetic and acquired obesity. Science **237**:405-408.

Fraser, H. B., and A. E. Hirsh. 2004. Evolutionary rate depends on number of protein-protein interactions independently of gene expression level. BMC Evol Biol **4**:13.

Fraser, H. B., A. E. Hirsh, L. M. Steinmetz, C. Scharfe, and M. W. Feldman. 2002. Evolutionary rate in the protein interaction network. Science **296**:750-752.

Fraser, H. B., D. P. Wall, and A. E. Hirsh. 2003. A simple dependence between protein evolution rate and the number of protein-protein interactions. BMC Evol Biol **3**:11.

Futreal, P. A., L. Coin, M. Marshall, T. Down, T. Hubbard, R. Wooster, N. Rahman, and M. R. Stratton. 2004. A census of human cancer genes. Nat Rev Cancer **4**:177-183.

Gearing, M., G. W. Rebeck, B. T. Hyman, J. Tigges, and S. S. Mirra. 1994. Neuropathology and apolipoprotein E profile of aged chimpanzees: implications for Alzheimer disease. Proc Natl Acad Sci U S A **91**:9382-9386.

Gearing, M., J. Tigges, H. Mori, and S. S. Mirra. 1996. A beta40 is a major form of beta-amyloid in nonhuman primates. Neurobiol Aging **17**:903-908.

Gene Ontology Consortium. 2008. The Gene Ontology project in 2008. Nucleic Acids Res **36**:D440-444.

Ghanem, M. E., T. Nakao, and M. Nishibori. 2006. Deficiency of uridine monophosphate synthase (DUMPS) and X-chromosome deletion in fetal mummification in cattle. Anim Reprod Sci **91**:45-54.

Gibbs, R., Rogers. A.J., Katze, M. G., Bumgarner, R., Weinstock G. M., Mardis E. R., Remington, K. A., Strausberg, R. L., Venter J. C., Wilson, R. K., et al. 2007. Evolutionary and biomedical insights from the rhesus macaque genome. Science **316**:222-234.

Gibbs, R. A.G. M. WeinstockM. L. MetzkerD. M. MuznyE. J. SodergrenS. SchererG. ScottD. SteffenK. C. WorleyP. E. Burch et al. 2004. Genome sequence of the Brown Norway rat yields insights into mammalian evolution. Nature **428**:493-521.

Giguere, V. 1999. Orphan nuclear receptors: from gene to function. Endocr Rev **20**:689-725.

Gilad, Y., O. Man, and G. Glusman. 2005. A comparison of the human and chimpanzee olfactory receptor gene repertoires. Genome Res **15**:224-230.

Gilad, Y., A. Oshlack, G. K. Smyth, T. P. Speed, and K. P. White. 2006. Expression profiling in primates reveals a rapid evolution of human transcription factors. Nature **440**:242-245.

Glazko, G., V. Veeramachaneni, M. Nei, and W. Makalowski. 2005. Eighty percent of proteins are different between humans and chimpanzees. Gene **346**:215-219.

Goldman, N., and Z. Yang. 1994. A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol Biol Evol **11**:725-736.

Greschik, H., and R. Schule. 1998. Germ cell nuclear factor: an orphan receptor with unexpected properties. J Mol Med **76**:800-810.

Grove, J., T. Huby, Z. Stamataki, T. Vanwolleghem, P. Meuleman, M. Farquhar, A. Schwarz, M. Moreau, J. S. Owen, G. Leroux-Roels et al. 2007. Scavenger receptor BI and BII expression levels modulate hepatitis C virus infectivity. J Virol **81**:3162-3169.

Guex, N., and M. C. Peitsch. 1997. SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. Electrophoresis **18**:2714-2723.

Guindon, S., F. Lethiec, P. Duroux, and O. Gascuel. 2005. PHYML Online--a web server for fast maximum likelihood-based phylogenetic inference. Nucleic Acids Res **33**:W557-559.

Gupta, P. K. 2008. Single-molecule DNA sequencing technologies for future genomics research. Trends Biotechnol.

Hahn, M. W. 2007. Bias in phylogenetic tree reconciliation methods: implications for vertebrate genome evolution. Genome Biol **8**:R141.

Hanlon, C. S., and D. C. Rubinsztein. 1995. Arginine residues at codons 112 and 158 in the apolipoprotein E gene correspond to the ancestral state in humans. Atherosclerosis **112**:85-90.

Hao, L., and M. Nei. 2005. Rapid expansion of killer cell immunoglobulin-like receptor genes in primates and their coevolution with MHC Class I genes. Gene **347**:149-159.

Harden, K. K., and J. L. Robinson. 1987. Deficiency of UMP synthase in dairy cattle: a model for hereditary orotic aciduria. J Inherit Metab Dis **10**:201-209.

Hardison, R. C. 2003. Comparative Genomics. PLoS Biology **1**:e58.

Harkany, T., M. B. Dobszay, F. Cayetanot, W. Hartig, T. Siegemund, F. Aujard, and K. Mackie. 2005. Redistribution of CB1 cannabinoid receptors during evolution of cholinergic basal forebrain territories and their cortical projection areas: A comparison between the gray mouse lemur (Microcebus murinus, primates) and rat. Neuroscience **135**:595.

Hasegawa, M., H. Kishino, and T. Yano. 1985. Dating the human-ape splitting by a molecular clock of mitochondrial DNA. J. Mol. Evol **22**:160-174.

Hawks, J., E. T. Wang, G. M. Cochran, H. C. Harpending, and R. K. Moyzis. 2007. Recent acceleration of human adaptive evolution. Proc Natl Acad Sci U S A **104**:20753-20758.

Heemskerk, J., A. J. Tobin, and B. Ravina. 2002. From chemical to drug: neurodegeneration drug screening and the ethics of clinical trials. Nat Neurosci **5 Suppl**:1027-1029.

Holbrook, J. D., and P. Sanseau. 2007. Drug discovery and computational evolutionary analysis. Drug Discov Today **12**:826-832.

Howlett, D. R., J. C. Richardson, A. Austin, A. A. Parsons, S. T. Bate, D. C. Davies, and M. I. Gonzalez. 2004. Cognitive correlates of Abeta deposition in male and female mice bearing amyloid precursor protein and presenilin-1 mutant transgenes. Brain Res **1017**:130-136.

Huang, H., E. E. Winter, H. Wang, K. G. Weinstock, H. Xing, L. Goodstadt, P. D. Stenson, D. N. Cooper, D. Smith, M. M. Alba et al. 2004. Evolutionary conservation and selection of human disease gene orthologs in the rat and mouse genomes. Genome Biol **5**:R47.

Hughes, A. L., and M. Nei. 1992. Models of host-parasite interaction and MHC polymorphism. Genetics **132**:863-864.

Hurles, M. 2004. Gene duplication: the genomic trade in spare parts. PLoS Biol **2**:E206.

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

International Chicken Genome Sequence Consortium. 2004. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. Nature **432**:695-716.

Izagirre, N., I. Garcia, C. Junquera, C. de la Rua, and S. Alonso. 2006. A scan for signatures of positive selection in candidate loci for skin pigmentation in humans. Mol Biol Evol **23**:1697-1706.

Jarrahian, A., V. J. Watts, and E. L. Barker. 2004. D2 dopamine receptors modulate Galpha-subunit coupling of the CB1 cannabinoid receptor. J Pharmacol Exp Ther **308**:880-886.

Jen, J. C., T. D. Graves, E. J. Hess, M. G. Hanna, R. C. Griggs, and R. W. Baloh. 2007. Primary episodic ataxias: diagnosis, pathogenesis and treatment. Brain **130**:2484-2493.

Jordan, I. K., Y. I. Wolf, and E. V. Koonin. 2003. No simple dependence between protein evolution rate and the number of protein-protein interactions: only the most prolific interactors tend to evolve slowly. BMC Evol Biol **3**:1.

Jouvenceau, A., L. H. Eunson, A. Spauschus, V. Ramesh, S. M. Zuberi, D. M. Kullmann, and M. G. Hanna. 2001. Human epilepsy associated with dysfunction of the brain P/Q-type calcium channel. Lancet **358**:801-807.

Jukes, T. H., and C. R. Cantor. 1969. Evolution of protein molecules. Pp. 21-132 *in* H. N. Munro, ed. Mammalian Protein Metabolism. Academic Press, New York.

Julian, M. D., A. B. Martin, B. Cuellar, F. Rodriguez De Fonseca, M. Navarro, R. Moratalla, and L. M. Garcia-Segura. 2003. Neuroanatomical relationship between type 1 cannabinoid receptors and dopaminergic systems in the rat basal ganglia. Neuroscience **119**:309-318.

Jungerius, B. J., M. L. Hoogendoorn, S. C. Bakker, R. Van't Slot, A. F. Bardoel, R. A. Ophoff, C. Wijmenga, R. S. Kahn, and R. J. Sinke. 2007. An association screen of myelin-related genes implicates the chromosome 22q11 PIK4CA gene in schizophrenia. Mol Psychiatry.

Kaessmann, H., V. Wiebe, G. Weiss, and S. Paabo. 2001. Great ape DNA sequences reveal a reduced diversity and an expansion in humans. Nat Genet **27**:155-156.

Katoh, M. 2002. GIPC gene family (Review). Int J Mol Med **9**:585-589.

Kehrer-Sawatzki, H., and D. N. Cooper. 2007. Understanding the recent evolution of the human genome: insights from human-chimpanzee genome comparisons. Hum Mutat **28**:99-130.

Keitges, E., M. Rivest, M. Siniscalco, and S. M. Gartler. 1985. X-linkage of steroid sulphatase in the mouse is evidence for a functional Y-linked allele. Nature **315**:226-227.

Keller, M. C., and G. Miller. 2006. Resolving the paradox of common, harmful, heritable mental disorders: which evolutionary genetic models work best? Behav Brain Sci **29**:385-404; discussion 405-352.

Kim, S. J., and D. J. Linden. 2007. Ubiquitous plasticity and memory storage. Neuron **56**:582-592.

Kimura, M. 1968. Evolutionary rate at the molecular level. Nature **217**:624-626.

King, J. L., and T. H. Jukes. 1969. Non-Darwinian evolution. Science **164**:788-798.

Klein, C., K. Schilling, R. J. Saunders-Pullman, J. Garrels, X. O. Breakefield, M. F. Brin, D. deLeon, D. Doheny, S. Fahn, J. S. Fink et al. 2000. A major locus for myoclonus-dystonia maps to chromosome 7q in eight families. Am J Hum Genet **67**:1314-1319.

Koike, A., and T. Takagi. 2005. PRIME: automatically extracted PRotein Interactions and Molecular Information databasE. In Silico Biol **5**:9-20.

Kola, I., and J. Landis. 2004. Can the pharmaceutical industry reduce attrition rates? Nat Rev Drug Discov **3**:711-715.

Kolchanov, N. A., E. V. Ignatieva, E. A. Ananko, O. A. Podkolodnaya, I. L. Stepanenko, T. I. Merkulova, M. A. Pozdnyakov, N. L. Podkolodny, A. N. Naumochkin, and A. G. Romashchenko. 2002. Transcription Regulatory Regions Database (TRRD): its status in 2002. Nucleic Acids Res **30**:312-317.

Krasowski, M. D., K. Yasuda, L. R. Hagey, and E. G. Schuetz. 2005. Evolutionary selection across the nuclear hormone receptor superfamily with a focus on the NR1I subfamily (vitamin D, pregnane X, and constitutive androstane receptors). Nucl Recept **3**:2.

Kremer, H., P. Zeeuwen, W. H. McLean, E. C. Mariman, E. B. Lane, C. M. van de Kerkhof, H. H. Ropers, and P. M. Steijlen. 1994. Ichthyosis bullosa of Siemens is caused by mutations in the keratin 2e gene. J Invest Dermatol **103**:286-289.

Laganiere, J., G. Deblois, C. Lefebvre, A. R. Bataille, F. Robert, and V. Giguere. 2005. From the Cover: Location analysis of estrogen receptor alpha target promoters reveals that FOXA1 defines a domain of the estrogen response. Proc Natl Acad Sci U S A **102**:11651-11656.

Lahav, R. 2005. Endothelin receptor B is required for the expansion of melanocyte precursors and malignant melanoma. Int J Dev Biol **49**:173-180.

Lalueza-Fox, C., H. Rompler, D. Caramelli, C. Staubert, G. Catalano, D. Hughes, N. Rohland, E. Pilli, L. Longo, S. Condemi et al. 2007. A melanocortin 1 receptor allele suggests varying pigmentation among Neanderthals. Science **318**:1453-1455.

Lan, Z. J., A. C. Chung, X. Xu, F. J. DeMayo, and A. J. Cooney. 2002. The embryonic function of germ cell nuclear factor is dependent on the DNA binding domain. J Biol Chem **277**:50660-50667.

Lao, O., J. M. de Gruijter, K. van Duijn, A. Navarro, and M. Kayser. 2007. Signatures of positive selection in genes associated with human skin pigmentation as revealed from analyses of single nucleotide polymorphisms. Ann Hum Genet **71**:354-369.

Larkin, M. A., G. Blackshields, N. P. Brown, R. Chenna, P. A. McGettigan, H. McWilliam, F. Valentin, I. M. Wallace, A. Wilm, R. Lopez et al. 2007. Clustal W and Clustal X version 2.0. Bioinformatics **23**:2947-2948.

Laudet, V., and H. Gronemeyer. 2002. The nuclear receptors factsbook. Academic Press, London.

Li, W. H. 1993. Unbiased estimation of the rates of synonymous and nonsynonymous substitution. J Mol Evol **36**:96-99.

Li, W. H., C. I. Wu, and C. C. Luo. 1985. A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. Mol Biol Evol **2**:150-174.

Li, Y., M. Wallis, and Y. P. Zhang. 2005. Episodic evolution of prolactin receptor gene in mammals: coevolution with its ligand. J Mol Endocrinol **35**:411-419.

Lindgren, N., Z. Q. Xu, M. Herrera-Marschitz, J. Haycock, T. Hokfelt, and G. Fisone. 2001. Dopamine D(2) receptors regulate tyrosine hydroxylase activity and phosphorylation at Ser40 in rat striatum. Eur J Neurosci **13**:773-780.

Loder, E. 2002. What is the evolutionary advantage of migraine? Cephalalgia **22**:624-632.

Luo, X., Y. Ikeda, D. Lala, D. Rice, M. Wong, and K. L. Parker. 1999. Steroidogenic factor 1 (SF-1) is essential for endocrine development and function. J Steroid Biochem Mol Biol **69**:13-18.

Mahley, R. W. 1988. Apolipoprotein E: cholesterol transport protein with expanding role in cell biology. Science **240**:622-630.

Mahley, R. W., and Y. Huang. 1999. Apolipoprotein E: from atherosclerosis to Alzheimer's disease and beyond. Curr Opin Lipidol **10**:207-217.

Makalowski, W., and M. S. Boguski. 1998. Evolutionary parameters of the transcribed mammalian genome: an analysis of 2,820 orthologous rodent and human sequences. Proc Natl Acad Sci U S A **95**:9407-9412.

Makuwa, M., S. Souquiere, P. Telfer, O. Bourry, P. Rouquet, M. Kazanji, P. Roques, and F. Simon. 2006. Hepatitis viruses in non-human primates. J Med Primatol **35**:384-387.

Mangelsdorf, D. J., C. Thummel, M. Beato, P. Herrlich, G. Schutz, K. Umesono, B. Blumberg, P. Kastner, M. Mark, P. Chambon et al. 1995. The nuclear receptor superfamily: the second decade. Cell **83**:835-839.

Matys, V., E. Fricke, R. Geffers, E. Gossling, M. Haubrock, R. Hehl, K. Hornischer, D. Karas, A. E. Kel, O. V. Kel-Margoulis et al. 2003. TRANSFAC: transcriptional regulation, from patterns to profiles. Nucleic Acids Res **31**:374-378.

McCallion, A. S., and A. Chakravarti. 2001. EDNRB/EDN3 and Hirschsprung disease type II. Pigment Cell Res **14**:161-169.

McClure, H. M. 1973. Tumors in nonhuman primates: observations during a six-year period in the Yerkes primate center colony. Am J Phys Anthropol **38**:425-429.

McDonald, J. H., and M. Kreitman. 1991. Adaptive protein evolution at the Adh locus in Drosophila. Nature **351**:652-654.

Messier, W., and C. B. Stewart. 1997. Episodic adaptive evolution of primate lysozymes. Nature **385**:151-154.

Miller, W., K. D. Makova, A. Nekrutenko, and R. C. Hardison. 2004. Comparative Genomics. Annual Review of Genomics and Human Genetics **5**:15-56.

Miranda-Vizuete, A., C. M. Sadek, A. Jimenez, W. J. Krause, P. Sutovsky, and R. Oko. 2004. The mammalian testis-specific thioredoxin system. Antioxid Redox Signal **6**:25-40.

Miyata, T., and T. Yasunaga. 1980. Molecular evolution of mRNA: a method for estimating evolutionary rates of synonymous and amino acid substitutions from homologous nucleotide sequences and its application. J Mol Evol **16**:23-36.

Moras, D., and H. Gronemeyer. 1998. The nuclear receptor ligand-binding domain: structure and function. Curr Opin Cell Biol **10**:384-391.

Morgenstern, B. 2007. Alignment of genomic sequences using DIALIGN. Methods Mol Biol **395**:195-204.

Mundy, N. I. and Kelly, J. 2003 Evolution of a pigmentation gene, the melanocortin-1 receptor, in primates. Am. J. Phys. Anthropol. **121**: 67–80.

Murphy, W. J., E. Eizirik, S. J. O'Brien, O. Madsen, M. Scally, C. J. Douady, E. Teeling, O. A. Ryder, M. J. Stanhope, W. W. de Jong et al. 2001. Resolution of the early placental mammal radiation using Bayesian phylogenetics. Science **294**:2348-2351.

Muse, S. V., and B. S. Gaut. 1994. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. Mol Biol Evol **11**:715-724.

Nei, M., and T. Gojobori. 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol Biol Evol **3**:418-426.

Nelson, D. R., D. C. Zeldin, S. M. Hoffman, L. J. Maltais, H. M. Wain, and D. W. Nebert. 2004. Comparison of cytochrome P450 (CYP) genes from the mouse and human genomes, including nomenclature recommendations for genes, pseudogenes and alternative-splice variants. Pharmacogenetics **14**:1-18.

Nesse, R. M., and G. C. Williams. 1995. Why we get sick: the new science of Darwinian medicine. Times Books, New York.

Nielsen, R., C. Bustamante, A. G. Clark, S. Glanowski, T. B. Sackton, M. J. Hubisz, A. Fledel-Alon, D. M. Tanenbaum, D. Civello, T. J. White et al. 2005. A

Scan for Positively Selected Genes in the Genomes of Humans and Chimpanzees. PLoS Biology **3**:e170.

Nielsen, R., and Z. Yang. 1998. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. Genetics **148**:929-936.

Norgren, R. B., Jr. 2004. Creation of non-human primate neurogenetic disease models by gene targeting and nuclear transfer. Reprod Biol Endocrinol **2**:40.

Nuclear Receptors Nomenclature Committee. 1999. A unified nomenclature system for the nuclear receptor superfamily. Cell **97**:161-163.

Ohno, S. 1999. Gene duplication and the uniqueness of vertebrate genomes circa 1970-1999. Semin Cell Dev Biol **10**:517-522.

Olefsky, J. M. 2001. Nuclear receptor minireview series. J Biol Chem **276**:36863-36864.

Olson, M. V., and A. Varki. 2003. Sequencing the chimpanzee genome: insights into human evolution and disease. Nat Rev Genet **4**:20-28.

Opazo, J. C., Palma, R., E., Melo, F. and E. P. Lessa. 2005 Adaptive evolution of the insulin gene in caviomorph rodents. Mol Biol Evol. **22**: 1290-1298.

Pamilo, P., and N. O. Bianchi. 1993. Evolution of the Zfx and Zfy genes: rates and interdependence between the genes. Mol Biol Evol **10**:271-281.

Pearson, W. 2004. Finding protein and nucleotide similarities with FASTA. Curr Protoc Bioinformatics **Chapter 3**:Unit3 9.

Perler, F., A. Efstratiadis, P. Lomedico, W. Gilbert, R. Kolodner, and J. Dodgson. 1980. The evolution of genes: the chicken preproinsulin gene. Cell **20**:555-566.

Pravenec, M., and T. W. Kurtz. 2007. Molecular genetics of experimental hypertension and the metabolic syndrome: from gene pathways to new therapies. Hypertension **49**:941-952.

Priest, B. T., and G. J. Kaczorowski. 2007. Blocking sodium channels to treat neuropathic pain. Expert Opin Ther Targets **11**:291-306.

Puente, X. S., G. Velasco, A. Gutierrez-Fernandez, J. Bertranpetit, M. C. King, and C. Lopez-Otin. 2006. Comparative analysis of cancer genes in the human and chimpanzee genomes. BMC Genomics **7**:15.

Rajagopalan, D., and P. Agarwal. 2005. Inferring pathways from gene lists using a literature-derived network of biological relationships. Bioinformatics **21**:788-793.

Rice, P., I. Longden, and A. Bleasby. 2000. EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet **16**:276-277.

Robinson-Rechavi, M., H. Escriva Garcia, and V. Laudet. 2003. The nuclear receptor superfamily. J Cell Sci **116**:585-586.

Rosen, E. D., and B. M. Spiegelman. 2001. PPARgamma: a nuclear regulator of metabolism, differentiation, and cell growth. J Biol Chem **276**:37731-37734.

Russo, K., S. Hoch, C. Dima, J. Varga, and M. Teodorescu. 2000. Circulating anticentromere CENP-A and CENP-B antibodies in patients with diffuse and limited systemic sclerosis, systemic lupus erythematosus, and rheumatoid arthritis. J Rheumatol **27**:142-148.

Saadat, M., N. Pakyari, and H. Farrashbandi. 2008. Genetic polymorphism in the DNA repair gene XRCC1 and susceptibility to schizophrenia. Psychiatry Res **157**:241-245.

Sachidanandam, R., D. Weissman, S. C. Schmidt, J. M. Kakol, L. D. Stein, G. Marth, S. Sherry, J. C. Mullikin, B. J. Mortimore, D. L. Willey et al. 2001. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature **409**:928-933.

Salido, E. C., X. M. Li, P. H. Yen, N. Martin, T. K. Mohandas, and L. J. Shapiro. 1996. Cloning and expression of the mouse pseudoautosomal steroid sulphatase gene (Sts). Nat Genet **13**:83-86.

Schweizer, J., P. E. Bowden, P. A. Coulombe, L. Langbein, E. B. Lane, T. M. Magin, L. Maltais, M. B. Omary, D. A. Parry, M. A. Rogers et al. 2006. New consensus nomenclature for mammalian keratins. J Cell Biol **174**:169-174.

Schwenger, B., S. Schober, and D. Simon. 1993. DUMPS cattle carry a point mutation in the uridine monophosphate synthase gene. Genomics **16**:241-244.

Searls, D. B. 2003. Pharmacophylogenomics: Genes, Evolution and Drug Targets. Nature Reviews Drug Discovery **2**:613.

Seeler, J. S., C. Muchardt, A. Suessle, and R. B. Gaynor. 1994. Transcription factor PRDII-BF1 activates human immunodeficiency virus type 1 gene expression. J Virol **68**:1002-1009.

Seibold, H. R., and R. H. Wolf. 1973. Neoplasms and proliferative lesions in 1065 nonhuman primate necropsies. Lab Anim Sci **23**:533-539.

Smith, N. G., and A. Eyre-Walker. 2003. Human disease genes: patterns and predictions. Gene **318**:169-175.

Stark, J. M., J. Locke, and R. V. Heatley. 1980. Immunogenicity of lipid-conjugated antigens. I. The influence of chain length and degree of conjugation on induction of antibody in mice. Immunology **39**:345-352.

Stopkova, P., T. Saito, and H. M. Lachman. 2003. Abstracts for the XIth World Congress of Psychiatric Genetics: Molecular Analysis of PIK3C3 and PIP5K2A, Candidate Genes for Bipolar Affective Disorder and Schizophrenia. American Journal of Medical Genetics Part B: Neuropsychiatric Genetics **122**:1-178.

Strittmatter, W. J., A. M. Saunders, D. Schmechel, M. Pericak-Vance, J. Enghild, G. S. Salvesen, and A. D. Roses. 1993. Apolipoprotein E: high-avidity binding to beta-amyloid and increased frequency of type 4 allele in late-onset familial Alzheimer disease. Proc Natl Acad Sci U S A **90**:1977-1981.

Sugimoto, K., H. Migita, Y. Hagishita, H. Yata, and M. Himeno. 1992. An antigenic determinant on human centromere protein B (CENP-B) available for production of human-specific anticentromere antibodies in mouse. Cell Struct Funct **17**:129-138.

Swanson, W. J., and V. D. Vacquier. 2002. The rapid evolution of reproductive proteins. Nat Rev Genet **3**:137-144.

Tamura, K., and M. Nei. 1993. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. Mol Biol Evol **10**:512-526.

Tang, K., K. R. Thornton, and M. Stoneking. 2007. A New Approach for Using Genome Scans to Detect Recent Positive Selection in the Human Genome. PLoS Biol **5**:e171.

Tascou, S., K. Nayernia, J. Uedelhoven, D. Bohm, R. Jalal, M. Ahmed, W. Engel, and P. Burfeind. 2001. Isolation and characterization of differentially expressed genes in invasive and non-invasive immortalized murine male germ cells in vitro. Int J Oncol **18**:567-574.

Thomas, P. D., M. J. Campbell, A. Kejariwal, H. Mi, B. Karlak, R. Daverman, K. Diemer, A. Muruganujan, and A. Narechania. 2003. PANTHER: a library of protein families and subfamilies indexed by function. Genome Res **13**:2129-2141.

Thomas, P. D., A. Kejariwal, M. J. Campbell, H. Mi, K. Diemer, N. Guo, I. Ladunga, B. Ulitsky-Lazareva, A. Muruganujan, S. Rabkin et al. 2003. PANTHER: a browsable database of gene products organized by biological function, using curated protein family and subfamily classification. Nucleic Acids Res **31**:334-341.

Thornton, J. W. 2001. Evolution of vertebrate steroid receptors from an ancestral estrogen receptor by ligand exploitation and serial genome expansions. Proc Natl Acad Sci U S A **98**:5671-5676.

Tiilikainen, A. 1980. Is there biological relevance in the statistically significant association between the HLA system and disease? Med Biol **58**:241-245.

Tishkoff, S. A., and B. C. Verrelli. 2003. Patterns of human genetic diversity: implications for human evolutionary history and disease. Annu Rev Genomics Hum Genet **4**:293-340.

Valentin, J. P., A. S. Bass, A. Atrakchi, K. Olejniczak, and F. Kannosuke. 2005. Challenges and lessons learned since implementation of the safety pharmacology guidance ICH S7A. J Pharmacol Toxicol Methods **52**:22-29.

Valverde, P., E. Healy, S. Sikkink, F. Haldane, A. J. Thody, A. Carothers, I. J. Jackson, and J. L. Rees. 1996. The Asp84Glu variant of the melanocortin 1 receptor (MC1R) is associated with melanoma. Hum Mol Genet **5**:1663-1666.

Vamathevan, J., J. D. Holbrook, and R. D. Emes. 2007. The Mouse Genome as a Rodent Model in Evolutionary Studies *in* L. John Wiley & Sons, ed. Encyclopedia of Life Sciences.

Varki, A. 2000. A chimpanzee genome project is a biomedical imperative. Genome Res **10**:1065-1070.

Varki, A., and T. K. Altheide. 2005. Comparing the human and chimpanzee genomes: searching for needles in a haystack. Genome Res **15**:1746-1758.

Voight, B. F., S. Kudaravalli, X. Wen, and J. K. Pritchard. 2006. A Map of Recent Positive Selection in the Human Genome. PLoS Biology **4**:e72.

Wang, E., Y. C. Ding, P. Flodman, J. R. Kidd, K. K. Kidd, D. L. Grady, O. A. Ryder, M. A. Spence, J. M. Swanson, and R. K. Moyzis. 2004. The genetic architecture of selection at the human dopamine receptor D4 (DRD4) gene locus. Am J Hum Genet **74**:931-944.

Wang, X., R. H. Baloh, J. Milbrandt, and K. C. Garcia. 2006. Structure of artemin complexed with its receptor GFRalpha3: convergent recognition of glial cell line-derived neurotrophic factors. Structure **14**:1083-1092.

Wang, X., W. E. Grus, and J. Zhang. 2006. Gene losses during human origins. PLoS Biol **4**:e52.

Warren, W. C.L. W. HillierJ. A. Marshall GravesE. BirneyC. P. PontingF. GrutznerK. BelovW. MillerL. ClarkeA. T. Chinwalla et al. 2008. Genome analysis of the platypus reveals unique signatures of evolution. Nature **455**:256.

Waterston, R. H.K. Lindblad-TohE. BirneyJ. RogersJ. F. AbrilP. AgarwalR. AgarwalaR. AinscoughM. AlexanderssonP. An et al. 2002. Initial sequencing and comparative analysis of the mouse genome. Nature **420**:520-562.

Wehrli, P., I. Viard, R. Bullani, J. Tschopp, and L. E. French. 2000. Death receptors in cutaneous biology and disease. J Invest Dermatol **115**:141-148.

Wernersson, R., and A. G. Pedersen. 2003. RevTrans: Multiple alignment of coding DNA from aligned amino acid sequences. Nucleic Acids Res **31**:3537-3539.

White, S. A., S. E. Fisher, D. H. Geschwind, C. Scharff, and T. E. Holy. 2006. Singing mice, songbirds, and more: models for FOXP2 function and dysfunction in human speech and language. J Neurosci **26**:10376-10379.

Yong-Jin Won, Y-J. and J. Hey. 2005. Divergence Population Genetics of Chimpanzees. Mol Biol Evol **22**: 297-307.

Wong, W. S., Z. Yang, N. Goldman, and R. Nielsen. 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.

Wyckoff, G. J., W. Wang, and C. I. Wu. 2000. Rapid evolution of male reproductive genes in the descent of man. Nature **403**:304-309.

Xie, W., and R. M. Evans. 2001. Orphan nuclear receptors: the exotics of xenobiotics. J Biol Chem **276**:37739-37742.

Yang, C., B. Yu, D. Zhou, and S. Chen. 2002. Regulation of aromatase promoter activity in human breast tissue by nuclear receptors. Oncogene **21**:2854-2863.

Yang S., Liu Y., Lin A.A., Cavalli-Sforza L.L., Zhao Z., Su B. 2005. Adaptive evolution of MRGX2, a human sensory neuron specific gene involved in nociception. Gene **352**: 30-35.

Yang, Z. 1994. Maximum Likelihood Phylogenetic Estimation from DNA Sequences with Variable Rates over Sites: Approximate Methods. J Mol Evol **39**:306-314.

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

Yang, Z. 1998. Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. Mol Biol Evol **15**:568-573.

Yang, Z. 2000. Maximum Likelihood Estimation on Large Phylogenies and Analysis of Adaptive Evolution in Human Influenza Virus A. J Mol Evol **51**:423-432.

Yang, Z., and J. P. Bielawski. 2000. Statistical methods for detecting molecular adaptation. Trends in Ecology & Evolution **15**:496.

Yang, Z., and R. Nielsen. 2000. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. Mol Biol Evol **17**:32-43.

Yang, Z., R. Nielsen, N. Goldman, and A.-M. K. Pedersen. 2000. Codon-substitution models for heterogeneous selection pressure at amino acid sites. Genetics **155**:431-449.

Yang, Z., and R. Nielsen. 2002. Codon-Substitution Models for Detecting Molecular Adaptation at Individual Sites Along Specific Lineages. Mol Biol Evol **19**:908-917.

Yang, Z., and W. J. Swanson. 2002. Codon-substitution models to detect adaptive evolution that account for heterogeneous selective pressures among site classes. Mol Biol Evol **19**:49-57.

Yang, Z. 2005. The power of phylogenetic comparison in revealing protein function. PNAS **102**:3179-3180.

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

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

Yen, P. H. 1998. A long-range restriction map of deletion interval 6 of the human Y chromosome: a region frequently deleted in azoospermic males. Genomics **54**:5-12.

Yoon, S. N., J. L. Ku, Y. K. Shin, K. H. Kim, J. S. Choi, E. J. Jang, H. C. Park, D. W. Kim, M. A. Kim, W. H. Kim et al. 2008. Hereditary nonpolyposis colorectal cancer in endometrial cancer patients. Int J Cancer **122**:1077-1081.

Young, J. H., Y. P. Chang, J. D. Kim, J. P. Chretien, M. J. Klag, M. A. Levine, C. B. Ruff, N. Y. Wang, and A. Chakravarti. 2005. Differential susceptibility to hypertension is due to selection during the out-of-Africa expansion. PLoS Genet **1**:e82.

Yu, X. J., H. K. Zheng, J. Wang, W. Wang, and B. Su. 2006. Detecting lineage-specific adaptive evolution of brain-expressed genes in human using rhesus macaque as outgroup. Genomics **88**:745-751.

Zhang, J. 2004. Frequent False Detection of Positive Selection by the Likelihood Method with Branch-Site Models. Mol Biol Evol **21**:1332-1339.

Zhang, J., R. Nielsen, and Z. Yang. 2005. Evaluation of an Improved Branch-Site Likelihood Method for Detecting Positive Selection at the Molecular Level. Mol Biol Evol **22**:1-8.

Zhang, Z., P. E. Burch, A. J. Cooney, R. B. Lanz, F. A. Pereira, J. Wu, R. A. Gibbs, G. Weinstock, and D. A. Wheeler. 2004. Genomic analysis of the nuclear receptor family: new insights into structure, regulation, and evolution from the rat genome. Genome Res **14**:580-590.

APPENDIX 1

Descriptions of selected Perl programs in the branch-site analyses pipeline

Program Name	Input files	Description
Dir.pl	<name of directory containing all alignment and tree files>	This program creates subdirectories containing the alignment and the tree file required for bandit.pl. Subdirectories are named as genename_spliceversion (eg. MEL102_1). Subdirectories are only created for alignments that have exactly five (depends on type of analysis) sequences in the alignment file and have a sequence length greater than 50 codons. File extensions assumed: paml-N-MEL* for alignment, tree-MEL* for tree file.
M0Log.pl	M0Log.pl <M0-R1> < M0-R2>	This program compares the lnL values in the mlc files from duplicate runs of the M0 analysis. It compares the values for the first and second runs for the following conditions: lnL_diff > 0.0001, $kappa_diff > 0.001, $tree_diff > 0.001. A message starting with CHECK is printed to the log file.
TreeBuilder.pl	<name of directory containing M0 analyses>	This program creates trees for the branch-site model analysis from the M0tree file. For each folder named MEL* in the directory provided, the program looks for the M0tree file, takes the second line from the file and matches up to the branch length. Number of trees equals the number of branch lengths. Branch tree files are created in a directory called 'BranchTrees' in the directory that the program is run from.
Dir-branches.pl	name of directory containing all alignment and branchtree files	This program creates subdirectories containing the alignment and the branchtree files required for bandit_jv1.pl. File extensions assumed: paml-N-MEL* for alignment, branchtree-MEL* for tree file.
BSAnalyser.pl	Directory names of three replicates of BS-fixed run and 3 replicates of BS-free run.	This program compares the lnL values in the mlc files from three runs of the branch-site model analysis. For each directory in each array, the mlc file in that directory is read looking for values in BSAnalyser.log (subroutine strip): Values are stored in a multi-dimenisonal array (row# corresponds to tree number minus 1 - must start at 0) All values except the two branch lengths are stored to 3 decimal places The numerical difference between the lowest value and the other two is compared. If the difference is >= 0.001, a message starting with 'CONVERGENCE PROB' is printed to the .log file. The likelihood-ratio test is conducted on the two best runs: The mlc data is printed for the best fixed run and the best free run. If the siginificance level is greater than or equal to 0.05, the BEB analysis positive sites are printed to the .log file. If $posprob > 0.9900, sites are stored with ** suffix; if $posprob > 0.9500, sites are stored with * suffix
SummariseBS.pl	BSAnalysis.log and any tree file	This program summarises the BSAnalysis.log file to a file called summary.out in table format giving a total of how many genes are present at each level of significance. Trees are labelled 1 through to 7 with the following branch names: Human, Chimp, Hominid, Mouse, Rat, Murid, Dog.

APPENDIX 2

Names of genes under positive selection in each lineage

Human	Chimpanzee	Hominid	Mouse	Rat	Murid	Dog
ABCF1	ABCF1	ABCC11	ADMR	ABTB2	ABCB10	ALB
ALPPL2	ACTN2	ADAD2	AQP9	ACTN2	ACRBP	ALS2CL
ANGEL1	ACVRL1	ADRB2	AVPR1B	AIM1	ADRB3	APBB1
ANKRD35	ADCY5	AMAC1	C11ORF34	APOF	ARMC3	B4GALT4
ARID2	ADCY6	APOE	C19ORF16	ARHGAP27	BLVRA	BCAS1
ATPBD3	ALG10	AZGP1	C1QA	ARHGEF17	C10ORF88	BCL2
C8ORF42	ALOX12	C11ORF34	C1R	ASPH	C10ORF93	BCL3
CA14	ALPPL2	C18ORF34	C20ORF102	ATP11C	C1ORF156	BMS1
CACNA1A	ANGEL1	C1QA	C20ORF186	ATRX	C5ORF32	C11ORF34
CACNA1S	ANKRD35	C9ORF75	CA6	C11ORF34	C6ORF170	C12ORF34
CEACAM20	AQP2	CCL19	CCDC83	C19ORF16	C6ORF194	C15ORF27
CENPB	ARHGEF17	CD86	CCDC95	C3	CACNA1A	C20ORF186
CNGA4	ARMC3	CDC42EP2	CD86	C8B	CCDC73	C6ORF182
COL11A1	ARMCX5	CDKN2B	CDC14B	CA6	CD86	C8A
CTAGE6	ARRB1	CLSTN2	CENPC1	CARD11	CDCA2	CACNA1S
EDNRB	ATP6AP1	COL11A1	DHDH	CAST	CLSTN2	CCDC66
EMB	BLK	COL4A4	DOCK3	CCDC108	CNR1	CD79A
FLJ40722	BMP4	COMP	DSPP	CCDC18	CX62	CDCA2
GFRA3	C10ORF93	CXYorf1	FAIM3	CCDC7	CXCL13	CDH17
GIPC2	C11ORF24	DRD2	FLJ40722	CDC14B	DAG1	CDH22
GPR111	C14ORF39	EMP1	FLT1	CDH22	EFCAB5	CFP
GPR83	C16ORF48	ENG	FZD6	CDKN1B	ELOVL4	CLCN1
GPRC6A	C17ORF28	ENSA	GIMAP8	CDKN2D	ENAM	COMP
HIVEP3	C1ORF129	F5	GPR83	CFD	ETV2	CRB2
IFRD2	C1ORF174	FLJ46266	H6PD	CHRNA7	F5	CREBL1
INPP5B	C21ORF13	FZD2	HECW1	CILP	FZD2	DAG1
KCNK5	C3	GDPD4	HLA-DQA2	COL11A1	GAS2L2	DBX1
KIAA0372	C8ORF42	GIPC2	HLA-DRB1	CYB561	GJC1	DPP6
LOC388969	CCDC27	GPR116	HOXC6	DAGLB	GP1BA	DSPP
LOC389072	CCDC88C	GPR97	IZUMO1	DHDH	GPR1	EFCAB4B
LOC619207	CCDC97	GSTO2	KIAA1949	DNM1	GPR111	ENSA
MC1R	CDH15	HSPA1B	KLF11	DPP6	GPR113	EPHA1
MGC50722	CHKA	HTR1D	KRT2	DSC2	HBD	EVI2A
MICALCL	CLTB	HTR2C	LOC253012	EIF2C3	HECW1	F5
MOV10	CNGA4	ITGAV	LOC388323	FLJ13305	HLA-C	FGF20
MYF5	COL11A2	LOC220686	LOC497190	FLJ40722	HOXA11	FLJ45187
NR5A1	COMP	LOC619207	MARCH3	FXYD1	HRH2	GALNS
OR4F17	CPNE9	MADCAM1	MGC71993	FZD2	HSPA1A	GAS2L2
PDE6A	CSTF1	MCAM	MMPL1	GPR141	HSPE1	GDPD4
PIK3C2G	CXORF38	MMPL1	MRVI1	HDAC4	IFIT2	GGTLA1
RBM16	DBX1	MRC2	MYH15	HLA-B	INSL3	GPRASP1
RDM1	DIP2C	MSH2	NDUFC1	ICAM1	ITGAV	GRID1
REPIN1	DOPEY1	MYCT1	NLRP9	IMPG1	KRT2	GRM3
RKHD1	DUSP2	NRAP	NOVA2	INPP4A	KRTAP3-3	HADHB
RUFY4	DYRK2	NUDT22	NXPH4	IQSEC3	LAMC2	HCLS1
SLC5A9	EEF1G	PHYHD1	PHYH	IQUB	LIPC	HDAC4
SRL	EFCAB4A	RUFY4	PZP	ITGB2	LYZ	HDC

ST8SIA3	EFCAB4B	SCML4	RAB11FIP2	KCNA4	MAGEB4	HLA-DMB
TMPRSS12	EHHADH	TFF1	RAPGEF2	KIFAP3	MAST3	HRG
TRIM67	ELF4	TFPT	RRAGA	KRT31	MCOLN2	HSPA6
UMPS	EMD	TH	SASP	LASS2	MDC1	IFT88
XRCC1	ENTPD5	TRAF6	SCD	LCTL	MRC2	IL18RAP
ZNF324B	EOMES	TXNDC3	SEPT1	LDHD	MRPL54	INPP5B
ZRSR2	ETAA1	WDR42B	SERINC5	MAN1A2	NLGN4Y	ITGA5
	FAM134A	ZNF384	SH2D6	MSL-1	NLRP5	ITPKA
	FLRT1	ZNF665	SLC1A5	NKX2-5	NLRP9	KIAA1727
	GALNT6		SSTR2	LOC619207	NR1I2	KRTAP2-4
	GDPD4		STS	OPN5	NUF2	LCP2
	GFPT2		SYCP3	PCDHB14	OR7C1	LRP5
	GIPC2		SYT4	PDE6C	OXSM	MCAM
	GIYD1		TARP	PELI3	PHACTR1	MDGA1
	GPC3		TIMD4	PIK3R5	PHYH	MGC50722
	GPD1L		TMF1	PIM1	PNLIP	MMP12
	GPR19		TST	PLIN	PSMB6	MTDH
	GPX2		UNQ9438	PLXNC1	PTGIR	MUSK
	GRIK5			PRSS1	RAPGEF1	NLRP5
	GSTP1			PRSS35	RBM16	NOS1AP
	HCRTR1			PRSS36	RP5-1054A22.3	NRTN
	HLA-DRB1			PSMB4	SAFB	PALM2-AKAP2
	ICA1L			RAB11FIP3	SCNN1G	PAX1
	IGFALS			RGSL1	SLC34A3	PCDHB6
	INPP5B			RP9	SNRPA	PCTK2
	IRAK2			RRAGA	SPTA1	PDE6C
	ISG15			RS1	TAS2R39	PLA1A
	ITGB6			SLCO2A1	TLR5	PRF1
	JUB			STON1	TRPC3	PTGFRN
	KIAA0372			STS	TXNDC3	PTX3
	KRT15			SYT4	VGLL2	RASGRF2
	KRT34			TAC4	ZC3H6	SCN8A
	LGALS7			TARP	ZNF658B	SEPP1
	LHB			TEKT4	ZNF665	SERPINB1
	LOC553158			THEM5		SIDT1
	MAGEH1			TMEM162		SIGLEC5
	MAP2K4			TRIM21		SIRT1
	MAPK4			UBR1		SLC17A8
	MAST3			UNC13A		SLC22A18
	MFAP4			ZBTB38		SLC26A2
	MGC50722			ZNF43		SLC2A2
	MICALCL			ZNF780B		SLC31A1
	MIPEP					SLCO4C1
	MORC2					SNTA1
	MSH2					TAL1
	MSI1					TRY1
	MYO18A					UGCGL1
	MYO1A					XRCC1
	NLRC3					ZFP36
	NPR1					ZNF282
	NTSR1					
	NUCB1					
	OR4F17					
	OTX1					

PAK2
PCSK5
PEX12
PEX19
PHOX2A
PI16
PIGV
PIK3C2G
PPP2R1A
PSD2
PSMB4
PTGS1
RAD23A
RBM16
RNF10
RNF145
RUFY4
SAFB
SALL1
SCUBE3
SERINC2
SERINC5
SERPINA5
SH3PXD2B
SLC14A1
SLC22A18
SLC45A1
SMC3
SNAPC1
SPATA1
SPATA21
SPERT
SPR
SREBF2
SYNC1
TBC1D10C
TEF
TEX264
TFR2
TKTL1
TLE2
TLR5
TMEM175
TPCN2
TRADD
TRIM65
UGT1A8
UPK3A
USP54
VMO1
WDR27
WDR34
WDR90
XPC

ZFP36L1

ZNF289

ZNF324B

ZNF43

ZNF653

ZNF768

ZRSR2

APPENDIX 3

Branch-site analysis of individual domains in nuclear receptors: names of species under positive selection

Gene	Domain									
AR	AB	MOD	OCU	HUM	MIC	ETE				
ESR1	AB	MLU	BTA	STO	OGA	ETE				
ESR2	AB	LAF								
ESRRA	AB	CAF	OCU							
ESRRB	AB	BTA								
ESRRG	AB	OAN								
HNF4A	AB	MLU	EEU	CPO	MMU	PTR				
NR1D2	AB	MOD								
NR1H3	AB	OAN	MMU							
NR1I2	AB	SUS	STO	OGA	ETE					
NR2C1	AB	DNO	MLU							
NR2C2	AB	OCU	LAF							
NR2E1	AB	MLU	SAR	TBE	MIC					
NR2F1	AB	BTA	CAF	EEU	SAR	STO	MUS	OPR	TBE	PTR
NR2F2	AB	OPR								
NR2F6	AB	MOD	CAF							
NR3C1	AB	DNO	FCA	STO	TBE	MIC	ETE			
NR3C2	AB	MOD	EQC	OCU	MIC	ETE				
NR4A1	AB	OCU	OPR							
NR4A2	AB	SAR								
NR4A3	AB	OAN	CAF	EEU	OCU	MIC				
NR5A2	AB	EQC								
NR6A1	AB	HUM								
PGR	AB	BTA	CAF	STO	OPR	TBE	PTR	OGA	MIC	
PPARA	AB	OAN								
PPARG	AB	CAF	OPR							
RARB	AB	MUS	MMU							
RARG	AB	CAF	EEU	MUS	PTR					
RORA	AB	MLU	OGA							
RXRB	AB	MOD	EQC	OGA	ETE					
RXRG	AB	OGA								
VDR	AB	CAF								
ESR2	C	LAF								
ESRRB	C	MMU	LAF							
HNF4A	C	EEU	STO							
HNF4G	C	STO								
NR1D1	C	MMU								
NR1D2	C	STO								
NR1H2	C	STO								
NR1H4	C	MLU								
NR1I3	C	MLU	OGA							
NR2E1	C	OAN	OCU	MIC						
NR2F1	C	BTA								
NR2F6	C	CAF								
NR3C1	C	OCU								
NR3C2	C	SAR	TBE	MMU						

NR4A2	C	LAF			
PGR	C	LAF			
PPARA	C	CAF			
RARA	C	EEU	CPO		
RARB	C	DNO	MMU	LAF	
RXRB	C	CAF	EEU	TBE	
RXRG	C	DNO			
THRA	C	EEU			
AR	D	STO	ETE		
ESRRA	D	DNO			
NR0B1	D	OCU			
NR0B2	D	ETE			
NR1D1	D	MMU			
NR1D2	D	MOD	MLU	STO	CPO
NR1H2	D	STO	HUM		
NR1H3	D	FCA			
NR1I2	D	CAF			
NR1I3	D	MLU	CAF		
NR2C2	D	FCA			
NR2E1	D	MIC			
NR2E3	D	OGA			
NR2F2	D	OGA			
NR2F6	D	CAF			
NR3C1	D	FCA			
NR3C2	D	STO			
NR4A2	D	MLU	OCU		
NR5A1	D	OAN	OPR		
NR5A2	D	MLU	OGA		
NR6A1	D	MOD	OCU		
PPARD	D	TBE			
PPARG	D	SAR	TBE	OGA	MIC
RORA	D	OAN			
RORB	D	OAN	CAF	OPR	
RORC	D	OAN	MOD		
RXRB	D	MOD	CAF		
RXRG	D	CPO			
THRA	D	OGA			
THRB	D	OAN	SAR		
VDR	D	CPO	PTR		
AR	E	CPO			
ESR1	E	OAN	RNO		
ESR2	E	EQC			
ESRRA	E	TBE			
ESRRB	E	MOD	OCU		
ESRRG	E	DNO	OCU	OGA	
HNF4A	E	STO			
HNF4G	E	OAN	EEU		
NR0B2	E	OAN	MIC		
NR1D1	E	MMU			
NR1H2	E	BTA	MIC		
NR1H3	E	CPO			
NR1H4	E	SAR			
NR1I2	E	MOD	OPR		
NR1I3	E	MLU	MUS		

NR2C1	E	OAN	OGA						
NR2C2	E	EEU	STO	MMU	OGA				
NR2E1	E	OAN							
NR2F1	E	MOD	TBE	MIC					
NR2F2	E	MLU							
NR3C1	E	FCA	MMU	OGA					
NR3C2	E	STO	RNO						
NR4A2	E	DNO							
NR4A3	E	BTA	MIC						
NR5A1	E	OAN	STO						
NR5A2	E	OGA	MIC						
NR6A1	E	MOD	DNO						
PGR	E	BTA							
PPARA	E	OAN	STO	MMU	OGA				
PPARD	E	CAF	MMU	OGA					
PPARG	E	OAN	BTA	TBE					
RARA	E	CAF	HUM	OGA					
RARB	E	FCA	ETE						
RARG	E	OAN	CPO						
RORA	E	DNO							
RORB	E	MOD	CAF	SAR	LAF				
RORC	E	OAN							
RXRA	E	CAF	FCA	RNO	OGA	LAF			
RXRB	E	MOD	CAF	LAF					
RXRG	E	FCA	STO	RNO					
THRA	E	OAN	CAF	OGA					
THRB	E	BTA							
VDR	E	OAN	PTR						
ESRRB	F	STO							
NR1H3	F	OAN							
NR3C2	F	RNO							
RARA	F	DNO	EEU	SAR	STO	OPR	TBE	MIC	LAF
RARB	F	EQC	OGA	MIC					
THRA	F	MOD	BTA	RNO					