Universidade de Lisboa
Faculdade de Ciências
Departamento de Informática

Signatures of natural selection in the adaptive immune system of primates

Bruno Eduardo Vasques Costa

Dissertação
Mestrado em Bioinformática e Biologia Computacional
Especialização em Biologia Computacional

2015

Universidade de Lisboa
Faculdade de Ciências
Departamento de Informática

Signatures of natural selection in the adaptive immune system of primates

Dissertação
Mestrado em Bioinformática e Biologia Computacional
Especialização em Biologia Computacional

Bruno Eduardo Vasques Costa

Orientador:
Prof. Doutor Octávio Fernando de Sousa Salgueiro Godinho Paulo (FCUL)
Prof. Doutora Hélia Cristina de Oliveira Neves (FMUL)

2015

1 Nota prévia

A presente dissertação foi escrita na língua inglesa, de forma a facilitar o processo de publicação de resultados sendo que esta é a língua oficial de disseminação de conhecimento pela comunidade científica.

2 Agradecimentos

Gostaria de agradecer ao Professor Doutor Octávio Paulo, à Professora Doutora Hélia Neves e à Professora Doutora Rita Zilhão, por me terem orientado durante este longo percurso. Agradecendo ao Professor Octávio pela frescura de abordagens a cada problema que surgia na construção desta tese, dando um apoio valioso para me apontar no caminho certo com otimíssimo, entusiasmo e segurança que o trabalho que estava a realizar tinha valor. Às Professoras Rita e Hélia pelo apoio que me deram a entender os fundamentos teóricos adjacentes ao trabalho prático, integrando-me no seu grupo de trabalho. Onde todas as semanas eram apresentados trabalhos com genes envolvidos na minha tese que me deram uma visão ampliada sobre o trabalho laboratorial relativamente ao trabalho *in sílica*. Todos este processos contribuíram enormemente para a minha evolução como cientista.

Quero agradecer à minha namorada:
- Raquel, muito obrigado pelo incentivo e apoio moral em momentos de dúvida deste longo processo de escrita, que não foi fácil.

Aos meus pais por acreditarem em mim e me darem o apoio necessário para realizar esta tese.
Aos colegas do CoBiG² pelo imenso e constante incentivo e resposta às minha dúvidas.
Ao Francisco pela paciência contínua, por me ensinar a trabalhar com um novo sistema operativo, me fornecer as bases para aprender a trabalhar em linha de comandos e apoio em imensas outras dúvidas técnicas, um amigo sempre disposto a ajudar. E à Joana e à Telma pela ajuda valiosa na análise dos resultados. Graças a vocês consegui exprimir com maior clareza as ideias expostas neste trabalho. Revelaram-se grandes amigas ao longo dos bons e especialmente nos maus momentos.

Um bem haja a todos que me deram um grande apoio e possa não ter enumerado.

3 Resumo

O desenvolvimento de uma resposta imunitária adaptativa possibilitou aos vertebrados montar uma defesa mais eficaz em resposta a agentes patogénicos. O sistema imunitário adaptativo tem a capacidade de reconhecer e guardar memória de agentes patogénicos específicos, conferindo ao sistema um poder de resposta mais rápido e eficaz aquando de uma reinfecção.

Com o sistema imunitário adaptativo surgem novas células, com o papel central na resposta imunitária. São exemplo dessas células, os linfócitos T e B, que produzem respectivamente os receptores das células T e B. Os receptores dos linfócitos T possuem grande capacidade de rearranjo das suas cadeias (α e β) e surgem *de novo* nos vertebrados mandibulados e em paralelo com o aparecimento dum novo órgão linfoide primário, o Timo. Este orgão é responsável pela maturação dos linfócitos T e tem a capacidade de eliminar linfócitos T autoreativos (isto é, que reconhecem o próprio).

Dada a importância do sistema imunitário adaptativo nos vertebrados, foi objectivo do presente estudo a analise bioinformática de um conjunto de 38 genes intimamente ligados ao desenvolvimento do sistema imunitário adaptativo. Estes, estão compreendidos no "processo de desenvolvimento do timo" e "processo do sistema imunitário", e foram analisados em busca de assinaturas de seleção positiva, através da aplicação de modelos estatísticos (PAML), que estimam pelo método de máxima verosimilhança, o rácio (ω) de mutações não sinónimas (d_N) versus sinónimas (d_S) .

No presente estudo, em genes ortólogos, de 11 espécies de primatas (incluindo *Homo sapiens*), encontraram-se sinais de seleção positiva em 7 genes que, após estudos complementares, foram reduzidos a 4 genes: CD4, IFNG, HOXA3 e PTCRA.

O mapeamento dos aminoácidos selecionados positivamente, por inferência Bayesiana, nas suas estruturas terciárias ou quaternárias, revelou que os aminoácidos selecionados positivamente se encontravam predominantemente na região de superfície, da respectiva proteína. Isto leva à formulação da hipótese, de que a superfície da proteína poderá estar sujeita a menores pressões seletivas purificantes do que o seu interior. Neste cenário, uma mutação terá menor impacto na conformação tridimensional, aquando do enrolamento da estrutura primária.

A ferramenta bioinformática SIFT, revelou que os aminoácidos selecionados positivamente surgem predominantemente em zonas putativamente menos conservadas.

Os resultados do presente estudo, sugerem que os genomas tidos como completos apresentam ainda zonas com baixa qualidade, ou baixa cobertura, que irão beneficiar grandemente da integração de *reads* produzidas pelos sequenciadores de 4ª geração, como a tecnologia *Nanopore*.

Palavras-Chave: Timo; Sistema imunitário adaptativo; PAML; Seleção Positiva; Genes ortológos; Primatas

4 Abstract

The emergence of an adaptive immune response has enabled vertebrates to respond more effectively to pathogenic infection. The adaptive immune system has the ability to recognize and memorize specific pathogens, allowing stronger responses each time the pathogen is encountered. In the adaptive immune system, T and B-lymphoid cells are central players, producing T-cell and B-cell receptors, respectively. The T-lymphoid cells arise a second time in vertebrates in the jawed lineage. These cells display a more random recombination process of the α and β chains of their receptors, which is followed by coevolution of a primary lymphoid organ (thymus), essential for the development T-lymphoid cells, allowing the elimination of self-reacting cells.

Given the importance of the adaptive immune system in vertebrates, the present study aimed to analyze, from a bioinformatics perspective, a set of 38 genes annotated to "thymic development process" and "immune system process" GO terms. These genes were studied in order to find signatures of positive selection. To accomplish this, a statistical model (PAML) was applied to estimate the ratio (ω) of nonsynonymous (d_N) versus synonymous substitutions (d_S), through maximum likelihood.

In the present study, in a set of orthologous genes, of 11 primate species (including *Homo sapiens*), signals of positive selection were found in 4 genes: CD4, IFNG, HOXA3 and PTCRA.

The amino acids identified with positive selection, through Bayesian inference, were mapped to their tertiary and quaternary structures, revealing that these were predominantly located on the protein surface. This leads to the formulation of the hypothesis that the protein surface is under lower purifying selective pressure than its core, with the consequent reduction of impact on the protein folding. The positively selected amino acids were mainly in regions putatively non-damaging or less conserved as predicted by the SIFT tool. This study brings to light problems in the so called complete genomes, that still bear regions of low quality, or low coverage, which will greatly benefit from fourth generation sequencing technology, like Nanopore.

Keywords: Thymus; Adaptive immune system; PAML; Positive selection;
Orthologous genes; Primates

5 Table of Contents

1 Introduction

1.1 General introduction

The evolution from single cell organisms into multicellular organisms, could not have occurred without a mechanism that allowed unicellular organisms to distinguish between food and other unicellular organisms of its kind, or even another part of itself[1,2]. This is accomplished by one of the most basic functions of the immune system: the ability to recognize specific surface receptors[3].
With the evolution of ever more complex organisms, comes the need for a stronger immune response to infection. The emergence of the adaptive immune system must have been a game changer on the fight against pathogens, bringing the ability to recognize and remember specific pathogens, allowing stronger responses, each time the pathogen is encountered.
T-cells become central players in the immune system. The T-cell progenitors depend on the interaction with epithelial cells (thymic epithelial cells) of a lymphoid organ (thymus), in order to develop into mature functional antigen specific T-lymphocytes[4].
Thus, the study of genes involved in the development of the adaptive immune system was indispensable to help provide answers to the question of how the adaptive immune system has been shaped in the last million years of evolution.

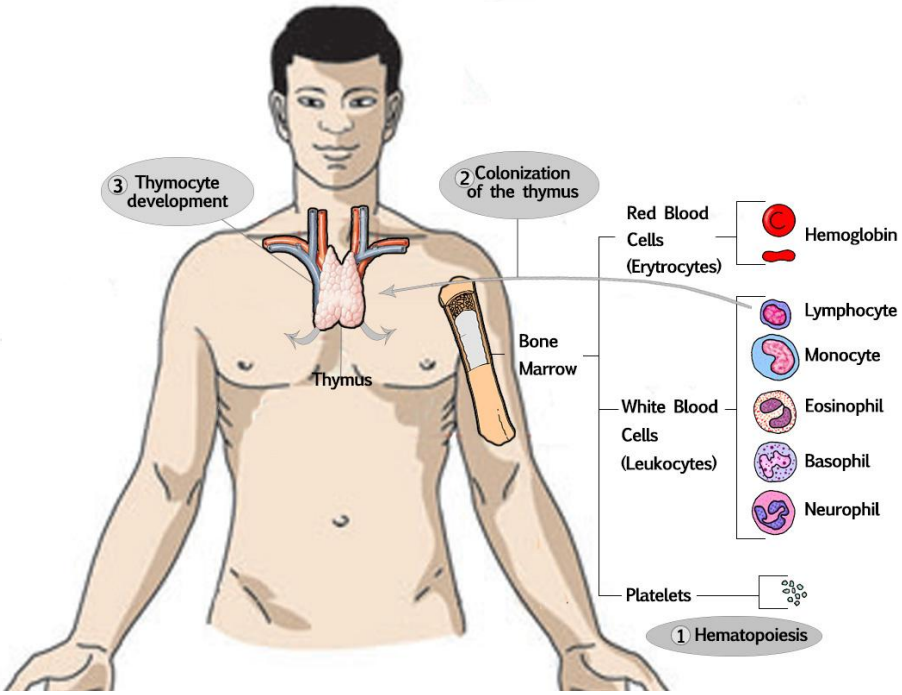


Figure 1.1.1 - Processes undergone for the formation of a Mature T Cell.

1.2 Immune system and how it responds to infection

The immune system is responsible for eliminating disease-causing microorganisms (pathogens), such as viruses, bacteria, fungi and parasites, and can be divided into innate and adaptive immune systems.

> The **innate immune system** provides a nonspecific rapid general response that doesn't improve with repeated exposure.
>
> The **adaptive immune system** produces a slower, highly specific and long lasting response.

While both immune systems are capable of distinguishing between self and non-self, the innate system, has a limited number of receptors encoded from the germline, which are able to recognize common features in pathogens. Conversely, the adaptive immune system uses a process of somatic cell rearrangement to generate a wide repertoire of antigen receptors[5].

1.2.1 Adaptive immune system function

The adaptive immune system comes into action, when the innate immune system alone, is incapable of dealing with an infection. Its major advantage over the innate immune system is its specificity, which allows a more targeted response against the pathogens and therefore is able to remove the threat with greater ease. This response produces antibodies (secreted by B-lymphocytes) and activated T-lymphocytes, which persist after the infection is eliminated and confer protection against reinfection[6].

1.2.2 Cells in the adaptive immune system

The main cells involved in the innate immune system, are granulocytes and dendritic cells, which originate from the bone marrow derived common myeloid progenitor. While in the adaptive immune system, the main cells are lymphocytes, derived from the common lymphoid progenitor. Both lineages are derived from bone marrow hematopoietic stem cells.

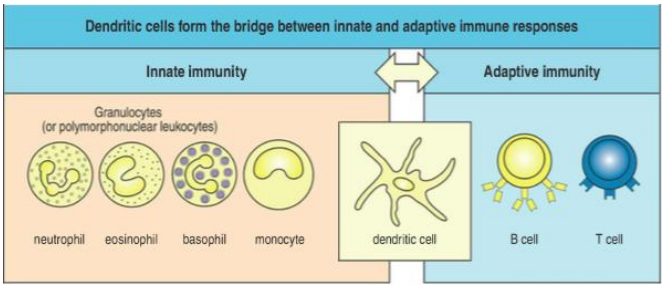


Figure 1.2.1 - Main cells involved in the adaptive immune system.
Adapted from K. Murphy 2011

There are two types of lymphocytes, B lymphocytes (B cells) and T lymphocytes (T cells). The B cells proliferate and differentiate into antibody producing plasma cells when an antigen binds to its B-cell receptor (BCR).
Mature naïve T cells circulate between blood and peripheral lymphoid tissues, until they encounter their specific antigen[7]. This encounter induces the proliferation and differentiation into effective T-cells (or activated). Activated T-cells differentiate into Cytotoxic, Helper or Regulatory effector T-cells depending on their cell markers.

In the peripheral lymphoid organ, the antigen (short peptide fragments of protein antigens) is presented by the major histocompatibility complex (MHC) on dendritic cells[8] (host cell). MHCs are transmembrane glycoproteins, that have a gap in the extracellular face of the molecule, where peptides can bind.
T-cells recognized the MHC presented antigen by the T-cell receptor (TCR). TCRs are membrane bound proteins, related to immunoglobulins, having both variable (V) and constant (C) regions. They are associated with an intracellular signaling complex.

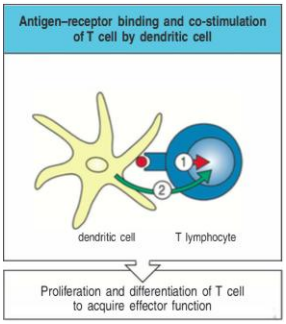
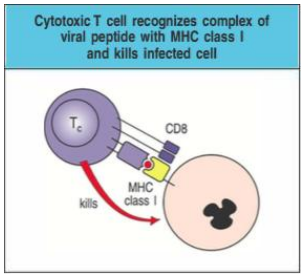


Figure 1.2.2 – Activation of a T cell by an antigen presenting dendritic cell. Adapted from K. Murphy 2011



The cytotoxic T cells express the CD8 and MHC (Class I) markers and kill infected cells that present its specific antigen. The cytotoxins (stored in specialized cytotoxic granules) released by CD8 cells can penetrate the lipid bilayer and trigger apoptosis in the target cell[9].

Figure 1.2.3 – Citotoxic T cell identifying an infected cell. Adapted from K. Murphy 2011

The Helper T cells (T_H1, T_H2, T_H17, T_{FH}) express the CD4 and MHC (Class II) markers and activate their target cells, or Regulatory T cells, which help control the immune response[10]. This cellular communication is mainly mediated by cytokine molecules.

	CD8 cytotoxic T cells	CD4 T_H1 cells	CD4 T_H2 cells	CD4 T_H17 cells	T_{FH} cells	CD4 regulatory T cells (various types)
Types of effector T cell						
Main functions in adaptive immune response	Kill virus-infected cells	Activate infected macrophages Provide help to B cells for antibody production	Provide help to B cells for antibody production, especially switching to IgE	Enhance neutrophil response Promote barrier integrity (skin, intestine)	B-cell help Isotype switching Antibody production	Suppress T-cell responses
Pathogens targeted	Viruses (e.g. influenza, rabies, vaccinia) Some intracellular bacteria	Microbes that persist in macrophage vesicles (e.g. mycobacteria, *Listeria*, *Leishmania donovani*, *Pneumocystis carinii*) Extracellular bacteria	Helminth parasites	*Klebsiella pneumoniae* Fungi (*Candida albicans*)	All types	

Figure 1.2.4 - Effector T cells and their respective function. Adapted from K. Murphy 2011.

Cytokines, are small soluble proteins that can alter the behavior or properties of the secreting cell or others[11].

CD8 T cells: peptide + MHC class I		CD4 T cells: peptide + MHC class II							
Cytotoxic (killer) T cells		T_H1 cells		T_H2 cells		T_H17 cells		T_{reg} cells	
Cytotoxic effector molecules	Others	Macrophage-activating effector molecules	Others	Barrier immunity activating effector molecules	Others	Neutrophil recruitment	Others	Suppressive cytokines	Others
Perforin Granzymes Granulysin Fas ligand	IFN-γ LT-α TNF-α	IFN-γ GM-CSF TNF-α CD40 ligand Fas ligand	IL-3 LT-α CXCL2 (GROβ)	IL-4 IL-5 IL-13 CD40 ligand	IL-3 GM-CSF IL-10 TGF-β CCL11 (eotaxin) CCL17 (TARC)	IL-17A IL-17F IL-6	TNF CXCL1 (GROα)	IL-10 TGF-β	GM-CSF

Figure 1.2.5 – Cytotoxins and cytokines produced by each effector T cell. Adapted from K. Murphy 2011.

The main cytokine produced by CD8 Cytotoxic T cells is interferon gamma (IFNG/IFN-γ), which can block viral replication or lead to the elimination of the virus from infected cells without causing their death. CD4 T_H1 also secretes IFNG in order to active macrophages that weren't able to destroy ingested pathogens and has become incapacitated[12,13].

1.3 Lymphocyte T development

All lymphocytes derive form bone marrow hematopoietic progenitor cells however T-lymphocyte differentiation (lymphopoiesis) takes place in another lymphoid organ. T cell differentiation depends on the interaction of hematopoietic progenitors and immature T-lymphocytes (thymocytes) with the thymic epithelium, which shapes the mature repertoire of T cells in order to ensure self-tolerance.

1.3.1 Thymocyte maturation in the thymus

T-cell development depends on cell-cell interactions, between the thymocytes and the thymic epithelial cells, which are critical for the complete morphological and functional maturation of both cell compartments[14], and determine thymocyte cell fate. If lymphoid progenitors at the T-cell/B-cell branch point received NOTCH1 signaling, these cells are prone to differentiate into T cells, whereas if presented with NOTCH2 the tendency is to differentiate to B-Cells[15,16,17]. Lymphocytes that arrive at the thymus lack most cell markers characteristic of the T cells and their receptor genes aren't rearranged yet. However through interaction with the thymic epithelium they start to differentiation towards the T cell lineage pathway.

There are two major populations of lymphocytes α:β and γ:δ. The rearrangement of the β and γ:δ chains determine the cell fate. Lymphocytes that have rearranged and express the γ:δ TCR shut off β chain rearrangement[18]. If injected in to peripheral circulation these cells can give rise to B cells.
Lymphocytes that complete β chain rearrangement, shut off γ:δ rearrangement and commit the cell to the α:β lineage.
Thymocytes that go down the α:β pathway pass through various double negative (DN) stages based on the expression adhesion molecules CD44, CD25 and Kit.

DN1 – Both chains of the TCR are in the germline configuration.
DN2 – Rearrangement of β chain begins.
DN3 – Expressed β chains pair with a surrogate pre-T-cell receptor α chain (pTα/PTCRA), and form a pre-TCR[19]. The pTα immunoglobulin domains makes two important contacts that help with further rearrangement of the β chains. If β chain is incapable of pairing with the pTα the thymocytes is eliminated by apopthosis.
DN4 – Proliferation of cell with functional β chains.

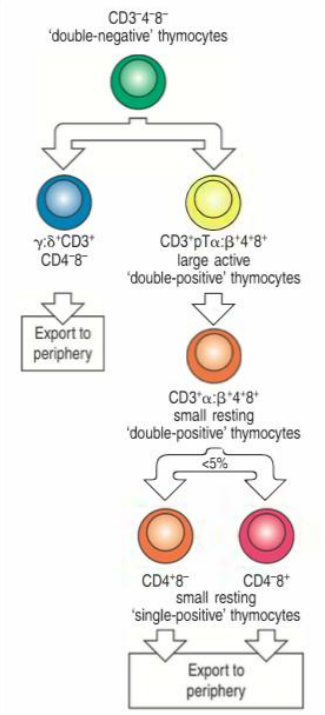


Figure 1.3.1 – Thymocyte fate based on TCR chain and surface cell marker expression. Adapted from K. Murphy, 2011.

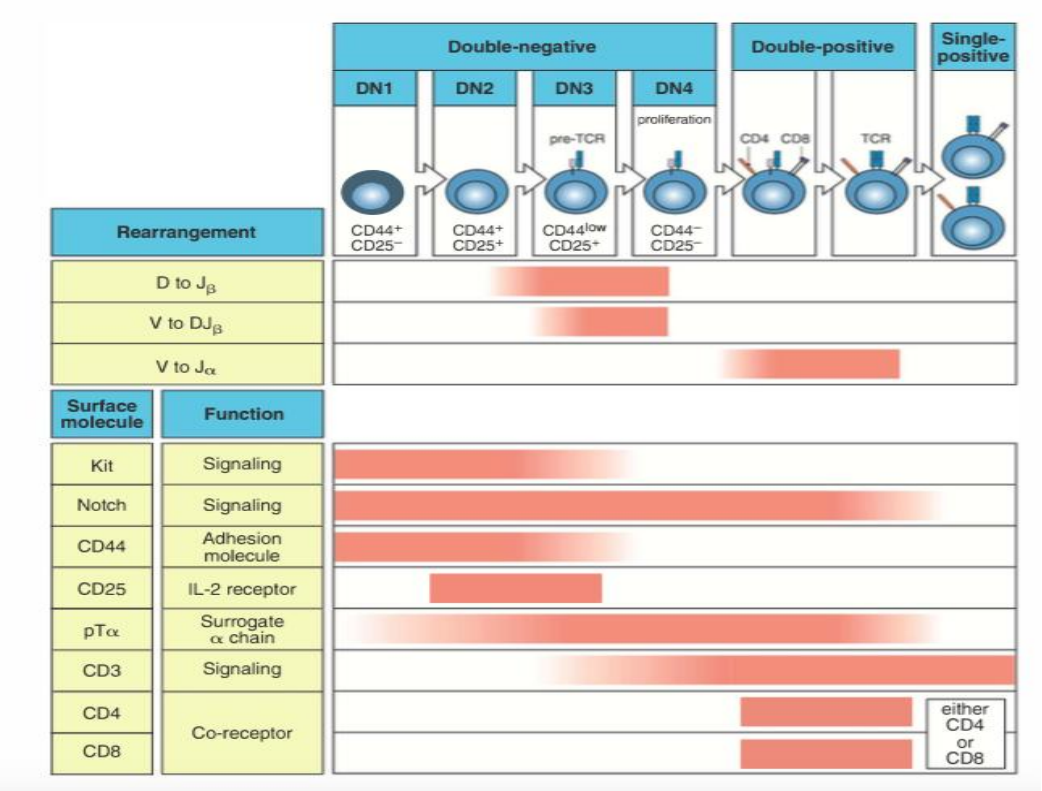


Figure 1.3.2 - Various phases of thymocyte development and surface marker expression timescale. In the Immunoglublin domain rearrangement, D stands for diverse genic segment, J for joining segment, V for variable segment and the subscript α/β refers to the chain.
Adapted from K. Murphy, 2011.

Subsequently, the DN thymocytes, express CD8 and CD4 and becomes double positive (DP). While β chains in DP cells cease further rearrangement, the pTα begins a series of rearrangement attempts to produce an α:β TCR.

Afterwards, the α:β TCR are positively selected by compatibility with self-MHC molecules[20]. Self-reactive receptors are given a death signal, which leads to their removal though cell death (negative selection).

Thymocytes that survive selection cease expression, of one, of the co-receptor molecules. Therefore, becoming either $CD4^{+}CD8^{-}$ or $CD4^{-}CD8^{+}$ single positive (SP) thymocytes, located in the medullar region that migrate to the periphery[21,22].

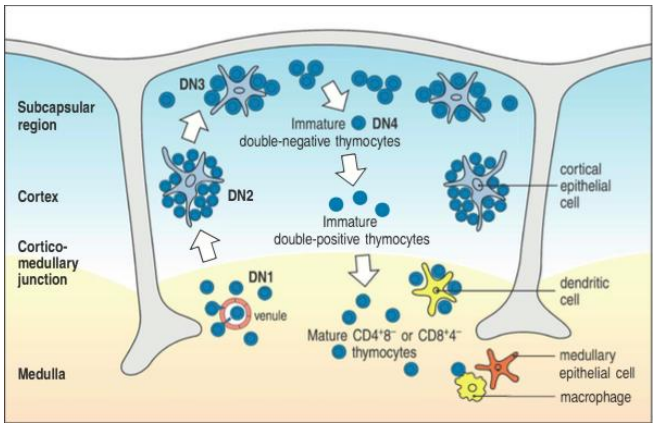


Figure 1.3.3 - Thymocyte maturation through interaction with thymic epithelium. Adapted from K. Murphy, 2011.

Since defects in the thymic epithelium can result in immunodeficiency or autoimmunity. The expression of many tissue-specific self-antigens requires the autoimmune transcription factor regulator AIRE[23]. AIRE interacts with many proteins involved in transcription, and is presumed to avoid termination of transcription from smaller promoter transcripts.

1.4 Thymus organogenesis

The thymus is composed by various lobules, and can be morphologically and functionally divided into an outer cortical region and an inner medulla. In young individuals, the thymus has a large number of developing T-cell precursors embedded in a network of epithelia. While in mature individuals, the development of new T cells in the thymus slows down, and numbers of these cells are maintained through long-lived individual T cells along with the division of mature T cells outside the central lymphoid organs.

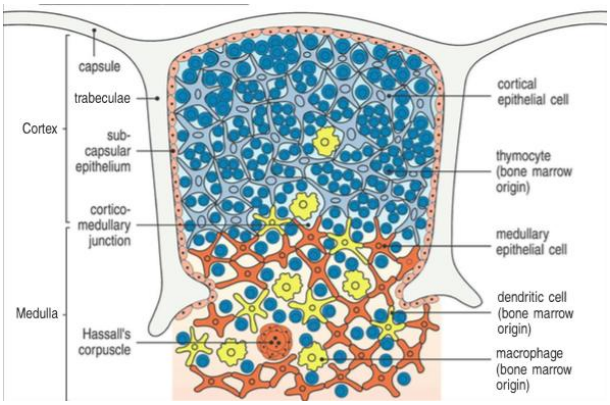


Figure 1.4.1 - The cellular network of the human thymus. Adapted from K. Murphy, 2011.

1.4.1 Thymus development in the mouse model and implied genes

Thymic epithelium rudiment arises early during embryonic development, from endoderm-derived segmented structures, the pharyngeal pouches (PP).

In mammalian and avian embryos four PP are produced. However in avian embryos (chick and quail) the thymus and parathyroid glands are derived from the third and fourth PP [24], and thus the thymus is formed from the third (3PP). Whereas in mammals, the thymus arises from a common primordium with parathyroid glands derived from the 3PP[25]. The formation of the 3PP both in mouse, as well as, in chick and quail are dependent on the expression of HOXA3 gene[26].

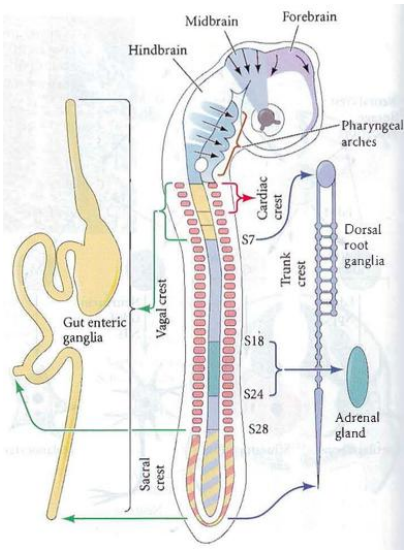


Figure 1.4.2 - Regions of the chick neural crest. Adapted from S. Gilbert, 2010[99].

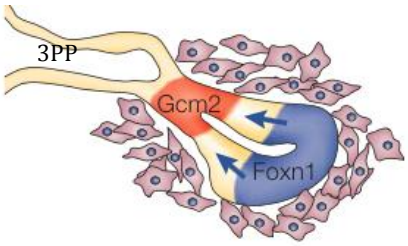


Figure 1.4.3 Segmentation of the third pharyngeal pouch by specific gene expression. Adapted from C. Blackburn, et al, 2004

In mammals, the 3PP begins to outgrow surrounded by a condensed population of neural crest cells (NCC) that will lead to the formation of the thymic capsule[27]. Shortly after, segmentation of the 3PP begins, though expression of GCM2 and FOXN1. GCM2 is responsible for the parathyroid[28] cell differentiation while FOXN1 is responsible for the thymus differentiation[29] Figure 1.4.2. FOXN1 is the earliest known thymus-specific marker.

The proliferation of the epithelium leads to the stratified organization of the thymus. Once the 3PP is completely patterned into both thymus and parathyroid domains, the thymic primordia is separated from the pharynx and begins to migrate to its final anatomical position[30].

These epithelial tissues form a rudimentary thymus, or thymic anlage that is ready to receive its first wave of thymocytes.

1.4.2 Colonization of the thymus by lymphoid progenitor cells

The colonization of the fetal thymus arises before its vascularization and occurs in two waves[31,32]. T cell precursors respond to a gradient of chemokines[33] (diffusible chemoattractant factors) that guide T-lymphoid progenitor cells out of the vasculature into the prevascular fetal thymus[34]. The second wave relies on the expression of FOXN1[34] to keep the constant in-flow of hemotopoietic precursors, into the thymus.

In the fetal primordium, thymic epithelial cells produces transcripts for several chemokines, such as CXCL12, CCL25 and CCL21[35]. However, the CXC12, or its

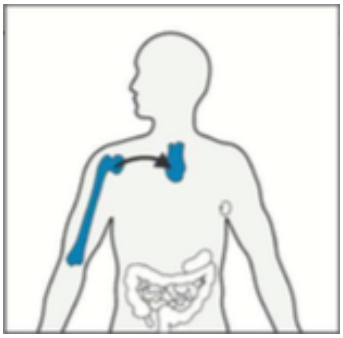


Figure 1.4.4 - Migration of lympoid progenitor cell to the thymus.
Adapted from K. Murphy 2011.

receptor CXCR4, mutant mice, are still able to colonize the thymic anlage with T-precursor cells[36]. In CCR9-deficient (a receptor for CCL25) mice, a threefold decrease in total thymocyte cellularity is exhibited when compared with wildtype animals[37]. Conversely, CCL21- and CCR7-deficient mice revealed that CCL21 is involved in the colonization of the prevascular fetal thymus[38].
Once the fetal thymus is fully vascularized, lymphocyte-progenitor cells have direct access to the thymus, via newly sprouted blood vessels[39], where integrins and CD44 are suggested to play a role in thymus seeding [40].

1.5 Emergence of thymopoiesis in Vertebrates

The adaptive immune system has only been documented in vertebrates and it has been shown to evolve independently in two basal vertebrates: the lineage that gave rise to jawless vertebrates, such as hagfish and lamprey, and the lineage that gave rise to all jawed vertebrates, represented by cartilaginous fishes[41].
While jawless vertebrates have lymphocytes with combinatorial diversity achieved through gene conversion, they still don't have an adaptive immune system. In jawed vertebrates, combinatorial diversity is achieved by VDJ recombination. Thus, the thymus emerges in the jawed lineage involved in the self-reactivity process, which would be problematic due to the diversity generated by VDJ recombination[42].

1.6 The immune system in primates

The immune system of nonhuman primates (NHP), shares a significant amount of homologous genes with the immune system of humans. The adaptive immune system of NHP species has been highly studied throughout the last decades and despite the similarity between species, the understanding of T cell repertoire dynamics is reduced. This is due to the lack of specific antibodies against human variable TCR and even less that cross-react with rhesus TCR[43].
The varieties of pathogens that invade NHP are generally similar to human, and therefore differences in the immune response can be investigated[44].
The study of genomic data of NHP will provide further insights into the immune system.
The high similarity and outbred nature of primates provides a great study model for further analysis of the adaptive immune function in order to provide new advances in the medical field[45].

1.7 Public Databases

Public databases of scientific data are becoming key tools for research in biology, especially in the field of bioinformatics, being essential for worldwide spread of information. Nowadays, bioinformaticians have a wide variety of public databases at their disposal: from nucleotide sequence databases, like GenBank[46], to whole genome databases, such as Ensembl[47]. There are also manually curated protein sequence databases, such as Swiss-Prot[48], or it's automatic counterpart – Uni-Prot, and even metabolic pathway and functional databases like KEGG[49] and many others.
There are three major worldwide molecular biology databases, the U*S National Center for Biotechnology Information* (NCBI) located in Bethesda, Maryland, USA, the

European Bioinformatics Institute (EBI) located in Hinxton, Cambridge UK, which is a part of the *European Molecular Biology Laboratory* (EMBL) and the *DNA Database of Japan* (DDBJ) operated by the *Center for Information Biology* (CIB) in Mishima, Japan. NCBI, EBI and CIB comprise the *International Nucleotide Sequence Database Collaboration* and synchronize their databases every 24h.

One of the best-known nucleotide sequence databases available at NCBI is GenBank. This database allows the query of billions of sequences and scientists can easily submit sequences to this database in order to accurately cite them in their publications though a unique record called *accession number*. NCBI uses the Entrez[50] system to allow users to query all NCBI associated databases and implements logical operators in queries. Another fundamental bioinformatics tool provided by the NCBI is the *Basic Local Alignment Search Tool* (BLAST)[51], which allows the calculation of sequence similarities, enabling the comparison of nucleotide and protein sequences to those available in entire databases.

Amongst the constant evolution of sequencing technology throughout the 90s due to the human genome project, the implementation of pyrosequencing in sequencing technology and its widespread to major laboratories lead to the passage from genetics to genomics. However, genomic sequencing technology became affordable due to the appearance of new companies in the sequencing market caused by the mass sequencing.

In 1999 the Ensembl project which is a joint project between EBI, and the *Wellcome Trust Sanger Institute* (WTSI) begins. Its mission is to provide automatically annotated genomes integrated with other available biological data.

The Ensembl 76 release is the latest available as of August 2014 and comprises a list of 79 species, all of which have their genome publicly available.

Subsequently, with the completion of the human genome sequence, in order to discover all crucial parts of the human genome biological function. The ***Enc****yclopedia* ***O****f* ***D****NA* ***E****lements* (ENCODE)[52], a public research consortium was launched in September 2003 by the National Human Genome Research Institute (NHGRI), in September 2003.

1.8 Evolution through mutation

Natural selection is one of the major evolutionary forces responsible for the diversity of organisms, making it one of the most important processes in biology. Identifying its action on the molecular basics, has become a current question and several statistical methods have been created to look for the "molecular footprints" left by Selection in genomes and protein sequences.
Molecular adaptation occurs due to the action of evolutionary forces of mutation, migration, natural selection, and genetic drift, which affect the allelic frequencies in a population. When a mutation arises, a new genetic variant appears in the populational genetic background. If it is advantageous it may become widespread and eventually fixate (positive selection). However since random mutations are typically deleterious, there is a constant purifying selection (negative selection) acting on mutations in order to remove them from the gene pool.
Besides positive and negative selection, balancing selection also acts to preserve multiple genetic variants within a population, for very long periods of time. In diploids, balancing selection[53] can be caused by **overdominance**, when the heterozygote at a particular locus is associated with greater fitness than both the homozygotes, thereby maintaining both alleles[54]. Alternatively, both haploids and diploids may display **frequency-dependent selection**, another form of balancing selection that occurs when a rare variant is associated with greater fitness than a more common one. Lastly, frequent environmental fluctuations, allow for multiple variants to be maintained since no single advantageous mutation has enough time to reach fixation, before the environment within which it is beneficial changes once again (fluctuating selection)[55].

1.8.1 Detection of natural selection

Molecular sequences encompass different types of information that can be used, individually or in combination, to infer the past action of selection[56]**.**

- **The frequency of observed polymorphisms,** depends on the action of selection and drift at a particular site. Deleterious mutations are more likely to be found at low frequencies since they are typically negatively selected, before they become widespread. Mutations that have become fixed are much more likely to represent neutral or beneficial changes[57].

- **The relative rate of silent and replacement fixations**.
  Non-synonymous nucleotide mutations are those that change the encoded amino acid, while Synonymous mutations maintain the coded amino acid unchanged, due to the redundancy inherent in the genetic code. A greater rate of fixation for Non-synonymous mutation, relative to the rate of Synonymous mutation, can be explained by action of positive selection.

- **Differences in genetic variation among genomic loci or among populations.**
  Fixation of a mutation by positive selection leads both to loss of genetic variation at the selected locus, but also at genetically linked loci that may be

evolving neutrally. Hence, the pattern of genetic variability among genomic loci can be used to infer selection in a recombining population. Similarly, differences in genetic variation at the same loci in different populations can also indicate action of natural selection.

1.8.2 Determining the occurrence of natural selection

To detect signatures of natural selection there are two major methods employed to detect positive selection, macroevolutionary methods and microevolutionary methods in which **summary statistic** methods are frequently used, to compare the observed frequency of polymorphisms with the null hypothesis of selective neutrality. The summary statistic is the simplest way to investigate selection, using a sequence alignment. Statistics that summarize the relative frequency of polymorphic sites are calculated from the alignment, they are then compared with the values expected to occur under a "null model" of neutral evolution. If the observed statistics are significantly different from their expected values, then the neutral model can be rejected.

1.8.3 Microevolutionary methods

Microevolutionary methods, focus on population genetics to identify positive selection within species.

Tajima's D summarizes the distribution of site frequencies of polymorphic sites.

Fay & Wu's H uses an outgroup sequence, from a closely related population or species to identify sites that have become fixed in the main study population.

1.8.4 Macroevolutionary methods

Macroevolutionary methods, which is the case of a method used in this study are used to identify past events of positive selection though comparative methods. Comparative methods use information of differences in genetic variation, among genomic loci, or among populations, frequently associated with the frequency of observed polymorphisms and the relative rate of synonymous and non-synonymous mutations.

McDonald-Kreitman test[58] measures the amount of adaptive evolution within a population by comparing it to an outgroup in order to distinguish fixations from polymorphisms. This is done by calculating the amount of polymorphisms in each species at neutral and non-neutral sites. A non-neutral site is one where the polymorphism is advantageous or deleterious, thus being prone to selection.
However, this test may be unreliable due to underestimation of degree of selection in presence of slightly deleterious mutations[59].

$\mathbf{D_N/d_S}$[60] methods, concentrate on the relative rate of synonymous and non-synonymous mutations[61] and is the method used to analyze the data in this study.

D_N/d_S methods, classify mutations in coding sequences as either synonymous or non-synonymous. Assuming that selection acts less strongly on silent mutations, observed differences between patterns of synonymous and non-synonymous mutations should reflect the action of natural selection.
If all non-synonymous mutations are neutral then, by definition the ratio of the two rates must equal one, indicating no selection.
D_N and d_S are calculated for every non-synonymous or synonymous site, taking into account the fact that random mutations generate more non-synonymous than synonymous mutations, due to the nature of the genetic code[62]. If the ratio is significantly greater than one, then positive selection is the most plausible scenario.
d_N/d_S methods are most successful in detecting adaptation when applied to genes that are under antagonistic co-evolution events, such as those generated by sexual conflict, predator-prey interactions, or host-parasite interactions[63].
However, this method is statistically weak and may fail to detect many instances of selection[64]. Notwithstanding, unlike the summary statistics introduced before, d_N/d_S methods do not require strong assumptions about the sampled population, and are therefore considered to be more robust.

1.9 Estimation of selection by maximum likelihood

The summary statistic (ω), which is a gene-based method is recurrently used to detect positive selection. One of the most used tools to calculate these ratio is PAML[65].
PAML is a **P**hylogenetic **A**nalysis by **M**aximum **L**ikelihood software that uses phylogenetic methods to preform comparative analysis of DNA and protein sequences by maximum likelihood. These phylogenetic methods are useful to estimate the evolutionary rates of genes and genomes, and to detect footprints of natural selection.

CODEML, a module of PAML, preforms comparisons and tests of phylogenetic trees by estimating parameters in sophisticated substitution models, including models for combined analysis of multiple genes. It estimates the synonymous (dS), and nonsynonymous (dN) substitution rates, permitting the detection of positive selection in protein-coding DNA sequences. To do this, CODEML uses to different methods: sites model, and branch site model.

1.9.1 Site models

The site models assume the same ω value for all branches. This method includes:

- M0 (one ratio) ignores chemical differences between amino acids and uses the same nonsynonymous/synonymous rate ratio ($\omega=d_N/d_S$) for all nonsynonymous substitutions;
- M1a (nearly neutral) assumes two site classes with $\omega_0=0$ and $\omega_1=1$, and does not allow sites with $\omega>1$;
- M2a (selection) adds a third site class and allows the presence of positively selected sites[66];
- M3 (discrete) allows three site classes ω_0, ω_1 and ω_2 that can take any value;
- M7 (beta) adopts a beta distribution for ω that is limited to the interval (0,1);
- M8 (beta & ω) adds one more site class to M7, with ω ratio estimated from the data[67].

For each model a log likelihood value is calculated by maximum likelihood ℓ. This value enables a comparison of an alternative model (positive selection allowed: H_1) to the nested statistical model (no positive selection allowed: H_0), through a likelihood ratio test (LRT). The LRT is calculated through twice the log likelihood of the difference between the two compared models ($2\Delta\ell$). If H_1 estimates $\omega>1$ and the LRT is greater than the critical values of the chi-square distribution with the appropriate degree of freedom (d.f.), then positive selection can be inferred. There are three pairs of models used to detect positive selection where a null model that doesn't allow $\omega>1$ is compared against a more general model that does:

H0: Uniform selection among sites (M0)
H1: Variable selective pressure among sites (M3)

H0: variable selective pressure but NO positive selection (M1)
H1: variable selective pressure with positive selection (M2)

H0: Beta distributed variable selective pressure (M7)
H1: Beta plus positive selection (M8)

When the likelihood ratio tests suggests positive selection, the Bayes empirical Bayes (BEB)[68] method can be used to calculate the posterior probabilities of each codon.

1.9.2 The branch-site models

The branch-site models assume different ω values among branches. This method includes different models to test particular lineages (foreground) for signals of positive selection:

- **Model 0**, applies one ω for all branches and is mainly used for site models or as a null hypotheses.
- **Model 1** (free ratios model), calculates separate ω for each branch in one run, however this model uses a big number of parameters.
- **Model 2,** allows the user to specify which branches to test for signals of positive selection, only allow one branch to be tested per run. In model 2 there are two tests implemented to check for branch specific positive selection[69], both of which use model A as the alternative hypotheses:
  - **Test 1,** uses as a null hypothesis the site model M1a (nearly neutral) that assumes two site classes with $0 < \omega_0 < 1$ and $\omega_1 = 1$, this test however can be misleading if relaxed selection acts on the foreground branch[70].
  - **Test 2,** uses as a null hypotheses branch-site model A with $\omega_2=1$ fixed. This allows sites evolving under negative selection on the background lineages to be released from constraint and to evolve neutrally on the foreground lineages.

Branch-site model A uses the parameters on Table 1.9.1.
By using test 2 it is possible to directly test weather a lineage evolves by positive selection if the null hypotheses is rejected based on the χ_1^2.

Table 1.9.1 - Branch sites Model A parameters

Site class	Proportion	Background	Foreground
0	p0	$\omega_0=0$	$\omega_0=0$
1	p1	$\omega_1=1$	$\omega_0=1$
2a	(1-p0-p1)p0/(p0+p1)	$\omega_0=0$	$\omega_2\geq1$
2b	(1-p0-p1)p1/(p0+p1)	$\omega_1=1$	$\omega_2\geq1$

1.10 Objectives

The main objectives of this work, was to study a network of cellular development genes related with the adaptive immune system in primates. This approach may shed light on how primates have evolved to adapt to pathogens and disease, and shed light on which genes are evolving due to positive selection. To accomplish this, three main goals have been set:

- Compilation of genetic data from public database for a large number of orthologous species, for a set of genes based on adaptive immune system development.
- Analysis of the presence of selection among the lineages of species, recurring to the selected set of genes
- Study the functional changes induced by the identified amino acid residues.

2 Methods

2.1 Collection and sorting of gene sequences from Ensembl

The identification of genes linked to thymus tissue development, was performed by searching Gene Ontology database (http://amigo.geneontology.org), for the go terms "thymus", "T cell" and "cytokine" in a list of 84 genes expressed by the QIAGEN Human Notch signaling pathway RT2 Profiler™ PCR Array. This search returned 24 genes, which were added to another 14 genes selected from the literature[4,71,24]. For each gene its respective coding sequences were downloaded from the Ensembl database (http://www.ensembl.org - release 76 - 15/08/2014), though the Ensembl API. A Perl script (https://github.com/netbofia/ensembl_sequence_getter) was used for this purpose. Once the human coding sequences were found, a list of orthologous genes (from primate species) was selected and their respective coding sequences, downloaded.

2.2 Selection of the species to be analyzed

After some tests with different types of species the final dataset was constructed using 11 different species of primates Table 2.2.1.

Table 2.2.1 - List of primate species.

Primate List	
Common name	Scientific name
Bushbaby	*Otolemur garnettii*
Chimpanzee	*Pan troglodytes*
Gibbon	*Nomascus leucogenys*
Gorilla	*Gorilla gorilla*
Human	*Homo Sapiens*
Macaque	*Macaca mulatta*
Marmoset	*Calithrix jacchus*
Mouse Lemur	*Microcebus murunus*
Olive baboon	*Papio anubis*
Orangutan	*Pongo abelii*
Tarsier	*Tarsius syrichta*

2.3 Go term enrichment clustering

All 38 human genes were blasted using "blastX" against the "nr" database using Blast2go. The most significant terms were complied and exported to SPSS to preform k-means clustering and hierarchical clustering.
The bioinformatics tool DAVID was used to obtain Go terms enrichment scores, and construct clusters.

2.4 Preparation for analysis by CODEML

All genes were filtered for primate species only and their transcripts were chosen based on size, using only one transcript per orthologous specie, where transcripts with a size difference, above 10%, from the chosen human transcript were excluded. All sequences were "blasted" using blastN against the "nt" database (accessed on 29/08/2014) to confirm their identity. Sequences were aligned with translatorX[72] tool using MAFFT[73] as the protein alignment method. The sequences were trimmed for stop codons using in house python scripts. Maximum likelihood phylogenetic trees were calculated with RAxML[74] using the GTRCAT model, with 1000 bootstrapped trees for each set of orthologous genes. The same process was conducted with Mr. Bayes to get the branch posterior probabilities.
All these methods were done by scripts, created specifically for this purpose (https://github.com/netbofia/paml_pipleline.git).

2.5 Detection of positive selection

CODEML from PAML 4.6[65] was used to test for positive selection. Firstly by applying to each set of orthologues genes, the free ratio model (Model: 1) where each branch is able to evolve at different omega ratios based on their calculated phylogenetic trees. Genes with evidence of positive selection were further analyzed with a series of branch specific models: first the omega ratio of previously identified branches was tested individually by indicating with a "#1" the branch under study in each run (Model:2 Nssites:0) and evaluating the statistical value compared to the one-ratio model of sites model (M0) through a likelihood ratio test. Then to confirm the previous test, a branch site model A (Test 2), (Model:2 Nssites:2) which allows sites evolving under negative selection on the background, was used. Furthermore, the sites model (Model:0 NSsites: 0.1,2.3,7.8) with 3 site classes (ncat=3) for M3 and 10 site classes (ncat=10) for M8 were also applied to see the distribution of selection among sites, throughout the originating protein. The models M3 vs. M0, M2 vs. M1 and M8 vs. M7 were tested through likelihood ratio tests, with d.f.=5, d.f.=1 and d.f.=1 respectively. The results were collected by a python script that clustered the relevant calculated parameters (dn/ds, kappa, omegas, probabilities, posterior probabilities and log likelihood) for each model into a table in excel per gene, along with the then calculated likelihood ratio test and respective p-value for each pair of models. Branch site model results (tress with nodes) were extracted with tree searching algorithms construted in python. This algorithm is recursive and makes recursive calls to itself, spaning a new instance for each branch until it reaches a tip. The collected information is merged with the branch in order to create the phylogenetic trees and to detect which genes have branches with values that merit further study (https://github.com/netbofia/paml_pipleline.git).

2.6 Functional analysis

The positively selected sites in genes under evolution were plotted against the average sift score calculated for all possible amino acid transitions. The sift scores were calculated on the SIFT[75] human protein webtool. The protein family domains were identified by Pfam a protein family database from EMBL-EBI[76].
The 3D human protein structures available were downloaded, for the protein sequences, in order to help in the visualization of the protein regions were the positively selected residues occurred. The protein structures were viewed using PyMol tool. When the protein structure wasn't available it was calculated though homology modeling, using similar proteins as templates with SWISS-MODEL[77].

2.7 CD4 Gene sequence validation

In order to confirm the chimpanzee sequence for CD4, the Ensembl chimpanzee mapped reads (.bam files) were downloaded from the Ensembl ftp server site. The chromosome 12 was extracted and visualized with samtools. Tables of nucleotide variation from the various reads were compiled using the BAM_to_TCS.py program from 4pipe4 tool (https://github.com/StuntsPT/4Pipe4, as of the commit `8ec3e53badbcdf97f940604095950686044edff7`).

The CD4 sequence was confirmed by blasting the Ensembl exon9 sequence against the reads of the chimpanzee genome downloaded from the Washington University server
(http://genome.wustl.edu/pub/organism/Primates/Pan_troglodytes/assembly/).
Then, to access their quality, its quality values were selected and surveyed using a in-house python script.

3 Results

3.1 Genes and species selection

In this study 38 genes Table 3.1.1 involved in tissue development of the immune system were used, with a total of 9412 GO terms, 7830 GO terms for biological process, 940 GO terms for cellular component and 642 terms for molecular function Table 3.1.2.

Table 3.1.1 Gene list with full name.

Gene symbol	Full name
ACKR2	Atypical Chemokine Binding Protein 2
ACKR3	Atypical Chemokine Receptor 3 Isoform x1
ADAM10	Disintegrin and Metalloproteinase Domain-containing protein 10
ADAM17	Disintegrin and metalloproteinase domain-containing protein 17
Aire	Autoimmune regulator
CCL25	C-C motif chemokine 25 isoform 2 precursor
CCR9	C-C chemokine receptor type 9 isoform x1
CD4	T-cell surface glycoprotein CD4 isoform 1 precursor
CD8	T-cell surface glycoprotein CD8 alpha chain isoform 2 precursor
CHUK	Conserved helix-loop-helix ubiquitous kinase
CTNNB1	Catenin beta-1 isoform x1
CXCL12	Stromal cell-derived factor 1 isoform x3
CXCR4	Chemokine (c-x-c motif) receptor 4
DLL1	Delta-like protein 1
DLL4	Delta-like protein 4
DTX1	E3 ubiquitin-protein ligase dtx1
FOXN1	Forkhead box protein n1
GCM2	Chorion-specific transcription factor gcmb
HES1	Hairly Enhancer of Split-1
HOXA3	Homeobox protein hox-a3
HOXB4	Homeobox protein hox-b4
IFNG	Interferon gamma
IL2RA	Interleukin 2 alpha
IL6ST	Interleukin 6 signal transducer (oncostatin m receptor)
IL17B	Interleukin 17b
JAG1	Jagged 1 (alagille syndrome)
JAG2	Jagged 2
LMO2	Rhombotin-2 isoform x1
NOTCH2	Neurogenic locus notch homolog protein 2 isoform 1
NOTCH4	Neurogenic locus notch homolog protein 4
NFKB1	Nuclear factor nf-kappa-b p105 subunit isoform x1
RUNX1	Runt-related transcription factor 1 isoform aml1b
RUNX1T1	Protein cbfa2t1 isoform x3
PTCRA	Pre t-cell antigen receptor alpha
PSEN1	Presenilin-1 isoform x1
PSEN2	Presenilin-2 isoform x1
SHH	Sonic hedgehog homolog
STAT6	Signal transducer and activator of transcription 6 isoform x1

3.2 Gene Selection

To detect genes, with accelerated evolutionary rates in the human specific branch the CODEML model 1 (Free ratio model) was applied to 38 genes off the 11 species of primates Table 2.2.1. From this, 7 genes revealed signals of positive selection - CD4, FOXN1, GCM2, HOXA3, IFNG, PTCRA and RUNX1T1 - that were further analyzed.

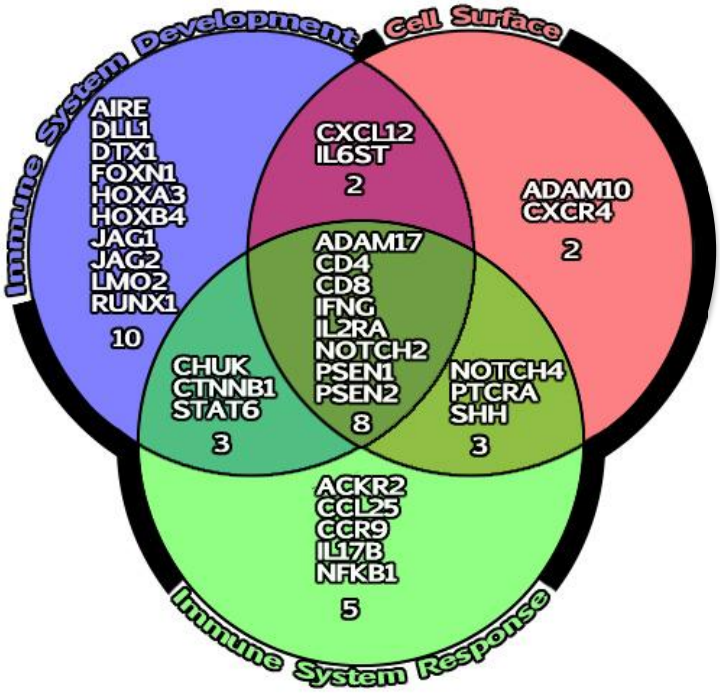


Figure 3.2.1 Venn diagram based on Go terms: Immune System response, Cell Surface and Immune system development.W

3.3 IFNG

(HGNC Symbol)

The free ratio model Figure 3.3.1 estimated that four branches are under positive selection, the Tarsier ancestral branch (Tarsier_p) with an ω ratio of 1.5703, the Marmoset ancestral branch (Marmoset_p) with an ω of 1.3478, the Gibbon ancestral branch (Gibbon_p) with an ω of 1.3092 and the macaque ancestral branch (macaque_p_p) with an ω of 1.4734.

Estimates were confirmed by testing each individual branch and preforming likelihood ratio tests, Table 3.3.1 both the Marsmoset_p (H4) and the Tarsier_p (H1) rejected the null hypothesis and estimate the omega ratio for the foreground branch at ω=2.17813 and 6.90653 respectively. While the Macaque_p and the Gibbon_p branch failed the likelihood ratio test.

The branch site model A Table 3.3.2 also followed the same pattern however upon correction of the values due to multiple testing the Tarsier_p *p*-value goes above 0.05 and fails the likelihood ratio test but has an ω ratio of 10.7596. The Marmoset_p branch has an estimate ω ratio of 27.57294. Both the Marmoset_p and the Tarsier_p branches detected positively selected sites though the BEB method both amino acid alignments of the positively selected site can be viewed in Figure 3.3.2 and Figure 3.3.3, respectively.

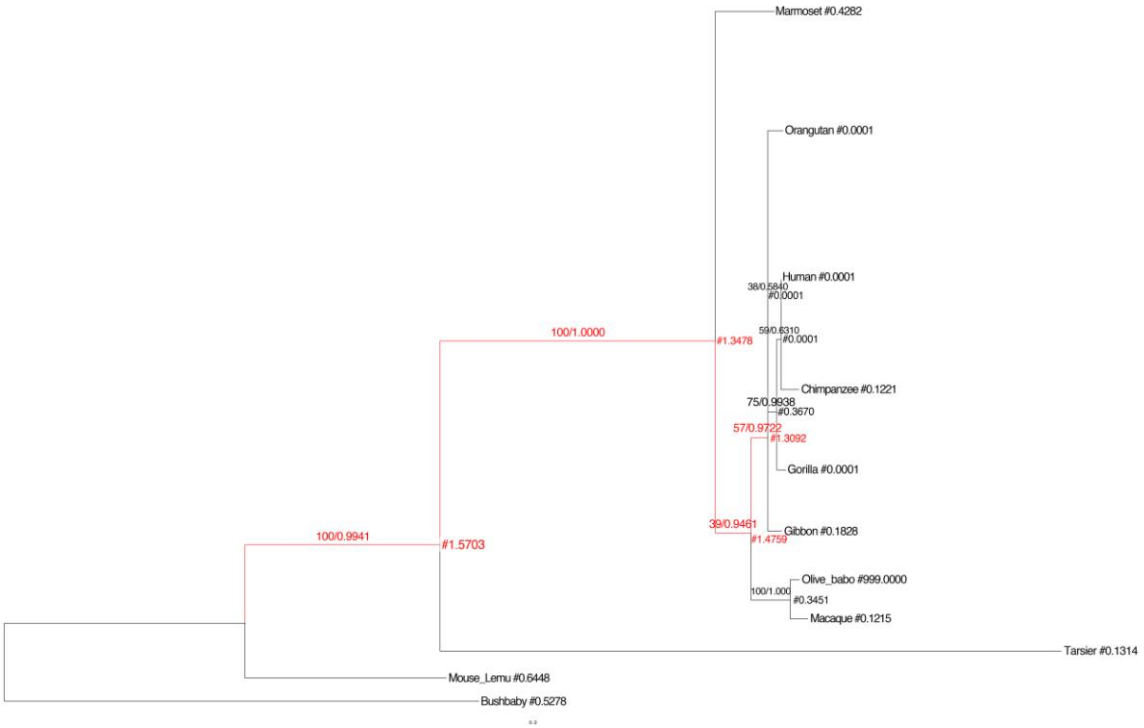


Figure 3.3.1 - Phylogenetic tree for gene IFNG with omega ratios on each node calculated by the free-ratio model in codeml. Bootsrtrap values calculated with RAxML and posterior probabilities calculated by Mr. Bayes are indicated on each branch respectively. Four branches that are under positive selection are indicated in red.

Table 3.3.1 - Parameter estimates under model of various omegas ratios among lineages and respective LRTs. LRT - Likelihood ratio test, FDR - False discovery rate correction, ω_b background omega. ω_f foreground omega.

Models	ω background	ω foreground	ℓ	LRT	*p*-value	FDR
H0: $\omega_b = \omega_{Bushbaby\_p} = \omega_{gibbon\_p} = \omega_{macaque\_p} = \omega_{tarsier\_p}$			-1905.4600			
H1: $\omega_{gibbon\_p} = \omega_{macaque\_p} = \omega_{marmoset\_p} = \omega_b \neq \omega_{tarsier\_p}$	0.36563	2.17813	-1902.1047	6.71058	0.009584255	0.0223929
H2: $\omega_{tarsier\_p} = \omega_{macaque\_p} = \omega_{marmoset\_p} = \omega_b \neq \omega_{gibbon\_p}$	0.42556	1.48708	-1905.0080	0.904092	0.34168686	0.3416869
H3: $\omega_{tarsier\_p} = \omega_{gibbon\_p} = \omega_{marmoset\_p} = \omega b \neq \omega_{macaque\_p}$	0.41768	1.78224	-1904.9690	0.981998	0.321706029	0.3416869
H4: $\omega_{tarsier\_p} = \omega_{gibbon\_p} = \omega_{macaque\_p} = \omega_b \neq \omega_{marmoset\_p}$	0.36688	6.90653	-1902.2431	6.433868	0.011196449	0.0223929

Table 3.3.2 - Branch site Model A estimates for Tarsier_p, Marmoset_p, Gibbon_p and Macaque_p, branches. LRT - Likelihood Ratio Test, FDR - False Discovery Rate correction applied to *p*-values.

Model A	ω background	ω foreground	ℓ	LRT	*p*-value	FDR
H_0: Tarsier_p	0.04568	1.00000	H_0-1883.640034	4.5609	0.03	0.06
			H_1-1881.359572			
H_0: Giboon_p	0.13090	1.00000	H_0-1884.533001	0.0451	0.83	0.83
			H_1-1884.510426			
H_0: Macaque_p	0.13208	1.00000	H_0-1884.406791	0.0791	0.78	0.83
			H_1-1884.367253			
H_0: Marmoset_p	0.13409	1.00000	H_0-1884.577027	7.0150	0.01	0.04
			H_1-1881.069509			

H_1: Tarsier_p				H_1:Giboon_p			
	Proportion	$\omega_{background}$	ω foreground		Proportion	$\omega_{background}$	ω foreground
Class site 0	0.48785	0.1098	0.1098	Class site 2a	0	0.13095	0.13095
Class site 1	0.31379	1.00000	1.00000	Class site 2b	0	1.0000	1.00000
Class site 2a	0.12072	0.1098	10.7596	Class site 2a	0.58683	0.13095	1.35164
Class site 2b	0.07765	1.00000	10.7596	Class site 2b	0.41317	1.00000	1.35164
H_1: Macaque_p				H_1: Marmoset_p			
	Proportion	$\omega_{background}$	ω foreground		Proportion	$\omega_{background}$	ω foreground
Class site 0	0	0.13244	0.13244	Class site 0	0.56201	0.13731	0.13731
Class site 1	0	1.00000	1.00000	Class site 1	0.29625	1.00000	1.00000
Class site 2a	0.60068	0.13244	1.79372	Class site 2a	0.09281	0.13731	27.57294
Class site 2b	0.39932	1.00000	1.79372	Class site 2b	0.04892	1.00000	27.57294

Species	29	48	57	65	83	88	102	108	109	116	134	137	150
											*		
Gorilla	K	N	K	R	F	S	V	N	K	E	H	I	G
Human	K	N	K	R	F	S	V	N	K	E	H	I	G
Mouse Lemur	-	G	K	K	H	T	A	S	R	Q	S	H	R
Gibbon	K	N	K	R	F	S	V	N	K	E	H	I	G
Bushbaby	A	G	K	-	H	T	A	S	Y	Q	S	Y	R
Chimpanzee	K	N	K	R	F	S	V	N	K	E	H	I	G
Orangutan	K	N	K	R	F	S	V	N	K	E	H	I	G
Tarsier	-	N	R	R	L	H	V	D	K	K	H	L	-
Marmoset	K	N	R	R	F	S	V	N	K	E	H	I	G
Macaque	K	N	R	R	F	R	V	N	K	E	H	I	G
Olive Baboon	K	N	R	R	F	R	V	N	K	E	H	I	G

Figure 3.3.2 - Alignment of Amino acid residues with positive selection on the Marmoset ancestral lineage according to branch site model A BEB analysis. * 0.95 < p.p. <0.99 - ** p.p. >0.99 .

Species	28	29	31	33	68	81	83	90	101	110	112	113	121	127	137	141	146	149	150	166
														*						
Gorilla_ENSGGOT00000006066	V	K	A	N	M	K	F	Q	N	K	R	D	Y	N	I	A	A	T	G	Q
Human_ENST00000229135	V	K	A	N	M	K	F	Q	N	K	R	D	Y	N	I	A	A	T	G	Q
Mouse_Lemu_ENSMICT00000004218	-	-	-	-	I	E	H	K	I	S	L	E	L	Q	H	N	R	Q	R	K
Gibbon_ENSNLET00000004455	V	K	A	N	M	K	F	Q	N	K	R	D	Y	N	I	A	A	T	G	Q
Bushbaby_ENSOGAT00000004301	S	A	I	Q	I	E	H	K	I	S	A	E	I	Q	Y	V	G	L	R	K
Chimpanzee_ENSPTRT00000009540	V	K	A	N	M	K	F	Q	N	K	R	D	Y	N	I	A	A	T	G	Q
Orangutan_ENSPPYT00000005616	V	K	A	N	M	K	F	Q	N	K	R	D	Y	N	I	A	A	T	G	Q
Tarsier_ENSTSYT00000001577	-	-	-	-	I	E	L	K	I	D	V	E	L	Q	L	L	R	L	-	K
Marmoset_ENSCJAT00000014072	V	K	A	N	M	K	F	Q	N	R	Q	D	Y	N	I	A	A	I	G	Q
Macaque_ENSMMUT00000027007	V	K	A	N	M	K	F	Q	N	K	R	D	Y	N	I	A	A	I	G	Q
Olive_babo_ENSPANT00000015498	V	K	A	N	M	K	F	Q	N	K	W	D	Y	N	I	A	A	I	G	Q

Figure 3.3.3 - Alignment of Amino acid residues with positive selection on the Tarsier ancestral lineage according to branch site model A BEB analysis. * 0.95 < p.p. <0.99 - ** p.p. >0.99

Site models analysis estimates that 1% of the amino acid sites, were under positive selection at an average ω ratio of 9.40 according to M3 and 8.25 according to M8 Table 3.3.3. The model M2a failed the likelihood ratio test. One amino acid residue with a posterior probability between 0.99 and 0.95 was detected and Figure 3.3.4. In order to visualize the distribution of the three classes identified by M3 a stacked histogram was plotted in Figure 6.0.2, where is possible to identify the selection sites.

Table 3.3.3 - Site model analysis, same omega for all branches, PPS – Positively selected sites, Likelihood, Likelihood ratio test, * 0.99 > p.p. >0.95 and ** p.p. > 0.99.

Model	dn/ds	kappa	Parameters				PSS	lnL	lnR	*p*-Value
M0 (one ratio)	0.43	3.27	ω= 0.43					-1905.46		
M1a(neutral)	0.49	3.45	p0= 0.59 ω0= 0.14	p1= 0.41 ω1= 1.00				-1885.51		
M2a(selection)	0.58	3.60	p0= 0.57 ω0= 0.14	p1= 0.42 ω1= 1.00	p2= 0.01 ω2= 9.65		2 *0 **0	-1882.71	5.597	0.0610
M3(Discrete)	0.57	3.58	p0= 0.55 ω0= 0.13	p1= 0.44 ω1= 0.96	p2= 0.01 ω2= 9.40		1 *1 **0	-1882.70	45.506	3.1205E-09
M7 (beta)	0.47	3.39	P= 0.37	q= 0.42				-1885.94		
M8(beta & ω)	0.55	3.52	p1= 0.01 p0= 0.99	ω= 8.25 P= 0.39	q= 0.43		5 *1 **0	-1882.74	6.390	0.0410

Species	112
	*
Gorilla	R
Human	R
Mouse_Lemur	L
Gibbon	R
Bushbaby	A
Chimpanzee	R
Orangutan	R
Tarsier	V
Marmoset	Q
Macaque	R
Olive_baboon	W

Figure 3.3.4 - Alignment of Amino acid residues with positive selection M3 NEB analysis. * 0.95 < p.p. <0.99 - ** p.p. >0.99

3.3.1 Functional Analysis

Sift scores for all the positions and the respective average of each position was calculated. Averages were used to plot the graph in Figure 3.3.5. Low sift score are associated to highly damaging (<0.1) mutations while higher scores are associated to tolerated mutations based on human protein structures[78]. Its possible to see the highest positively selected residues identified earlier, represented in zones with tolerated mutations.

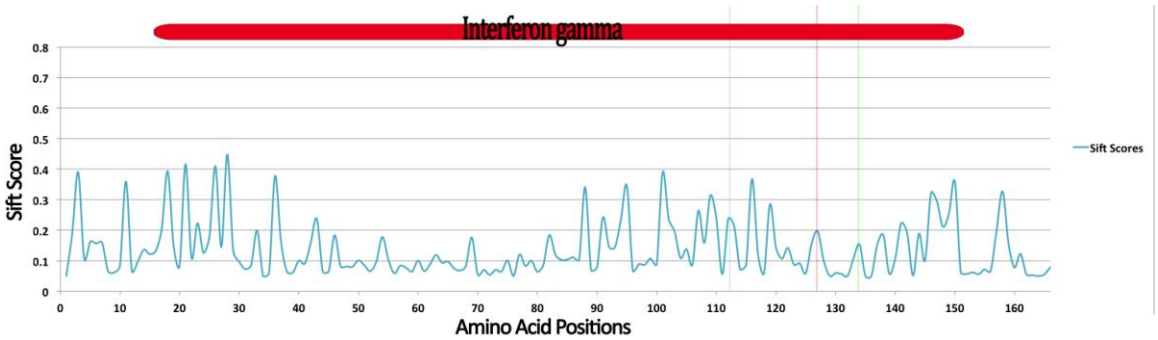


Figure 3.3.5 - Average sift score for each possible amino acid mutation throughout IFNG, with Pfam domain types positioned in red over graph and highest positively selected residues shown with colored vertical lines. Green line corresponds to the marmoset ancestral branch, red line corresponds to the tarsier ancestral branch and blue to the sites model M3 NEB posterior probability.

The tertiary structure was calculated by modeling the protein sequence to the 1eku.1[79] template of IFNG calculated by x-ray crystallography. The tertiary structure Figure 3.3.6 shows that the selected residues were found on the protein surface region.

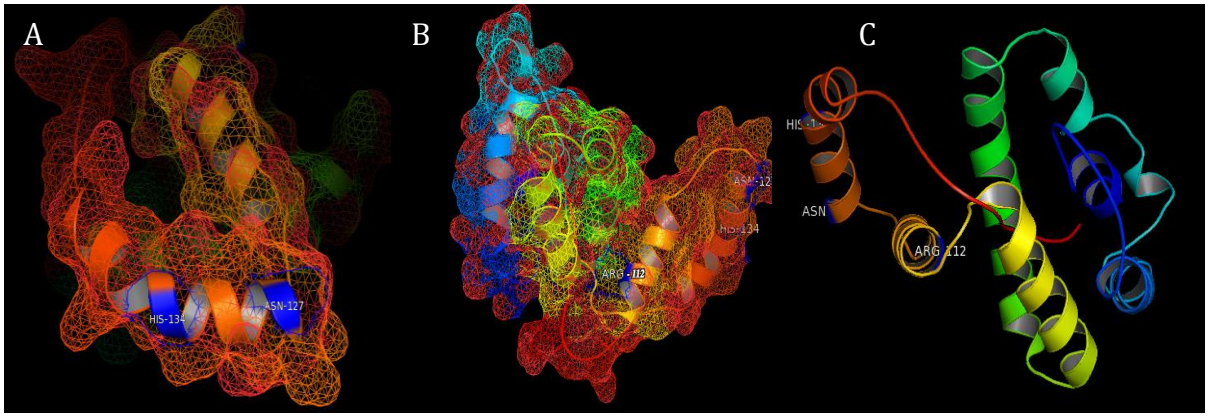


Figure 3.3.6 – IFNG tertiary structure with surface area rendered. A shows the a.a. residues HIS-134 and ASN-127 on the surface of the protein, B shows ARG-112 in region of the protein, C IFNG tertiary structure without surface area rendering

3.4 PTCRA

(HGNC Symbol)

The free ratio model estimated five branches Figure 3.4.1 to be evolving under positive selection: the Human ancestral branch (Human_p) with an omega ratio of 1.0704, the Gorilla branch with an omega ratio of 1.3474 the Macaque ancestral branch (Macaque_p) with an omega of 3.8731, the Gibbon branch with an omega ratio of 1.1572 and the Macaque branch with an omega ratio of 1.1621. Each branch was tested individually to correct for over estimation by the free ratio model. None of the tested branches passed the likelihood ratio test.
Besides that, testing the branch sites model A, Table 3.4.2 also suggests, that no particular branch has evolved though positive selection.
Nevertheless the site models Table 3.4.3 suggest that 31% of the amino acid residues evolved though positive selection had an average ω ratio of 1.81. An alignment of the amino acid residues, identified with positive selection though the NEB method, were plotted in Figure 6.0.2 and a stacked histogram with the distribution of the amino acids though the 3 class sites identified by M3 are represented in Figure 3.4.2, where it is possible to identify the selection sites.

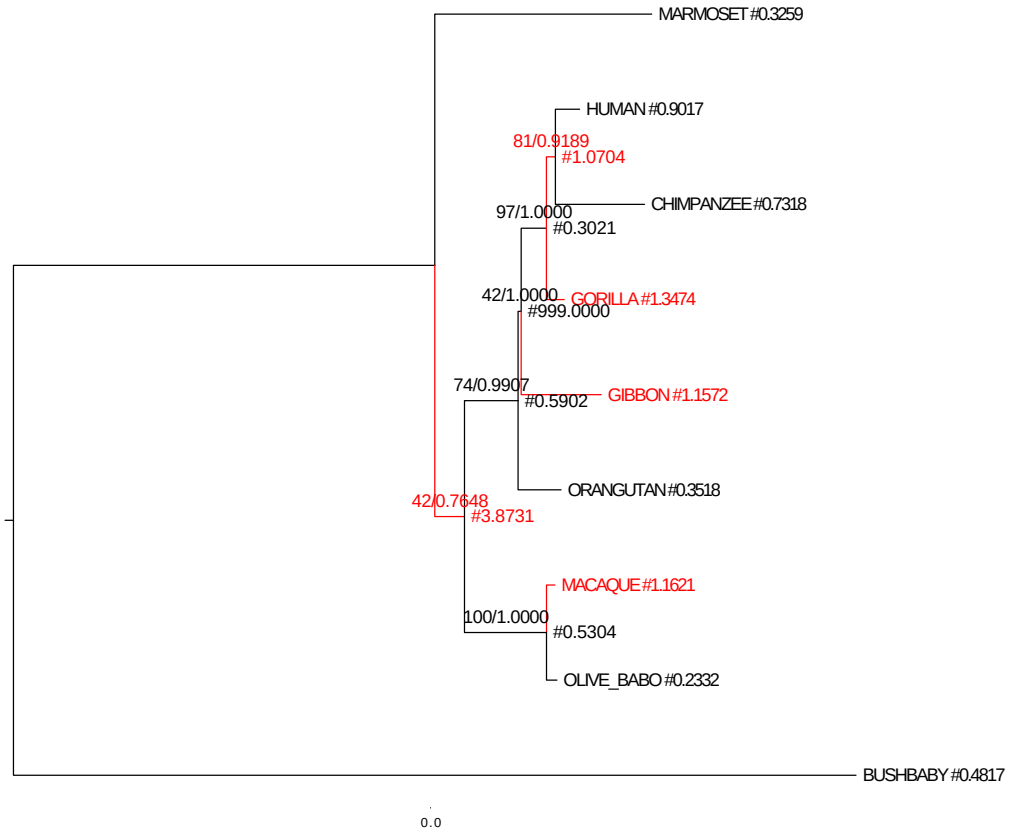


Figure 3.4.1 - Phylogenetic tree for gene PTCRA with omega ratios on each node calculated by the free-ratio model in CODEML. Bootstrap values calculated with RAxML and posterior probabilities calculated by Mr. Bayes are indicated on each branch respectively. Five branches that are under positive selection are indicated in red.

Table 3.4.1 - Parameter estimates under model of various omegas ratios among lineages and respective LRTs. LRT - Likelihood ratio test, FDR - False discovery rate correction, ωb background omega. ωf foreground omega.

Models	ω_b	ω_f	ℓ	LRT	*p*-value	FDR
H0: $\omega_b = \omega_{Gibbon} = \omega_{Gorilla} = \omega_{Human\_p} = \omega_{Macaque} = \omega_{Macaque\_p}$	0.56000		-2790.5200			
H1: $\omega_{Gorilla} = \omega_{Human\_p} = \omega_{Macaque\_p} = \omega_{Macaque} = \omega_b \neq \omega_{Gibbon}$	0.52697	1.14497	-2789.0804	2.8792	0.0900	0.4486
H2: $\omega_{Gibbon} = \omega_{Human\_p} = \omega_{Macaque\_p} = \omega_{Macaque} = \omega_b \neq \omega_{Gorilla}$	0.54959	1.33816	-2790.1280	0.7840	0.3760	0.8390
H3: $\omega_{Gibbon} = \omega_{Gorilla} = \omega_{Macaque\_p} = \omega_{Macaque} = \omega b \neq \omega_{Human\_p}$	0.55323	1.21049	-2790.3371	0.3657	0.5453	0.8390
H4: $\omega_{Gibbon} = \omega_{Gorilla} = \omega_{Human\_p} = \omega_{Macaque} = \omega_b \neq \omega_{Macaque\_p}$	0.55564	0.59626	-2790.5116	0.0169	0.89660	0.8966
H5: $\omega_{Gibbon} = \omega_{Gorilla} = \omega_{Human\_p} = \omega_{Macaque\_p} = \omega_b \neq \omega_{Macaque}$	0.5547	0.97395	-2790.4299	0.1801	0.6712	0.8390

Table 3.4.2 - Branch site Model A estimates for the Gibbon, Gorilla, Human_p, Macaque_p_p and Macaque branches. LRT - Likelihood Ratio Test, FDR - False Discovery Rate correction applied to *p*-values.

Model A	$\omega_{background}$	$\omega_{foreground}$	ℓ	LRT	*p*-value	FDR
H_0: Macaque	0.0294	1.0000	H_0-2766.575738 H_1-2766.575738	0.0000	1.00	1.0000000
H_0: Macaque_p_p	0.02505	1.0000	H_0-2766.419449 H_1-2765.963223	0.9125	0.34	0.9334262
H_0: Human_p	0.02877	1.0000	H_0-2766.548784 H_1-2766.496639	0.1043	0.75	0.9334262
H_0: Gorilla	0.02597	1.0000	H_0-2766.503899 H_1-2766.397872	0.2121	0.65	0.9334262
H_0: Gibbon	0.02219	1.0000	H_0-2765.245652 H_1-2764.907976	0.6754	0.41	0.9334262

H_1: Gibbon				H_1:Gorilla			
	Proportion	$\omega_{background}$	$\omega_{foreground}$		Proportion	$\omega_{background}$	$\omega_{foreground}$
Class site 0	0.36859	0.02393	0.02393	Class site 2a	0.43222	0.02544	0.02544
Class site 1	0.39963	1.00000	1.00000	Class site 2b	0.50378	1.00000	1.00000
Class site 2a	0.11121	0.02393	2.29525	Class site 2a	0.02956	0.02544	5.20777
Class site 2b	0.12057	1.00000	2.29525	Class site 2b	0.03445	1.00000	5.20777
H_1: Human_p				H_1: Macaque_p_p			
	Proportion	$\omega_{background}$	$\omega_{foreground}$		Proportion	$\omega_{background}$	$\omega_{foreground}$
Class site 0	0.43302	0.02895	0.02895	Class site 0	0.46001	0.02488	0.024880
Class site 1	0.49912	1.00000	1.00000	Class site 1	0.52843	1.00000	1.000000
Class site 2a	0.03153	0.02895	5.23545	Class site 2a	0.00538	0.02488	27.57668
Class site 2b	0.03634	1.00000	5.23545	Class site 2b	0.00618	1.00000	27.57668
H_1: Macaque							
	Proportion	$\omega_{background}$	$\omega_{foreground}$		Proportion	$\omega_{background}$	$\omega_{foreground}$
Class site 0	0.46324	0.0294	0.0294	Class site 2a	0	0.0294	1
Class site 1	0.53676	1.0000	1.0000	Class site 2b	0	1	1

Table 3.4.3 - Site model analysis, same omega for all branches, PPS – Positively selected sites, Likelihood, Likelihood ratio test, * 0.99 > p.p. >0.95 and ** p.p. > 0.99.

Model	dn/ds	kappa	Parameters						PSS	lnL	lnR	*p*-Value
M0 (one ratio)	0.56	3.78	ω=	0.56						-2790.52		
M1a(neutral)	0.55	4.11	p0=	0.46	p1=	0.54				-2766.58		
			ω0=	0.03	ω1=	1.00						
M2a(selection)	0.68	4.28	p0=	0.68	p1=	0.01	p2=	0.31	27 *0 **0	-2763.49	6.17	0.0457
			ω0=	0.17	ω1=	1.00	ω2=	1.82				
M3(Discrete)	0.68	4.28	p0=	0.12	p1=	0.57	p2=	0.31	70 *15 **11	-2763.49	54.06	5.124E-11
			ω0=	0.17	ω1=	0.17	ω2=	1.81				
M7 (beta)	0.53	4.11	P=	0.02	q=	0.02				-2766.87		
M8(beta & ω)	0.68	4.28	p1=	0.31	ω=	1.81			52 *1 **0	-2763.49	6.755622	0.0341
			p0=	0.69	P=	20.56	q=	99.00				

Species	3	16	19	23	24	25	29	51	52	55	57	70	88	100	101	113	123	137	141	162	163	166	169	190	191	192	193	194	196	200	201	203	207	218	219
				*						*						*		*							*		*		*	*		*	*	*	*
GORILLA	G	A	T	G	T	P	L	D	V	P	L	A	N	A	S	A	V	R	R	G	G	W	V	R	D	P	A	G	P	A	T	T	Q	H	L
HUMAN	G	A	T	G	T	P	L	D	V	P	L	A	N	A	S	A	M	Q	R	G	G	W	V	C	D	P	A	G	L	A	T	T	R	H	P
Gibbon	R	A	T	G	T	L	L	D	V	P	F	A	N	A	S	A	L	R	S	S	G	W	V	R	D	P	A	G	P	T	A	T	R	H	L
Bushbaby	R	A	R	S	T	P	L	D	V	T	L	A	R	T	A	R	L	Q	S	G	Q	R	A	-	-	C	A	H	P	P	A	I	P	A	V
Chimpanzee	G	A	T	P	V	S	S	E	A	T	-	A	N	A	S	A	V	Q	R	G	G	W	V	C	D	P	A	G	L	A	T	T	R	H	L
Orangutan	G	D	T	G	T	P	L	D	V	P	L	T	N	A	S	A	L	R	R	G	G	W	V	R	D	P	A	G	L	A	T	A	R	H	L
Marmoset	G	A	T	G	T	P	L	D	A	S	L	A	S	A	F	T	L	Q	R	G	R	R	A	R	R	P	M	G	P	T	A	A	R	H	P
Macaque	G	T	T	G	T	P	L	D	V	P	F	A	S	A	S	A	L	W	R	G	G	W	V	R	H	P	E	G	L	A	A	A	R	Y	L
Olive Baboon	G	T	T	G	T	P	L	D	V	P	F	A	S	A	S	A	L	W	R	G	G	W	V	R	H	P	A	G	L	A	A	A	R	Y	L

Species	221	223	225	226	227	231	237	239	241	242	244	247	250	251	261	263	264	265	269	270	271	272	273	278	279	280	282	283	284	285	286	290	293	303	304
		*	*	*				*	*		*											*		*	*	*	*				*		*	*	
Gorilla	T	T	G	R	E	S	Q	R	H	R	G	P	R	K	S	L	S	S	C	P	A	R	A	S	A	L	T	P	S	S	S	F	G	A	A
Human	T	T	G	R	E	S	Q	R	R	R	G	P	R	K	S	L	S	S	C	P	A	Q	A	S	A	L	A	P	S	S	S	F	G	A	A
Gibbon	T	N	R	R	E	S	Q	W	R	R	G	P	R	K	S	L	S	S	Y	P	A	R	G	S	A	L	T	P	S	S	S	F	G	A	A
Bushbaby	E	Q	A	A	H	W	H	R	S	C	N	R	R	Q	T	-	-	-	R	R	S	W	G	S	I	S	S	S	T	E	P	V	F	Q	S
Chimpanzee	T	T	G	R	E	S	Q	R	C	R	G	P	R	K	S	L	S	S	C	P	A	R	A	S	A	L	T	P	S	S	S	F	G	A	A
Orangutan	T	T	G	R	E	S	Q	W	R	R	S	P	R	K	S	L	S	S	C	P	A	R	A	S	A	R	T	P	S	S	S	F	G	A	A
Marmoset	T	T	G	R	E	S	L	W	Y	R	G	P	W	E	S	P	T	S	C	P	A	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Macaque	T	T	G	R	E	S	Q	R	H	G	S	P	W	K	S	L	S	R	C	P	A	Q	A	S	D	L	I	P	S	S	S	F	C	A	A
Olive Baboon	T	T	G	R	E	S	Q	R	H	G	S	P	R	K	S	L	S	R	C	P	A	Q	A	S	D	L	I	P	S	S	S	F	C	A	A

Figure 3.4.2 - Alignment of Amino acid residues with positive selection M3 NEB analysis. * 0.95 < p.p. <0.99 - ** p.p. >0.99

3.4.1 Functional analysis

The tertiary structure was calculated by modeling the protein sequence to the 3of6.2.C[80] template of PTCRA calculated by x-ray crystallography. The tertiary structure Figure 3.4.3 shows that the selected residues are on the protein surface region.

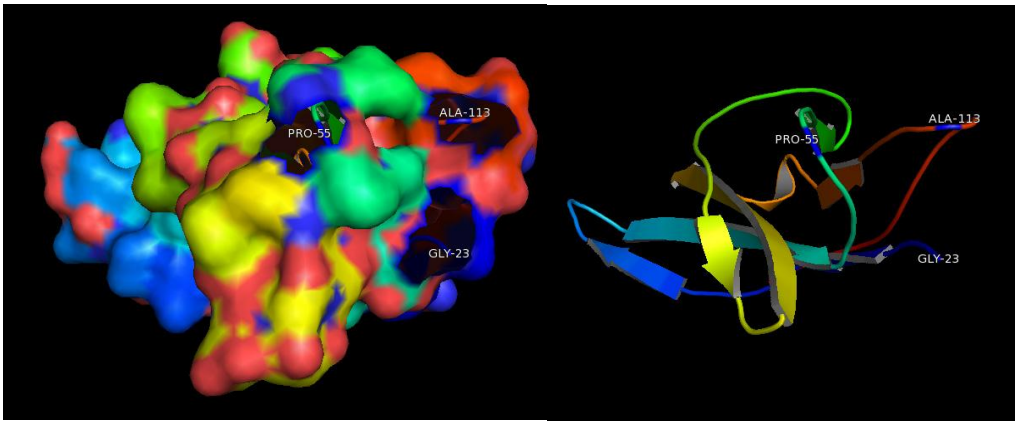


Figure 3.4.3 - PTCRA tertiary structure of the amino acid residues from 23-126. On the left with surface area represented on the right without surface area. The a.a. residues identified by M3 to be under positive selection with posterior probabilities bigger than 0.95 are identified with the three letter protein symbol and respective position.

Analysis of the quaternary structure[80] (3of6.2.C) shows a complex of 2 biomacromolecules, T cell receptor beta chain (A,B,C chains) and Pre T-cell antigen receptor alpha (D,E,F chains). The positions of the positively selected amino acids were shifted 16 positions, in the primary structure reference, of the 3D structures, however maintaining the same relative distance to each other.

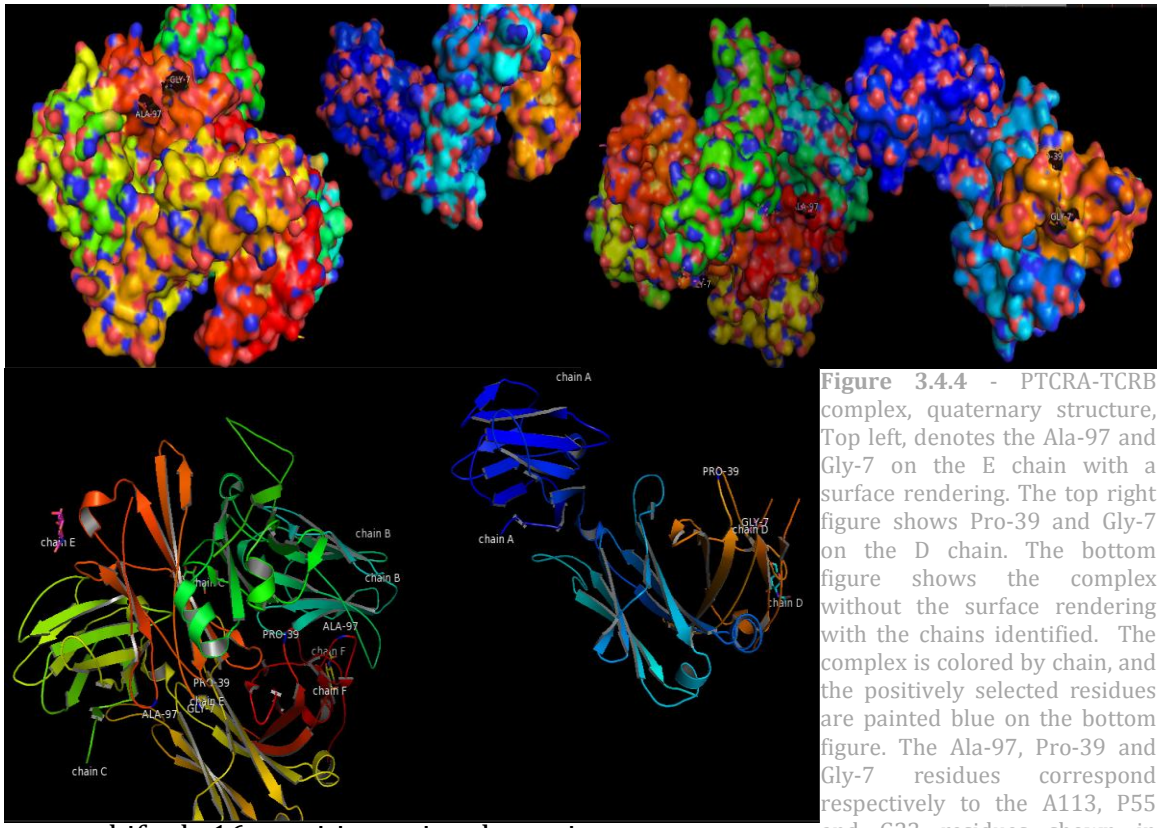


Figure 3.4.4 - PTCRA-TCRB complex, quaternary structure, Top left, denotes the Ala-97 and Gly-7 on the E chain with a surface rendering. The top right figure shows Pro-39 and Gly-7 on the D chain. The bottom figure shows the complex without the surface rendering with the chains identified. The complex is colored by chain, and the positively selected residues are painted blue on the bottom figure. The Ala-97, Pro-39 and Gly-7 residues correspond respectively to the A113, P55 and G23 residues shown in **Figure 3.4.3** and **Figure 3.4.2**

3.5 HOXA3

(HGNC Symbol)

Analysis of the results, from the free ratio model Figure 3.5.1 suggest that the, Human branch is evolving due to positive selection. In order to exclude the over calculation, the Human branch was tested individually and compared with the one-ratio model. This test Table 3.5.1 rejected the one-ratio model null hypotheses with an omega ratio of 1.01062 Table 3.5.2. Also, the branch site model A test failed to reject the null model.

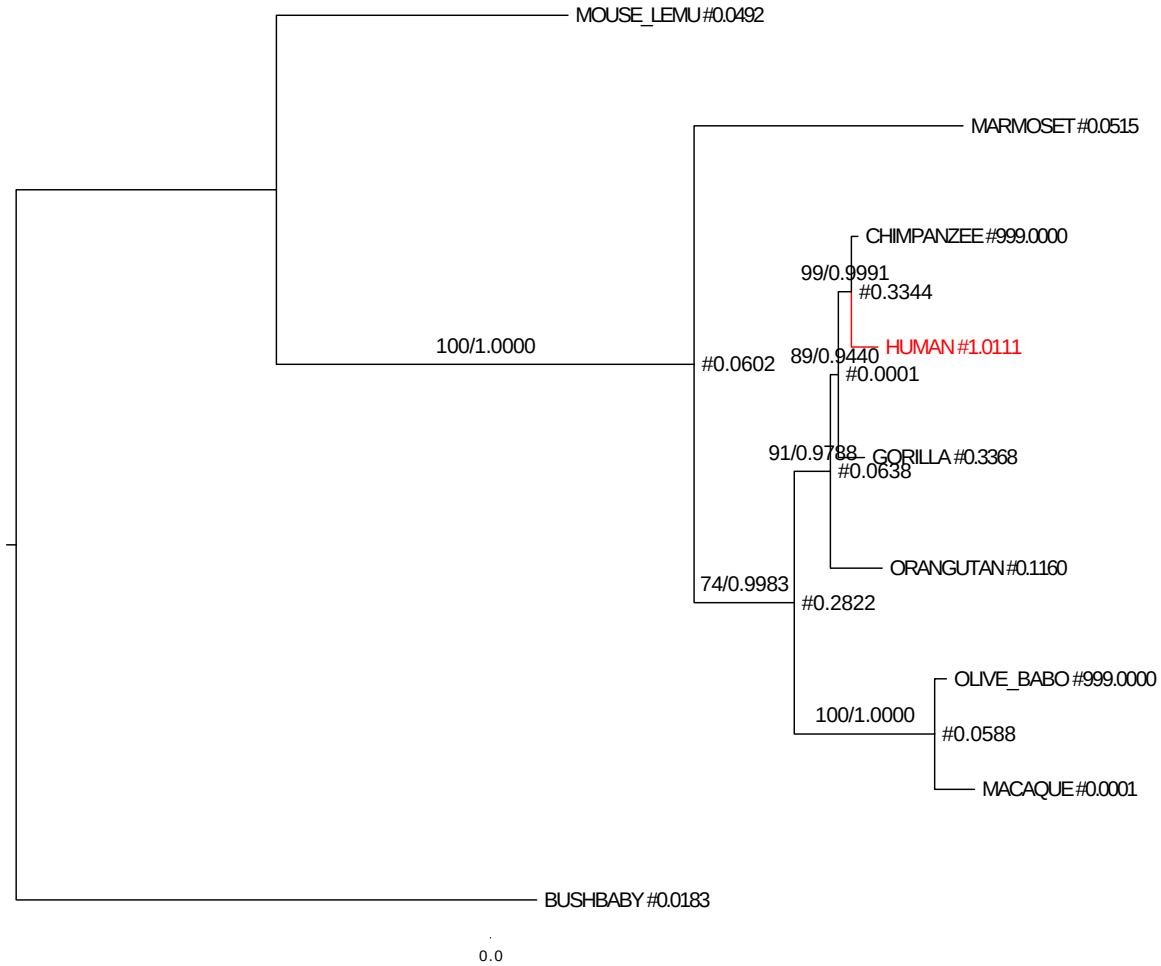


Figure 3.5.1 - Phylogenetic tree for gene HOXA3, with omega ratios on each node calculated by the free-ratio model in codeml. Bootstrap values calculated with RAxML and posterior probabilities calculated by Mr. Bayes are indicated on each branch respectively. One branch that is under positive selection is indicated in red.

Table 3.5.1 - Parameter estimates under model of various omegas ratios among lineages and respective LRTs. LRT - Likelihood ratio test, FDR - False discovery rate correction, ω_b background omega. ω_f foreground omega.

Models	ω_b	ω_f	ℓ	LRT	*p*-value	FDR
H_0: $\omega_b = \omega_{Human}$			-2880.4300			
H_1: $\omega_b \neq \omega_{Human}$	0.06024	1.01062	-2876.9505	6.9590	0.0083	

Table 3.5.2 - Branch site Model A estimates for Human branch. LRT - Likelihood Ratio Test, FDR - False Discovery Rate correction applied to *p*-values.

Model A		ω background	ω foreground	ℓ	LRT	*p*-value	FDR
H_0: Human		0.04568	1.00000	H_0-2873.042 H_1-2873.007	0.0704	0.79	-
H_1:Human							
	Proportion	ω background	ω foreground		Proportion	ω background	ω foreground
Class site 0	0.76831	0.04543	0.04543	Class site 2a	0.21238	0.04543	4.20396
Class site 1	0.01513	1.00000	1.00000	Class site 2b	0.00418	1.00000	4.20396

The site models M2a and M8 also fail the likelihood ratio test, while M3 detects approximately one site with an average omega ratio of 13.42. One amino acid residue with a posterior probability between 0.99 and 0.95 was detected and Figure 3.5.2 shows a representation of the aligned amino acid residues for the positively selected site. In order to visualize the distribution of the three classes identified by M3 a stacked histogram was plotted Figure 6.0.2.

Table 3.5.3 Site model analysis, same omega for all branches, PPS – Positively selected sites, LRT - Likelihood ratio test, * 0.99 > p.p. >0.95 and ** p.p. > 0.99.

Model	dn/ds	kappa	Parameters						PSS	lnL	lnR	*p*-Value
M0 (one ratio)	0.06	4.74	ω=	0.06						-2880.43		
M1a(neutral)	0.07	4.80	p0=	0.98	p1=	0.02				-2875.94		
			ω0=	0.05	ω1=	1.00						
M2a(selection)	0.09	4.85	p0=	0.98	p1=	0.02	p2=	0.00	1 *0 **0	-2874.78	2.31862	0.3137
			ω0=	0.05	ω1=	1.00	ω2=	13.42				
M3(Discrete)	0.09	4.83	p0=	0.62	p1=	0.38	p2=	0.00	1 *1 **0	-2873.81	13.2410	0.0102
			ω0=	0.00	ω1=	0.16	ω2=	13.42				
M7 (beta)	0.07	4.78	P=	0.20	q=	2.58				-2876.36		
M8(beta & ω)	0.09	4.83	p1=	0.00	ω=	13.41			4 *0 **0	-2873.90	4.92035	0.0854
			p0=	1.00	P=	0.38	q=	5.44				

species	337
	*
GORILLA/1-444	T
HUMAN/1-444	A
MOUSE_LEMU/1-388	A
BUSHBABY/1-444	S
CHIMPANZEE/1-444	A
ORANGUTAN/1-444	A
MARMOSET/1-444	S
MACAQUE/1-444	A
OLIVE_BABO/1-444	A

Figure 3.5.2 Alignment of Amino acid residues with positive selection according to sites model M3 NEB analysis. * 0.95 < p.p. <0.99 - ** p.p. >0.99.

3.6 FOXN1

(HGNC Symbol)

Application of the free-ratio model to the FOXN1 gene revealed one branch with positive selection the Chimpanzee branch (1.2650). Due to the nature of this parameter rich model to rule out an overestimation the values were tested against the one ratio model M0 from site model Table 3.6.1.
Study of the individual branch reveals positive selection on the chimpanzee branch with a foreground omega of 1.277 at a 0.001 significance.
Branch site model A Table 3.6.2 however fails to confirm positive selection being unable to reject the null hypotheses.
Also, both the site models LRTs, for M8 and M2 both fail at a 5% significance, and only M3 passes the LRT. Although M3 does not detect positive selection on any site.

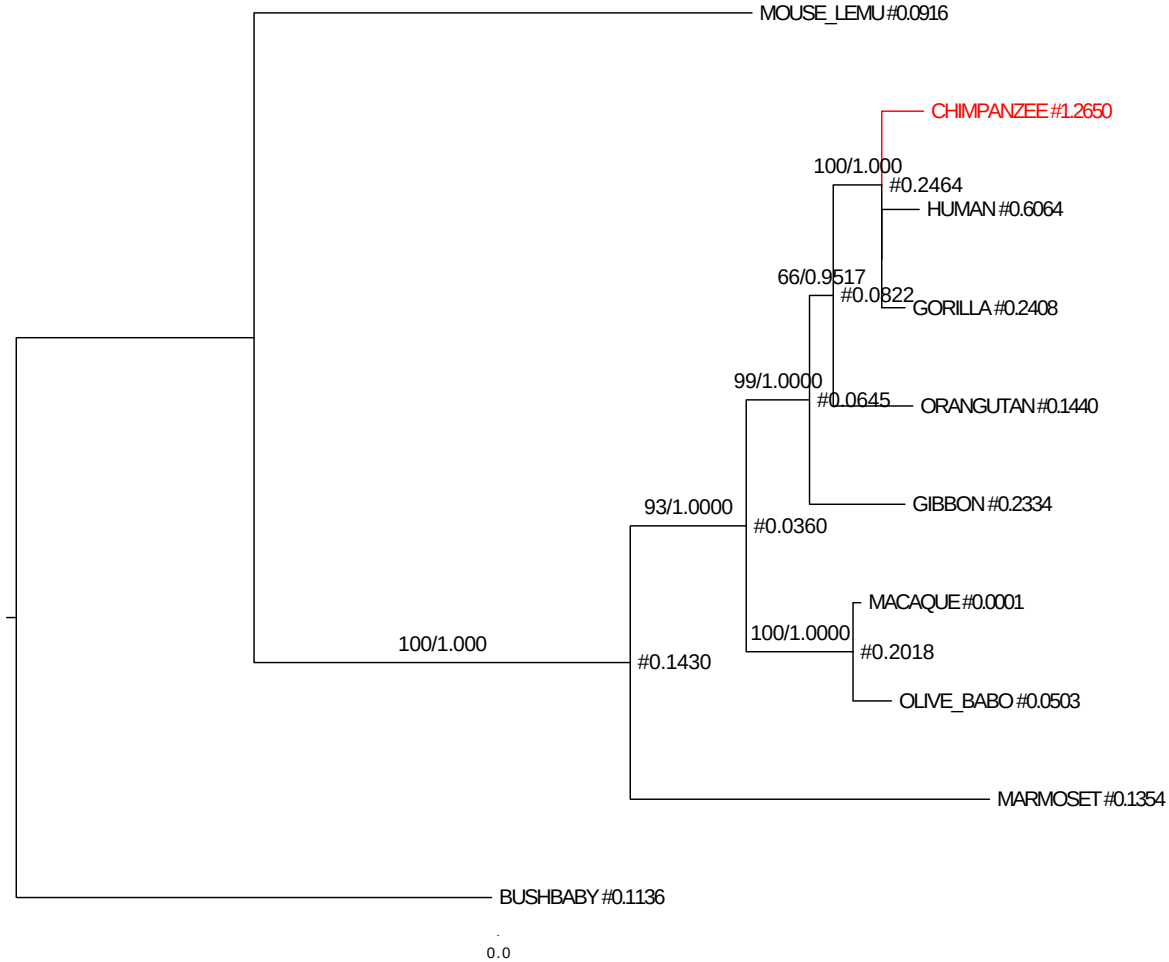


Figure 3.6.1 - Phylogenetic tree for gene FOXN1 with omega ratios on each node calculated by the free-ratio model in codeml. Bootstrap values calculated with RAxML and posterior probabilities calculated by Mr. Bayes are indicated on each branch respectively. One branch that is under positive selection is indicated in red.

Table 3.6.1 - Parameter estimates under model of various omegas ratios among lineages and respective LRTs. LRT - Likelihood ratio test, FDR - False discovery rate correction, ω_b background omega. ω_f foreground omega

Models	ω_b	ω_f	ℓ	LRT	*p*-value	FDR
H0: ω_b :q= ω_{chimp}			-4800.5500			
H1: $\omega_b \neq \omega_{chimp}$	0.12298	1.27749	-4795.2759	10.548138	0.001163052	-

Table 3.6.2 - Branch site Model A estimates for Chimpanzee branch, Gorilla ancestor and Human ancestor. LRT - Likelihood Ratio Test, FDR - False Discovery Rate correction applied to *p*-values.

Model A		$\omega_{background}$	$\omega_{foreground}$	ℓ	LRT	*p*-value	FDR
H_0: Chimpanzee			1	H_0-4783.159424 H_1-4782.815515	0.6878	0.41	-
H_1: Chimpanzee							
	Proportion	$\omega_{background}$	$\omega_{foreground}$		Proportion	$\omega_{background}$	$\omega_{foreground}$
Class site 0	0.76332	0.06898	0.06898	Class site 2a	0.15908	0.06898	5.46564
Class site 1	0.06422	1	1	Class site 2b	0.01338	1	5.46564

Table 3.6.3 - Site model analysis, same omega for all branches, PPS – Positively selected sites, LRT - Likelihood ratio test, * 0.99 > p.p. >0.95 and ** p.p. > 0.99.

Model	dn/ds	kappa	Parameters			PSS	lnL	lnR	*p*-Value
M0 (one ratio)	0.13	5.08	ω= 0.13				-4800.55		
M1a(neutral)	0.15	5.18	p0= 0.92 ω0= 0.07	p1= 0.08 ω1= 1.00			-4787.24		
M2a(selection)	0.15	5.18	p0= 0.92 ω0= 0.07	p1= 0.04 ω1= 1.00	p2= 0.04 ω2= 1.00	12 *0 **0	-4787.24	0	1
M3(Discrete)	0.15	5.18	p0= 0.49 ω0= 0.07	p1= 0.43 ω1= 0.07	p2= 0.09 ω2= 0.93	0 *0 **0	-4787.23	26.629	2.362E-05
M7 (beta)	0.14	5.16	P= 0.24	q= 1.44			-4787.79		
M8(beta & ω)	0.15	5.18	p1= 0.08 p0= 0.92	ω= 1.00 P= 8.01	q= 99.00	17 *0 **0	-4787.25	1.093	0.579

3.7 GCM2

(HGNC Symbol)

The free ratio model Figure 3.7.1 was employed to search for branches that might have evolved through positive selection in the GCM2 gene. Three branches with positive selection were detected, the Chimpanzee, the Gorilla and the Marmoset specific branches, with respective omegas of 2.3572, 1.3831, and 1.0259. The branches were tested individually against the one ratio model, to refute overexpression of the free ratio model.

The Table 3.7.1 Shows the results of the individual test preformed on each branch, where only Marmoset (H3) is able to reject the one-ratio model.

Similarly the branch site tests performed with model A Table 3.7.2, all fail to provide reliable estimates of positive selection.

Nevertheless, site models M2a and M8 Table 3.7.3 also fail to reject their respective null models. While model 3 rejects its null model the omega estimates of all three classes are below 1 indicating only purifying selection.

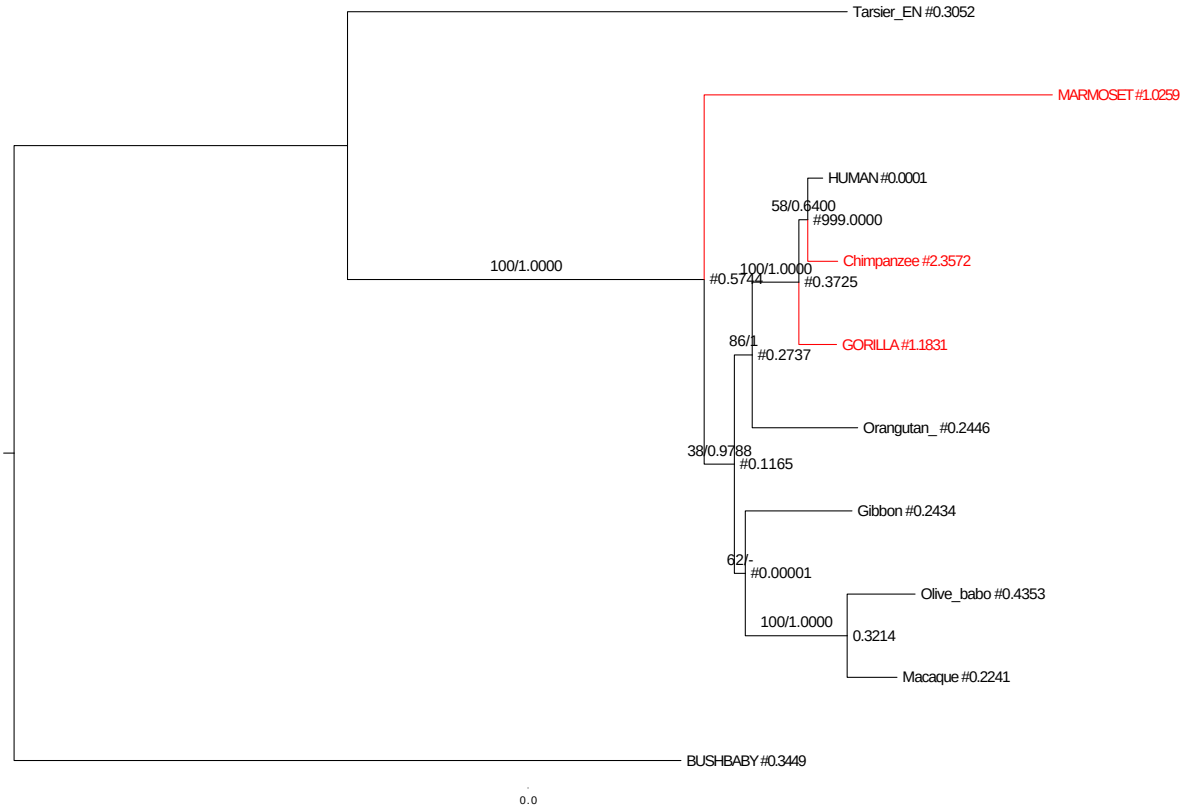


Figure 3.7.1 - Phylogenetic tree for gene GCM2 with omega ratios on each node calculated by the free-ratio model in codeml. Bootsrtrap values calculated with RAxML and posterior probabilities calculated by Mr. Bayes are indicated on each branch respectively. Three branches that are under positive selection are indicated in red.

Table 3.7.1 - Parameter estimates under model of various omegas ratios among lineages and respective LRTs. LRT - Likelihood ratio test, FDR - False discovery rate correction, ω_b background omega. ω_f foreground omega.

Models	ω_b	ω_f	ℓ	LRT	*p*-value	FDR
H0: $\omega_{b} = \omega_{Chimp} = \omega_{Gorilla} = \omega_{Marmoset}$		0.4	-4438.5100			
H1: $\omega_{Marmoset} = \omega_{Gorilla} = \omega_{b} \neq \omega_{Chimp}$	0.38644	2.2989	-4436.6307	3.758564	0.05253768	0.07880652
H2: $\omega_{Chimp} = \omega_{Marmoset} = \omega_{b} \neq \omega_{Gorilla}$	0.38707	1.31533	-4437.2686	2.482806	0.115096749	0.1150967
H3: $\omega_{Chimp} = \omega_{Gorilla} = \omega_{b} \neq \omega_{Marmoset}$	0.35144	0.9286	-4433.5944	9.831192	0.001715771	0.005147313

Table 3.7.2 - Branch site Model A estimates for Chimpanzee, Gorilla and Marmoset branches. LRT - Likelihood Ratio Test, FDR - False Discovery Rate correction applied to *p*-values.

Model A		$\omega_{background}$	$\omega_{foreground}$	ℓ	LRT	*p*-value	FDR
H_0: Chimpanzee		0.18072	1	H_0-4428.056429 H_1-4427.538361	1.0361	0.31	0.465
H_0: Gorilla		0.17167	1	H_0-4428.237028 H_1-4428.128185	0.2177	0.64	0.64
H_0: Marmoset		0.16626	1	H_0-4424.334711 H_1-4422.774077	3.1213	0.08	0.24
H_1: Chimpanzee				H_1:Gorilla			
	Proportion	$\omega_{background}$	$\omega_{foreground}$		Proportion	$\omega_{background}$	$\omega_{foreground}$
Class site 0	0.58803	0.18296	0.18296	Class site 0	0.68649	0.17039	0.17039
Class site 1	0.23695	1	1	Class site 1	0.30711	1	1
Class site 2a	0.12475	0.18296	6.84649	Class site 2a	0.00443	0.17039	40.90796
Class site 2b	0.05027	1	6.84649	Class site 2b	0.00198	1	40.90796
H_1: Marmoset				H_1: Marmoset			
	Proportion	$\omega_{background}$	$\omega_{foreground}$		Proportion	$\omega_{background}$	$\omega_{foreground}$
Class site 0	0.63644	0.16437	0.16437	Class site 2a	0.10085	0.16437	4.91219
Class site 1	0.22677	1	1	Class site 2b	0.03594	1	4.91219

Table 3.7.3 Site model analysis, same omega for all branches, PPS – Positively selected sites, Likelihood, Likelihood ratio test, * 0.99 > p.p. >0.95 and ** p.p. > 0.99.

Model	dn/ds	kappa	Parameters						PSS	lnL	lnR	*p*-Value
M0 (one ratio)	0.40	3.73	ω=	0.40						-4438.51		
M1a(neutral)	0.43	3.80	p0=	0.69	p1=	0.31				-4428.24		
			ω0=	0.17	ω1=	1.00						
M2a(selection)	0.43	3.80	p0=	0.69	p1=	0.18	p2=	0.13	1 *0 **0	-4428.24	0	1
			ω0=	0.17	ω1=	1.00	ω2=	1.00				
M3(Discrete)	0.42	3.78	p0=	0.59	p1=	0.23	p2=	0.18	0 *0 **0	-4428.12	20.7812	0.0003
			ω0=	0.13	ω1=	0.85	ω2=	0.85				
M7 (beta)	0.42	3.78	P=	0.43	q=	0.60				-4428.19		
M8(beta & ω)	0.42	3.78	p1=	0.19	ω=	1.00			1 *0 **0	-4428.18	0.02973	0.9852
			p0=	0.81	P=	0.67	q=	1.70				

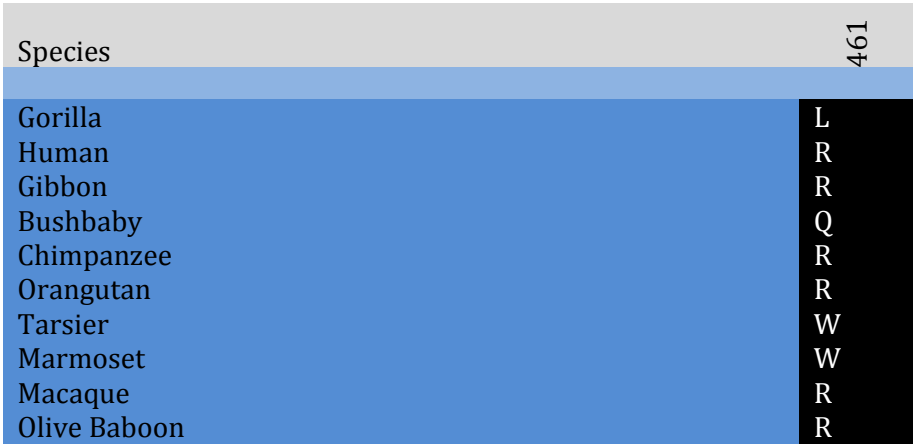


Figure 3.7.2 - Alignment of Amino acid residues with positive selection according to sites model M8 BEB analysis. * 0.95 < p.p. <0.99 - ** p.p. >0.99.

3.8 RUNX1T1

(HGNC Symbol)

The free ratio model Figure 3.8.1 estimated one branch with positive selection, the Human ancestral branch, with a ω ratio of 7.7370.
Testing the individual branch Table 3.8.1 against the one ratio model did not reject the null model.
Furthermore, the branch sites model A, did not detect positive selection on the Human ancestral branch, which also failed the likelihood ratio test Table 3.82.

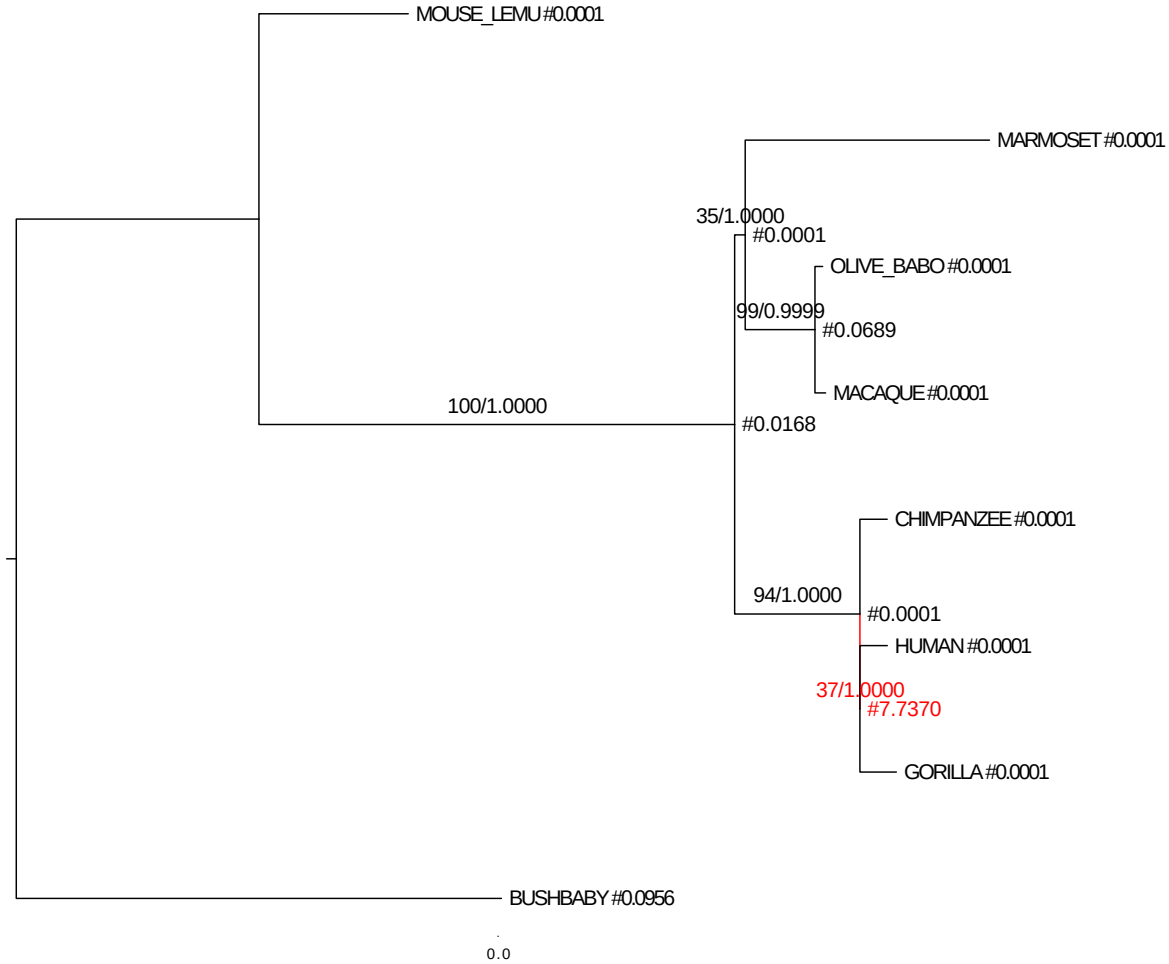


Figure 3.8.1 - Phylogenetic tree for gene RUNX1T1 with omega ratios on each node calculated by the free-ratio model in codeml. Bootstrap values calculated with RAxML and posterior probabilities calculated by Mr. Bayes are indicated on each branch respectively. One branch that is under positive selection is indicated in red.

Table 3.8.1 - Parameter estimates under model of various omegas ratios among lineages and respective LRTs.

Models	ω_b	ω_f		LRT	*p*-value	FDR
H0: $\omega_b = \omega_{human\_p}$	0.03		-3281.5600			
H1: $\omega_b \neq \omega_{human\_p}$	0.02956	3.00126	-3281.5646	-0.009278	-	-

Table 3.8.2 - Branch site Model A estimates for the Human_p branch. LRT - Likelihood Ratio Test, FDR - False Discovery Rate correction applied to *p*-values.

Model A		ω background	ω foreground		LRT	*p*-value	FDR
H0: Human_p		0.00591	1.00000	H0- 3276.639183 H1- 3276.639196	0.0000	-	-
H1: Human_p				H1:Human_p			
	Proportion	ω background	ω foreground		Proportion	ω background	ω foreground
Class site 0	0.89411	0.00591	0.00591	Class site 2a	0.07926	0.00591	1.90653
Class site 1	0.02446	1.00000	1.00000	Class site 2b	0.00217	1.00000	1.90653

Table 3.8.3 - Site model analysis, same omega for all branches, PPS – Positively selected sites, LRT- Likelihood ratio test, * 0.99 > p.p. >0.95 and ** p.p. > 0.99.

Model	dn/ds	kappa	Parameters						PSS	lnL	lnR	*p*-Value
M0 (one ratio)	0.03	3.39	ω=	0.03						-3281.56		
M1a(neutral)	0.03	3.39	p0=	0.97	p1=	0.03				-3276.64		
			ω0=	0.01	ω1=	1.00						
M2a(selection)	0.03	3.39	p0=	0.97	p1=	0.01	p2=	0.01	2 *0 **0	-3276.64	0	1
			ω0=	0.01	ω1=	1.00	ω2=	1.00				
M3(Discrete)	0.03	3.39	p0=	0.39	p1=	0.58	p2=	0.03	0 *0 **0	-3276.64	9.8514	0.0430
			ω0=	0.00	ω1=	0.00	ω2=	0.91				
M7 (beta)	0.03	3.39	P=	0.01	q=	0.26				-3277.37		
M8(beta & ω)	0.03	3.39	p1=	0.03	ω=	1.00			2 *0 **0	-3276.64	1.4619	0.4814
			p0=	0.97	P=	0.03	q=	2.11				

3.9 CD4

(HGNC Symbol)

Application of the free-ratio model on the CD4 gene Figure 3.9.1 revealed two branches with positive selection, on the *Hominiae* branch, specifically on the Chimpanzee and Gorilla specific branches, with omega values of 1.017 and 2.78, respectively. Since the free-ratio model tends to overestimate values, these values were confirmed by testing each branch individually, and comparing to the one ratio model (M0).

Study of the individual branches reveals that the chimpanzee [H1] branch at a 5% significance, did not confirm the presence of positive selection, while the gorilla [H2] branch had a *p*-value near 0.05, which after correction for false discovery rate did not reject the null hypothesis. In order to confirm the previous results, both branches were tested simultaneously [H3] and compared to M0, which also fails to reject the null hypothesis. A random sample of other branches, which indicated omega values lower then one, were also tested, with the various omega ratios used in Table 3.9.1, which all confirmed omega values lower then one.

Furthermore, branch site models A was employed to confirm the distribution of amino acid residues on the Chimpanzee and Gorilla lineages.

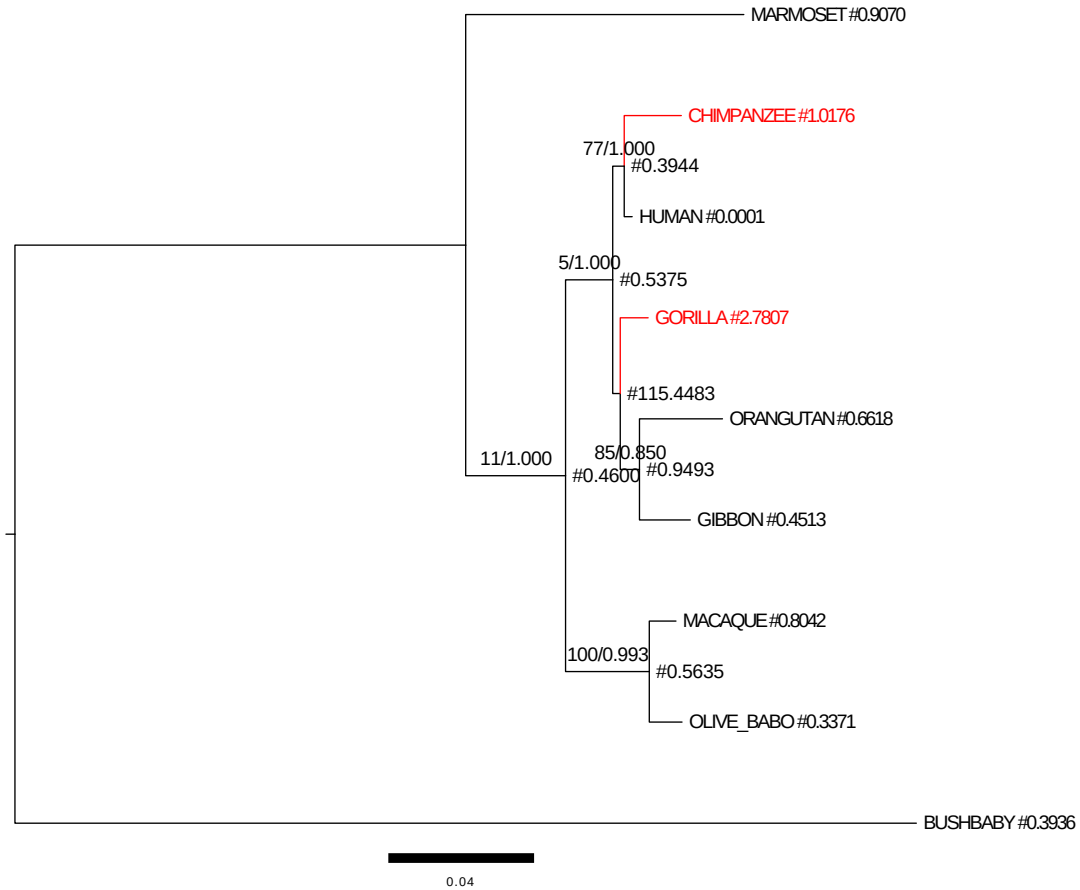


Figure 3.9.1 - Phylogenetic tree for gene CD4 with omega ratios on each node calculated by the free-ratio model in codeml. Bootstrap values calculated with RAxML and posterior probabilities calculated by Mr. bayes are indicated on each branch respectively.

The Table 3.8.2 shows that the likelihood ratio tests for model A, fails at a 0.05 significance, when testing gorilla specific branch, while the chimpanzee specific branch rejects the null hypothesis, suggesting that the branch has residues that are

under positive selection. Figure 3.9.2 shows the alignment of positively selected amino acids, specific to the chimpanzee branch.
Complementary to branch site mode, site models Table 3.9.3 were also calculated to estimate positive selection along all lineages. The selection model M2a identified 30% of amino acids with an average ω of 1.81. The log likelihood ratio test between M1 and M2 was 8.72, which, besides that, gave a *p*-value of 0.012,when compared with the X^2 distribution.
The discrete model M3 shows 30% of amino acids with a positive selection average of 1.81, and the log likelihood ratio test between M0 and M3 of 74.7 with a *p*-value < 0.001.
Concordantly, the M8 model also suggests that 30% of the amino acids were under positive selection, with an average ω of 1.82. The log likelihood ratio test between M7 and M8 was of 11.8 with a *p*-value of 0.0027.
The site-specific likelihood models used to detect positive sites, yielded 57 positively selected sites with posterior probabilities between 0.5 and 0.095 for M2.
The model M3 detected, 107 positively selected sites Figure 3.9.4 with posterior probabilities between 0.5 and 0.95, 22 between 0.95 and 0.99 and 12 amino acids between 0.99 and 1. Model M8 detected 79 amino acids with posterior probabilities between 0.5 and 0.95. The staked histogram Figure 3.9.3 also depicts the distribution of the three site classes calculated in M3.

Table 3.9.1 - Parameter estimates under model of various omegas ratios among lineages and respective LRTs. FDR - False discovery rate correction of *p*-values. H_0vH_1=1d.f.,H_0vH_2=1d.f.,H_0vH_3=2d.f.

Models	ω_b	ω_{chimp}	$\omega_{gorilla}$	ℓ	LRT	*p*-value	FDR
H0: $\omega_b = \omega_{chimp} = \omega_{gorilla}$	0.54	0.54	0.54	-3977.15296			
H1: $\omega_{gorilla} = \omega_b \neq \omega_{chimp}$	0.52019	1.01633	0.52019	-3976.28311	1.739702	0.1872	0.1872
H2: $\omega_{gorilla} \neq \omega_b = \omega_{chimp}$	0.51926	0.51926	2.88888	-3975.15560	3.994714	0.0456	0.0913
H3: $\omega_{gorilla} \neq \omega_b \neq \omega_{chimp}$	0.50198	1.01635	2.89584	-3974.18790	5.930114	0.0516	-

Table 3.9.2 - Branch site Models A for Chimpanzee and Gorilla lineages. LRT – Likelihood ratio test, FDR - False discovery rate correction

Model A		$\omega_{background}$	$\omega_{foreground}$	ℓ	LRT	*p*-value	FDR
H_0: Chimpanzee		0.04381	1.00000	H_0-3937.2040 H_1-3927.2161	19.9756	7.8435E-06	1.5687e-05
H_0: Gorilla		0.07092	1.00000	H_0-3944.1301 H_1-3944.1031	0.0540	0.8162	0.8162
H_1: Chimpanzee				H_1:Gorilla			
	Proportion	$\omega_{background}$	$\omega_{foreground}$		Proportion	$\omega_{background}$	$\omega_{foreground}$
Class site 0	0.5080	0.0546	0.0546	Class site 0	0.4932	0.0719	0.0719
Class site 1	0.4690	1.0000	1.0000	Class site 1	0.4710	1.0000	1.0000
Class site 2a	0.0120	0.0546	107.6919	Class site 2a	0.0183	0.0719	4.5711
Class site 2b	0.0111	1.0000	107.6919	Class site 2b	0.0175	1.0000	4.5711

Species	59	80	93	112	431	439	442	443	444	445	447	450
						**		*	*			
HUMAN	I	A	P	E	E	L	E	K	K	T	Q	H
BUSHBABY	-	V	P	E	E	L	E	K	K	T	Q	H
CHIMPANZEE	T	V	T	G	Q	K	V	W	P	S	R	R
ORANGUTAN	-	A	P	E	E	L	E	K	K	T	Q	H
GIBBON	I	A	P	E	E	L	E	K	K	T	Q	H
GORILLA	M	A	P	E	E	L	E	K	K	T	Q	H
OLIVE_BABO	I	A	S	E	E	L	E	K	K	T	Q	H
MARMOSET	-	I	P	E	E	L	E	K	K	T	Q	H
MACAQUE	I	A	S	E	E	L	E	K	K	T	Q	H

Figure 3.9.2 – Alignment of Amino acid residues with positive selection on the chimpanzee lineage according to branch site model A BEB analysis. * 0.95 < p.p. <0.99 - ** p.p. >0.99

Table 3.9.3 - Site model analysis, same omega for all branches, PPS – Positively selected sites, LRT - Likelihood ratio test, * 0.99 > p.p. >0.95 and ** p.p. > 0.99.

Model	dn/ds	kappa	Parameters						PSS	lnL	lnR	*p*-value
M0 (one ratio)	0.54	3.48	ω=	0.54						-3977.15		
M1a(neutral)	0.53	3.57	p0=	0.51	p1=	0.49				-3944.13		
			ω0=	0.07	ω1=	1.00						
M2a(selection)	0.68	3.79	p0=	0.70	p1=	0.00	p2=	0.30	57 *0 **0	-3939.77	8.721	0.0128
			ω0=	0.18	ω1=	1.00	ω2=	1.81				
M3(Discrete)	0.68	3.79	p0=	0.70	p1=	0.30	p2=	0.00	107 *22 **12	-3939.77	74.77	2.22E-15
			ω0=	0.18	ω1=	1.81	ω2=	2.44				
M7 (beta)	0.54	3.27	P=	0.11	q=	0.09				-3945.68		
M8(beta & ω)	0.68	3.79	p1=	0.30	ω=	1.82			79 *0 **0	-3939.78	11.80	0.00274
			p0=	0.70	P=	22.75	q=	99.00				

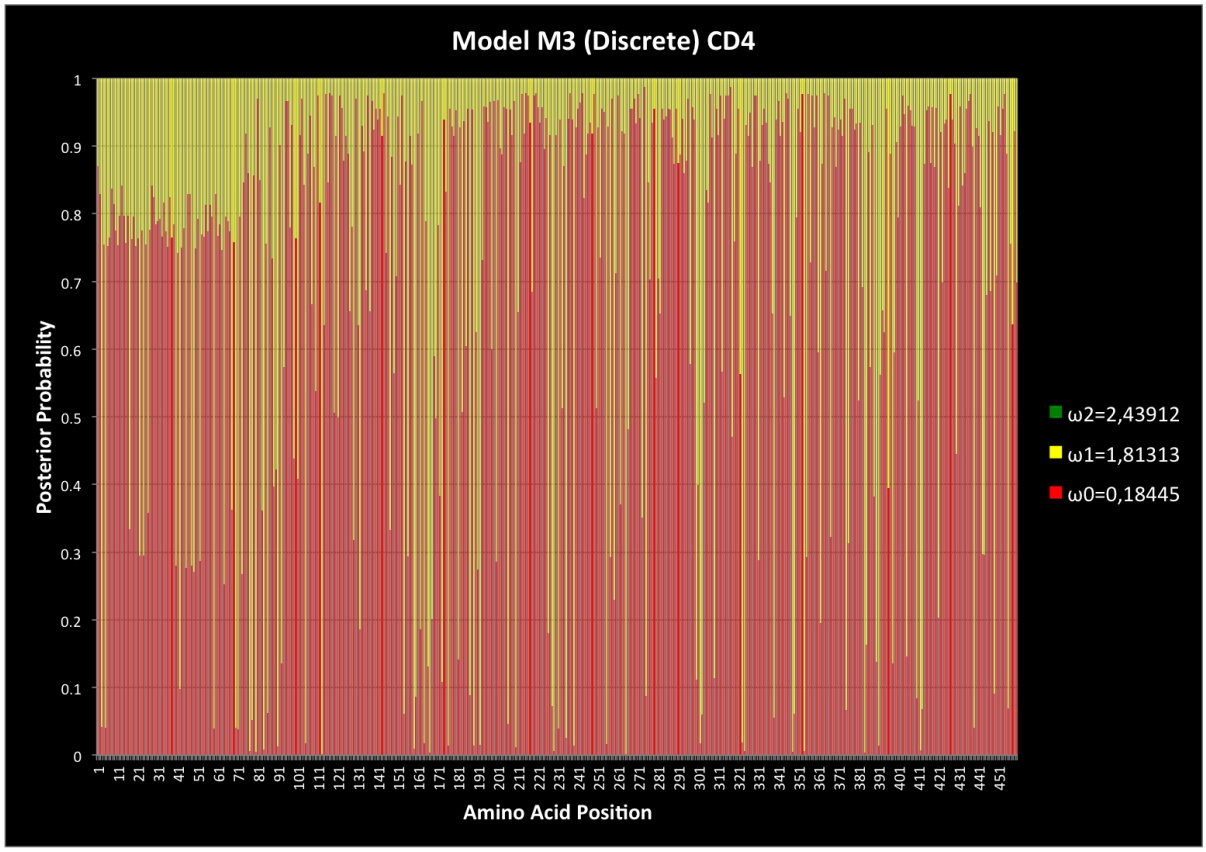
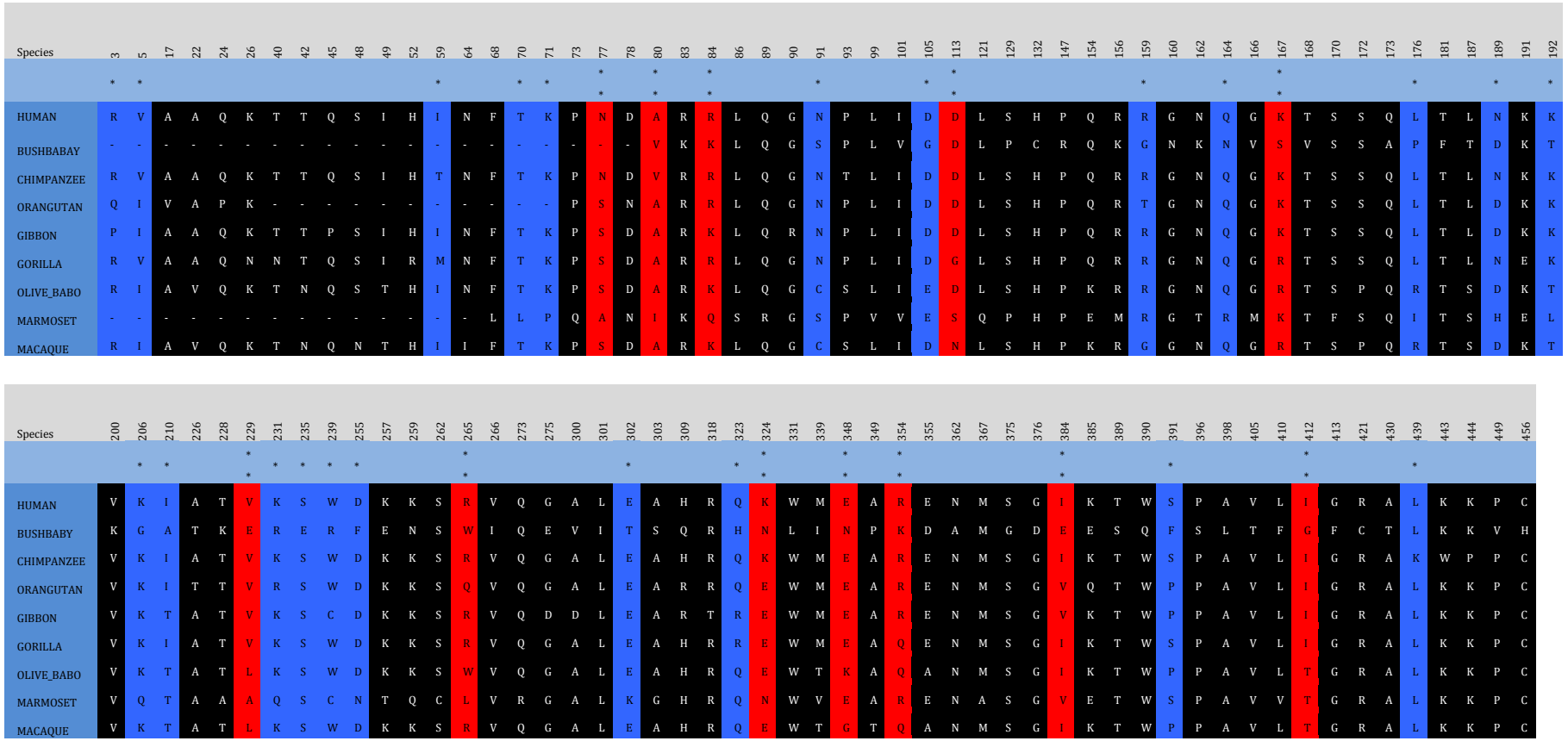


Figure 3.9.3 - Stacked histogram representing the posterior probabilities for the three site classes with different selective pressures identified by the CODEML model M3 (discrete).



Species	3	5	17	22	24	26	40	42	45	48	49	52	59	64	68	70	71	73	77	78	80	83	84	86	89	90	91	93	99	101	105	113	121	129	132	147	154	156	159	160	162	164	166	167	168	170	172	173	176	181	187	189	191	192
	*	*											*			*	*		* *		* *		* *				*				*	* *						*			*	* *					*			*		*
HUMAN	R	V	A	A	Q	K	T	T	Q	S	I	H	I	N	F	T	K	P	N	D	A	R	R	L	Q	G	N	P	L	I	D	D	L	S	H	P	Q	R	R	G	N	Q	G	K	T	S	S	Q	L	T	L	N	K	K
BUSHBABAY	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	V	K	K	L	Q	G	S	P	L	V	G	D	L	P	C	R	Q	K	G	N	K	N	V	S	V	S	S	A	P	F	T	D	K	T
CHIMPANZEE	R	V	A	A	Q	K	T	T	Q	S	I	H	T	N	F	T	K	P	N	D	V	R	R	L	Q	G	N	T	L	I	D	D	L	S	H	P	Q	R	R	G	N	Q	G	K	T	S	S	Q	L	T	L	N	K	K
ORANGUTAN	Q	I	V	A	P	K	-	-	-	-	-	-	-	-	-	-	-	P	S	N	A	R	R	L	Q	G	N	P	L	I	D	D	L	S	H	P	Q	R	T	G	N	Q	G	K	T	S	S	Q	L	T	L	D	K	K
GIBBON	P	I	A	A	Q	K	T	T	P	S	I	H	I	N	F	T	K	P	S	D	A	R	K	L	Q	R	N	P	L	I	D	D	L	S	H	P	Q	R	R	G	N	Q	G	K	T	S	S	Q	L	T	L	D	K	K
GORILLA	R	V	A	A	Q	N	N	T	Q	S	I	R	M	N	F	T	K	P	S	D	A	R	R	L	Q	G	N	P	L	I	D	G	L	S	H	P	Q	R	R	G	N	Q	G	R	T	S	S	Q	L	T	L	N	E	K
OLIVE_BABO	R	I	A	V	Q	K	T	N	Q	S	T	H	I	N	F	T	K	P	S	D	A	R	K	L	Q	G	C	S	L	I	E	D	L	S	H	P	K	R	R	G	N	Q	G	R	T	S	P	Q	R	T	S	D	K	T
MARMOSET	-	-	-	-	-	-	-	-	-	-	-	-	-	-	L	L	P	Q	A	N	I	K	Q	S	R	G	S	P	V	V	E	S	Q	P	H	P	E	M	R	G	T	R	M	K	T	F	S	Q	I	T	S	H	E	L
MACAQUE	R	I	A	V	Q	K	T	N	Q	N	T	H	I	I	F	T	K	P	S	D	A	R	K	L	Q	G	C	S	L	I	D	N	L	S	H	P	K	R	G	G	N	Q	G	R	T	S	P	Q	R	T	S	D	K	T

Species	200	206	210	226	228	229	231	235	239	255	257	259	262	265	266	273	275	300	301	302	303	309	318	323	324	331	339	348	349	354	355	362	367	375	376	384	385	389	390	391	396	398	405	410	412	413	421	430	439	443	444	449	456
		*	*			* *	*	*	*	*				* *							*				*	*			* *	* *							*						* *						*					
HUMAN	V	K	I	A	T	V	K	S	W	D	K	K	S	R	V	Q	G	A	L	E	A	H	R	Q	K	W	M	E	A	R	E	N	M	S	G	I	K	T	W	S	P	A	V	L	I	G	R	A	L	K	K	P	C
BUSHBABY	K	G	A	T	K	E	R	E	R	F	E	N	S	W	I	Q	E	V	I	T	S	Q	R	H	N	L	I	N	P	K	D	A	M	G	D	E	E	S	Q	F	S	L	T	F	G	F	C	T	L	K	K	V	H
CHIMPANZEE	V	K	I	A	T	V	K	S	W	D	K	K	S	R	V	Q	G	A	L	E	A	H	R	Q	K	W	M	E	A	R	E	N	M	S	G	I	K	T	W	S	P	A	V	L	I	G	R	A	K	W	P	P	C
ORANGUTAN	V	K	I	T	T	V	R	S	W	D	K	K	S	Q	V	Q	G	A	L	E	A	R	R	Q	E	W	M	E	A	R	E	N	M	S	G	V	Q	T	W	P	P	A	V	L	I	G	R	A	L	K	K	P	C
GIBBON	V	K	T	A	T	V	K	S	C	D	K	K	S	R	V	Q	D	D	L	E	A	R	T	R	E	W	M	E	A	R	E	N	M	S	G	V	K	T	W	P	P	A	V	L	I	G	R	A	L	K	K	P	C
GORILLA	V	K	I	A	T	V	K	S	W	D	K	K	S	R	V	Q	G	A	L	E	A	H	R	R	E	W	M	E	A	Q	E	N	M	S	G	I	K	T	W	S	P	A	V	L	I	G	R	A	L	K	K	P	C
OLIVE_BABO	V	K	T	A	T	L	K	S	W	D	K	K	S	W	V	Q	G	A	L	E	A	H	R	Q	E	W	T	K	A	Q	A	N	M	S	G	I	K	T	W	P	P	A	V	L	T	G	R	A	L	K	K	P	C
MARMOSET	V	Q	T	A	A	A	Q	S	C	N	T	Q	C	L	V	R	G	A	L	K	G	H	R	Q	N	W	V	E	A	R	E	N	A	S	G	V	E	T	W	S	P	A	V	V	T	G	R	A	L	K	K	P	C
MACAQUE	V	K	T	A	T	L	K	S	W	D	K	K	S	R	V	Q	G	A	L	E	A	H	R	Q	E	W	T	G	T	Q	A	N	M	S	G	I	K	T	W	P	P	A	V	L	T	G	R	A	L	K	K	P	C

Figure 3.9.4 - Amino acid alignment of residues with positive selection, * 0.95 >p.p. >0.99, ** p.p. > 0.99, others 0.50 > p.p. > 0.95

3.9.1 Functional analysis

Sift scores for all the positions and their respective average of each position were calculated to measure the effect of the amino acid change in the protein structure, specifically for the chimpanzee branch, in comparison to human. Two of the three positions showed above as positively selected, with posterior probabilities higher than 0.95, were identified as potentially damaging to the protein structure Table 3.9.4. Besides those, another position was identified as damaging, although that same position had a low posterior probability.

When the selected sites were combined with data of different databases and literature, it was possible to identify several overlapping zones of mutation, which reinforced the possibility of these sites being truly positively selected sites Figures 3.9.5, 3.9.6.

The tertiary structure of the CD4 protein Figure 3.9.7 shows that the selected residues are on the protein surface region.

Finally, the average sift score along the protein showed that the zones with tolerated mutations overlapped with the positively selected sites Figure 3.9.8.

Table 3.9.4 - Measurement of the effect of the amino acid change in protein tridimensional conformation. ENSP – protein identification; Pos – Position of the residue; Ref – residue in the reference; Subst – substitute residue; Prediction – damaging or tolerate if it changes or not the protein; SIFT Score – Varies between 0 and 1. The smaller the number, higher the effect on protein folding. The marked lines correspond to the sites earlier identified with positive selection.

User Input	ENSP	Pos	Ref	Subst	Prediction	SIFT Score
ENSP00000011653,I59TM	ENSP00000011653	59	I	T	TOLERATED	0.63
ENSP00000011653,I59TM	ENSP00000011653	59	I	M	TOLERATED	0.1
ENSP00000011653,A80VI	ENSP00000011653	80	A	V	TOLERATED	1
ENSP00000011653,A80VI	ENSP00000011653	80	A	I	TOLERATED	0.54
ENSP00000011653,P93TS	ENSP00000011653	93	P	T	TOLERATED	0.12
ENSP00000011653,P93TS	ENSP00000011653	93	P	S	TOLERATED	1
ENSP00000011653,E112GE	ENSP00000011653	112	E	G	TOLERATED	0.2
ENSP00000011653,E112GE	ENSP00000011653	112	E	E	TOLERATED	1
ENSP00000011653,E430Q	ENSP00000011653	430	E	Q	TOLERATED	0.44
ENSP00000011653,L438K	ENSP00000011653	438	L	K	TOLERATED	1
ENSP00000011653,E441V	ENSP00000011653	441	E	V	TOLERATED	0.08
ENSP00000011653,K442W	ENSP00000011653	442	K	W	DAMAGING	0.02
ENSP00000011653,K443P	ENSP00000011653	443	K	P	DAMAGING	0
ENSP00000011653,T444S	ENSP00000011653	444	T	S	TOLERATED	0.15
ENSP00000011653,Q446R	**ENSP00000011653**	**446**	**Q**	**R**	**DAMAGING**	**0.03**
ENSP00000011653,H449R	ENSP00000011653	449	H	R	TOLERATED	0.21

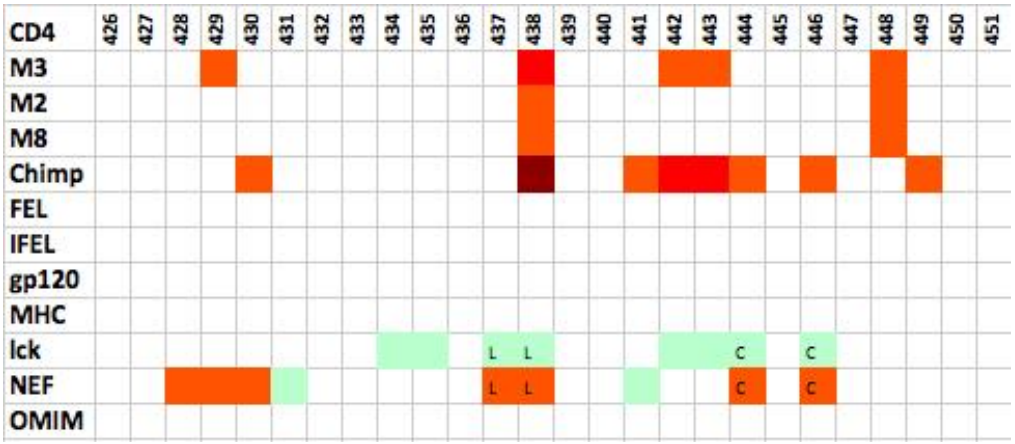


Figure 3.9.5 - CD4 Cytoplasmatic tail region, positively selection sites identified by codeml, m3 m2 and m8 models, Hyphy server Fel and IFEL methods, Chimp corresponds to the chimpanzee specific p.p. from model A. Gp120,MHC,lck,NEF are the binding sites of the respective protein molecules. OMIM is the snps associated with disease identified by the online mendelian inheritance in man database (omim.org).

Figure 3.9.6 - CD4 extracellular domains, positively selection sites identified by codeml, m3 m2 and m8 models, Hyphy server Fel and IFEL methods, Chimp corresponds to the chimpanzee specific p.p. from model A. Gp120,MHC,lck,NEF are the binding sites of the respective protein molecules. OMIM is the snps associated with disease identified by the online mendelian inheritance in man database (omim.org).

.

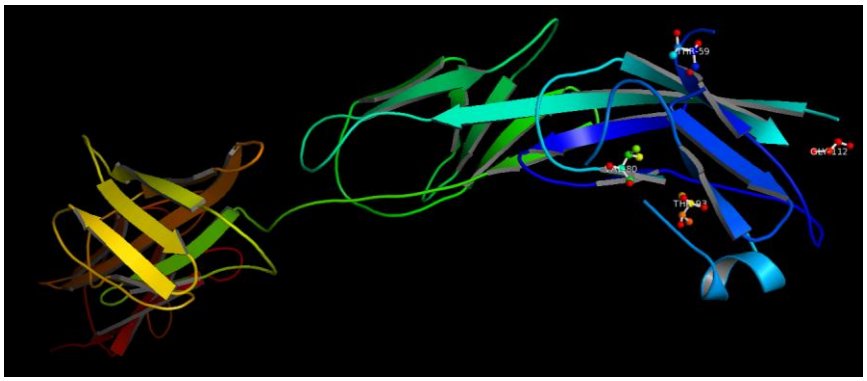


Figure 3.9.7 - 3D Protein confirmation of a.a. 26-386

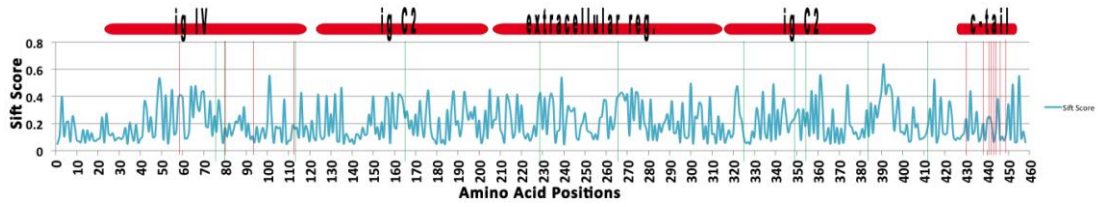


Figure 3.9.8 Average sift score for each possible amino acid mutation throughout CD4, with Pfam domain types positioned in red over graph and highest positively selected residues shown with colored vertical lines. Green lines corresponds highest p.p. of M3 sites model and red lines corresponds to the chimpanzee branch.

3.9.2 Validation of chimpanzee CD4 sequence

To validate the consensus sequence of CD4 of chimpanzee, the Non human primate reference transcriptome (NHPRT) assembly was surveyed with the purpose of identify the reads that originated that same sequence. It was possible to determine only one read with the expected nucleic acid, crucial to the origin of the reference sequence Table 3.9.5. On the other hand, on the merged assembly Table 3.9.6 some reads with the expected nucleic were found.

Table 3.9.5 - Sum up of nucleotides per position for NHPRT assembly.

Chr	Position	Consensus M	Expected	tcov	covA	covC	CovG	covT
12	6982895	C	A	138	1	137	0	0
12	6982896	T	A	138	0	1	0	137
12	6982897	C	A	138	0	136	0	2
12	6982905	A	T	117	115	1	0	1
12	6982906	G	G	117	0	0	117	0
12	6982907	A	T	117	117	0	0	0
12	6982908	A	G	117	117	0	0	0
12	6982909	G	G	110	0	0	110	0
12	6982910	A	C	110	110	0	0	0
12	6982911	A	C	110	110	0	0	0
12	6982912	G	G	110	0	0	110	0
12	6982913	A	T	110	110	0	0	0
12	6982920	A	G	104	104	0	0	0

The Figure 3.9.9 shows the alignment of the RNAseq reads from Henrik Kaessmann aligned to the reference genome from which the sequence in study was obtained plus the relative position in the genome. While Figure 3.9.10 shows, the translation of the nucleotide sequence to amino acids for three species, denoting the exons in the sequences.

Table 3.9.6 - Sum up of nucleic acids per position for the Merged (M), C6_Antoine (A), C5_Koos (K), C1_Herman (H) and C2_Japie (J) assemblies from Henrik Kaessmann (University of Lausanne) RNAseq data. Highlighted lines are positions were consensus differ from database reported red columns are the expected nucleic acid.

					M	A	K	J	H		M	A	K	J	H		M	A	K	J	H		M	A	K	J	H		M	A	K	J	H
Chr	Position	Consensus M	Expected		tcov						covA						covC						covG						covT				
12	6982895	B	A		6	1	2	-	-		0	0	0	-	-		2	0	1	-	-		2	1	0	-	-		2	0	1	-	-
12	6982896	C	A		6	1	2	-	-		0	0	0	-	-		4	1	1	-	-		2	0	1	-	-		0	0	0	-	-
12	6982897	G	A		6	1	2	-	-		0	0	0	-	-		2	0	1	-	-		4	1	1	-	-		0	0	0	-	-
12	6982905	G	T		6	1	2	-	-		2	0	1	-	-		0	0	0	-	-		4	1	1	-	-		0	0	0	-	-
12	6982906	G	G		6	1	2				0	0	0				2	1	0	-	-		4	0	2	-	-		0	0	0	-	-
12	6982907	H	T		6	1	2	-	-		2	0	1	-	-		2	0	1	-	-		0	0	0	-	-		2	1	0	-	-
12	6982908	A	G		6	1	2	-	-		4	0	2	-	-		0	0	0	-	-		2	1	0	-	-		0	0	0	-	-
12	6982909	G	40		6	1	2	-	-		2	0	1	-	-		0	0	0	-	-		4	1	1	-	-		0	0	0	-	-
12	6982910	C	C		6	1	2	-	-		2	0	1	-	-		4	1	1	-	-		0	0	0	-	-		0	0	0	-	-
12	6982911	G	C		6	1	2	-	-		2	0	1	-	-		0	0	0	-	-		4	1	1	-	-		0	0	0	-	-
12	6982912	G	G		6	1	2	-	-		0	0	0				0	0	0	-	-		6	1	2	-	-		0	0	0	-	-
12	6982913	D	T		6	1	2	-	-		2	0	1	-	-		0	0	0	-	-		2	1	0	-	-		2	0	1	-	-
12	6982920	V	G		6	1	2	-	-		2	0	1	-	-		2	0	1	-	-		2	1	0	-	-		0	0	0	-	-

Figure 3.9.9 - RNAseq sequences from Henrik Kaessmann aligned with the CHIMP2.

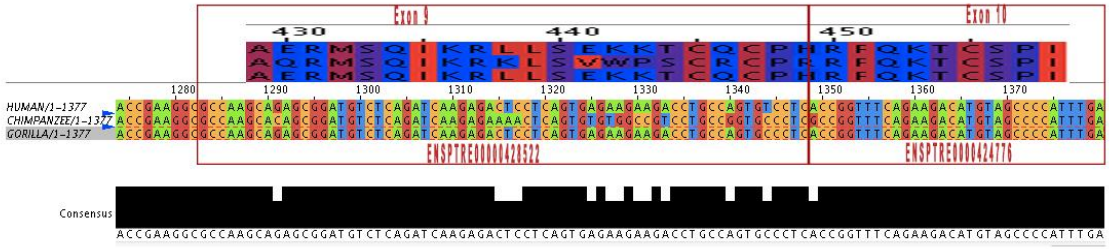


Figure 3.9.10 - Visualization of the cytoplasmic tail with respective translation and annotation of the corresponding exons.

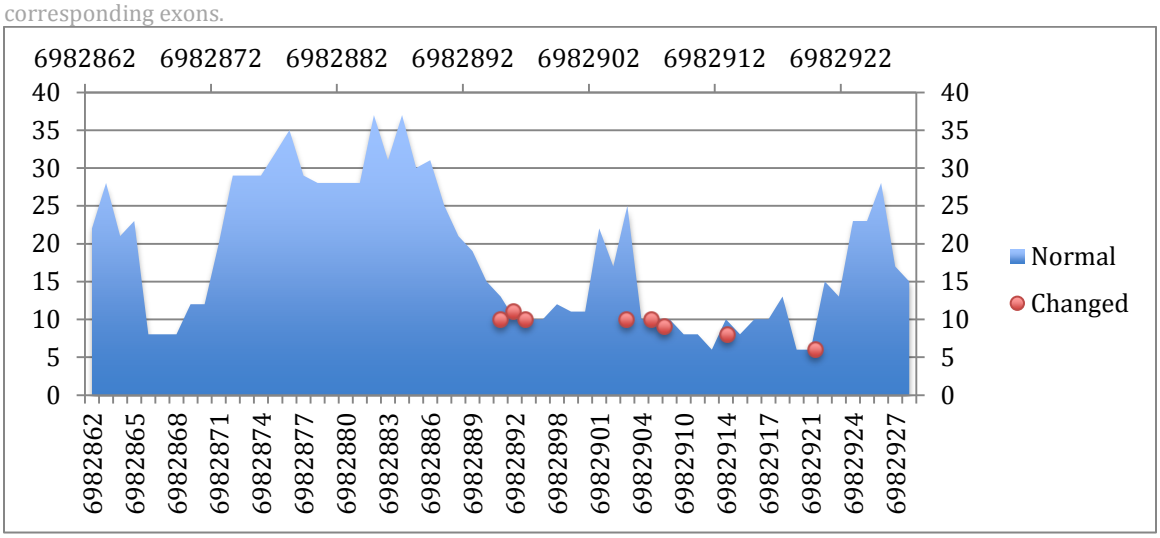


Figure 3.9.11 – Plot of the quality scores (range 0-100, in y axis) of the CD4 exon9 region the blue area chart shows the score of the normal positions. The red scatter plot shows the quality scores of the changed nucleotides. The scales of the axis of both graphs were set to the same values so that the positions match.

4 Discussion

4.1 General discussion

The Ensembl database was the sole source of sequences in this work, since it presents a fully documented API to retrieve sequences *programmatically*, while the NCBI alternative uses a url based call to the server that returns a XML, which has to be further parsed with a sequence of procedures. Furthermore, urls used to call sequence pages have been reported to change in updates without the support for the old urls.
Once all the sequences were downloaded, each gene sequences of each species had to be parsed trough multiple procedures, which would be extremely time consuming, without the creation of scripts, especially since the inputs required many parameters. Due to the repetitive nature of most of the tasks, scripting simplified the process, and when options were necessary, the scripts prompted the specific parameters, and provided summary context to the user in order to simplify the procedure. The creation of these scripts constitutes a pipeline that with slight adjustments can be used for other datasets. The construction of the pipeline though time-consuming was essential, since the analysis was run several times on the dataset, to provide optimization of parameters and to integrate sequences retrieved at a later time-point.

The initial number of species considered for analyzed was around fifty vertebrates. However, the huge diversity of sequences in such a large group increased the amount of "noise", causing synonymous substitutions to reach saturation, as well as generating alignment difficulties and differences in codon usage patterns, which would cause the branch-site test to generate false positives[81].

Classification of the genes into clusters, based on go terms, proved to be extremely difficult, even though various clustering methods, such as k-means and hierarchical clustering, were used in the attempt. Regardless, due to the lack of a semantic relation between terms, it proved impossible to differentiate genes into meaningful groups. Usage of the bioinformatics resource DAVID[82] allowed clustering of annotations based on their associations. However, this produced clusters that were not mutually exclusive. The reason for the lack of a successful clustering may be that these genes are widespread throughout the immune system processes and intimately related with its development, comprising, therefore, a single cluster.

From the 38 initially selected genes, only seven showed branches with omega ratios greater than one, as revealed by the free ratio model. The advantage of the free ratio model is its ability to screen all branches in one test, rather than having to preform one test for each branch. However, the free model ratio test is a parameter rich model and is prone to over estimation, therefore requiring further analyses, and an exclusion of estimations with high omega ratios[83] which result from a dS estimate tending to zero.

For each of these seven genes, the branches that suggested having omega ratios larger than one, were tested individually for positive selection, in order to reduce false positives, recurrent to a model that allows the foreground to assume omega ratios greater than one. This method is more rigorous than the free ratio model, but is still unrealistic, since it assumes a constant omega ratio over all codon sites[84]. In a functional protein the majority of amino acid sites are conserved to maintain its structure and function and are, therefore, subject to different selective pressures. Consequently, the branch sites model A was applied to further confirm the branches with omega ratios greater than one, estimated by the free ratio model. The branch sites model A allows various classes with different types of constraints, and is thus a direct test for positive selection on the foreground lineages[69].
Of the seven genes with suspected positive selection, the branch site model A, only two genes confirmed the presence of one positively selected branch in CD4, and two positively selected branches in IFNG. Another two genes PTCRA and HOXA3 also confirmed the presence of positive selection throughout the protein while assuming the same omega ratio to all branches.

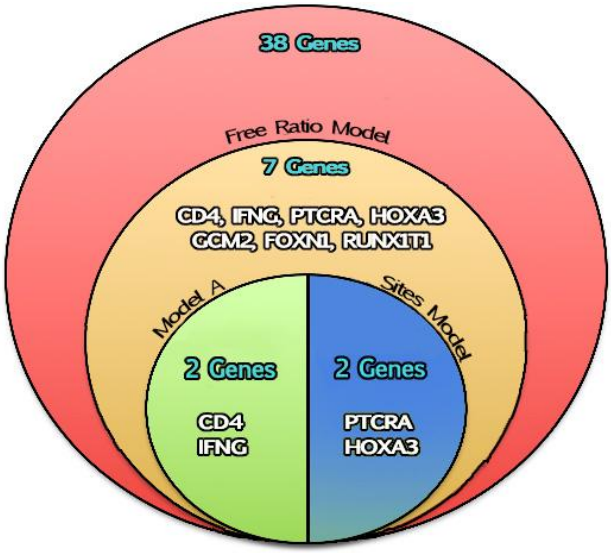


Figure 4.1.1 - A stacked Venn diagram of the classification of sequences with positive selection. Each circle shows the number of genes subjected to each Model and the resulting genes are shown in the next circle.

Analysis of 3D structures of genes, with positively selected sites shows that, three out of the four genes, with positively selected sites were on the protein surface rather than its core.

4.2 IFNG

Results regarding the gene IFNG suggest positive selection on the Tarsier ancestor branch (Tarsiiformes), and Marmoset ancestral branch (Simiiformes). These events of positive selection suggest coevolution of host and pathogen. Through their interaction, individuals that acquired the new found capability to defend against an infection were more likely to be the fittest of their populations, and transmit the

mutation throughout generations, eventually becoming fixed in the genetic background.
The tarsiiformes and simiiformes both of the suborder Haplorrhini are of the primates closest related to rodents, where positive selection has also been found[85]. The same study[85], also supports the idea that positive selection may occur in response to intracellular infectious agents, which may be responsible for the selective sweep found in Mus musculus domesticus.
As a defense against infection, IFNG induces the enzyme IDO to catabolize intracellular pools of tryptophan. When *Chamydia muridarum* (a mouse strain) infects IFNG treated human cells, there is no bacterial growth, since this strain lacks a functional tryptophan synthase and thus, enters a non-replicating, persistent state. However, *C. trachomatis* (human strain) can overcome IDO-dependent growth restriction in human cells, while in murine cell lines IFNG treatment reduces its growth[86]. This highlights the differences in IFNG function in human and murine lines, showing how it has undergone various rounds of selection, propelled by various pathogens.
Overlaying SIFT scores throughout IFNG's amino acid structure Figure 3.3.5 is facing inward on a α helix positioned at the surface of the protein. This is compatible with the fact that the effect of the amino acid modification on the protein surface should have a smaller effect on the protein folding than an amino acid modification at the protein core[87].

4.3 PTCRA

Although PTCRA gene does not show signs of branch specific selection, a great number of sites were predicted to be evolving under positive selection, which indicates that this gene is a putative selection target. Comparison of the phylogeny generated for PTCRA differs from the consensus for the whole genome. The PTCRA phylogeny places the gibbon closer to gorillas than the orangutan, whereas in the consensus phylogenetic tree[88] the orangutan is closer, followed by the gibbon. This could be due to some selection occurring specifically on this gene on the gibbon branch, which caused it to converge towards gorillas. However the positive selection detected on the gibbon branch was not significant. Thus, it cannot be argued that this is due to evolution by positive selection. While looking at the protein tertiary structure Figure 3.4.3 the amino acid residues that where mapped on the structure were also situated on the surface of the protein. However, only three amino acids, of the ones identified by the M3 model, were present in the range of amino acids for which there is a known structure.

The quaternary structure, which contains tree chains of PTRCA and three chains of TCRB, shows that the positively selected amino acids are still on the protein surface, and are not in the inter-chains region. This gives strength to the hypostasis that these mutations occur in regions that are subject o weaker purifying selective pressures than the rest of the protein, and therefore does not affect the folding or arrangement of the protein quaternary structure.

The quaternary structure was not obtained through protein modeling. Therefore, the positions of these amino acids in the quaternary structure (3of6.2.C) were confirmed by aligning the primary structure of the PTCRA gene with the primary structure of this quaternary structure (3of6.2.C).

However, all three amino acids were not present in all three chains. In chain D the positions 96-100 (112-116 in Figure 3.4.2) were omitted from the structure, which removed the positively selected amino acid ALA-97 (ALA-113 in Figure 3.4.2) and on the F chain the structure starts on amino acid 8, leaving out GLY-7 (GLY-23 in Figure 3.4.2).

4.4 HOXA3

The Hoxa3 gene shows one positively selected site, identified by M3 at the position 337. However, there is no quaternary structure described for this gene, nor is there any template closely related, which limited the ability to construct a tertiary structure of this protein, through protein modeling, to infer with accuracy the region were this amino acids would be placed. Nevertheless Hoxa3 is an extremely conserved transcription factor, responsible for position identity in the thymus organogenesis. This conservation therefore assumes a low plasticity in its structure. Search for proteins linked to this position revealed no data, suggesting that this region may not be involved in any major function, and that purifying selective pressure might be more relaxed on this region.

4.5 CD4

In the CD4 the various omega per lineage proved flawed while detecting positive selection for both the gorilla branch and the chimpanzee branch. However the likelihood ratio test (LRT) for the gorilla branch was considered more significant than the LRT for the chimpanzee branch contrary to results from Model A Table 3.9.2. Notwithstanding, in the analysis of all eight genes, this was the only case of contradiction between a broader model and a more specific model. The positively selected sites calculated through Bayesian method BEB for the chimpanzee lineage clearly show that chimpanzee amino acid residues differ from the consensus of the remaining species.
This led to the detection of a positive selection hotspot in the cytoplamatic tail of the CD4, in the chimpanzee sequence. At first hand this seems to occur due to a miss-annotation of the Ensembl sequence, since it differs from the sequences in NCBI. However, a Blastp of the sequence reveals a 98% identity with CD4 of *Pan troglodytes,* where the conserved cysteine residues of the cytoplasmatic tail are present. This fact tends to rule out an error in the Ensemble sequence, since an error in less conserved sites by chance seems less likely. Nevertheless, validation of this chimpanzee sequence revealed difficult, since none of the reported nucleotide

polymorphisms were found in non-human primate reference transcriptome resource RNAseq data[89], and only a few of those polymorphisms were found in RNAseq data from paired-end reads from chimpanzees, obtained from Hernrik Kaessmann and aligned to the CHIMP2.1.4 assembly using Burrows-Wheeler Aligner[90].

When attempting to uncover assembly files of the CHIMP2.1.4 assembly[91] only FASTA files of the chromosomes were found on NCBI server and on the Washington University server. In the last, a file with the quality of the reads was found, showing low quality scores for the CD4 region, along with the observed low coverage denoted by the amount of reads from the RNAseq that were unable to validate the sequence used in this study. The chimpanzee genome was sequenced primarily from a captive born male *Pan trogolodytes,* presumed *verus,* from Yerkes Primate Research Center in the USA named Clint.

Functional analysis of the amino acid substitution was preformed using SIFT which both identified transitions of the human amino acid residue to the chimpanzee form to be damaging in the 438 – 446 region in the cytoplasmatic tail. Interspecies studies of domesticated rice show that damaging predictions occurred in higher frequency in positively selected regions[92]. However, the loss of function does not necessarily mean disadvantage. As long as the mutation does not imply complete deletion of the gene, the mutated allele will persist in the genome, and can be reverted if the selective environment changes[93]. The most known example of gene loss of function as an advantage comes from the stickle-cell disease, which is associated with resistance to malaria infection.
The 438, 442 and 443 positions are annotated as residues where intermolecular Nuclear Overhauser effets (NOE) have been observed[94] in the CD4-LCK-Zn^{++}.
The residue 441 is responsible for the binding to the HIV nef protein, which promotes Lck dissociation
The di-leuciene residues (437 e 438) are necessary for internalization of the clarathrin adapter AP2, responsible for CD4 mobilization. A study of the inactivation of the Lck protein[95] illustrates its role in the gp120-CD4 complex's signaling suggesting a possible requirement for the binding of the nef viral protein. A mechanism for the lack of mobilization of CD4, during the HIV infection, could be the blockage of the sphingosine-1-phosphate G coupled receptor, by the binding of the HIV nef protein to CD4, and dissociation of the Lck protein.
Since CD4 is the primary HIV and SIV receptor, the study of sequence diversity can be a key candidate to uncover the lack of immunodeficiency symptoms in chimpanzees infected by SIV[96].
Comparison of the positively selected sites, detected by the Bayesian method BEB in M8, with the gp120 binding sites, shows that sites 68 and 84 are high affinity regions. Site 84 has specifically been shown to form a salt-bridge with ASP-368 residue, in the gp120[97].
Two OMIM annotated SNPs, Lys191Glu and Arg265Trp, were also identified as residues with probability of being under weaker selective pressure.

The online HyPhy server[98] was also used to identify residues subject to positive selection. Even though, FEL model positively identified three residues, and IFEL four residues, at a 0.05 significance, results show that PAML identified these same residues with similar posterior probabilities. The main reason for the preference in use of PAML instead of HyPhy, was the fact that command line use of PAML favors batch use, while batch use in HyPhy is more complicated, producing similar results.

4.6 Final remarks

Only a few of the initially selected genes, were confirmed to be under positive selection, as shown in the stacked Venn diagram Figure 4.1.1. This result is not unexpected if the nature of the genes under analysis is taken into account. The majority are "master genes", implicated in numerous signaling pathways involved in the developmental processes besides their role in the immune system. They are also conserved across vertebrates. Therefore, a non-synonymous mutation in these genes is likely to have a deleterious nature, impairing their function and probably affecting several developmental processes. CD4 and IFNG are genes with more circumscribed functions. Therefore, a non-synonymous deleterious mutation could impair fitness but not survival, contrarily to what would probably happen in SHH, which is, for example, highly involved in organogenesis.
Many of these genes are currently under study, in order to understand the complete genetic pathways in which they are involved, spanning a great variety of developmental and evolutionary questions, reflecting their many roles.

With the uprising of the NGS technology, genome acquisition is becoming popular, even for non-model organisms. However, genomes in databases assumed to be complete may not be as accurate as expected, showing ambiguities and regions with low coverage, and low quality. This is the case of CD4 gene in the chimpanzee genome analyzed in the present study. This fact highlights the necessity of the curation of bioinfomatic material available in databases.

The analysis of the initially selected genes proved to be a difficult task, since manual download of sequences demands tremendous effort, and both the API's and batch methods applied in databases are not straightforward.
An alternative approach to this difficulty was the learning of a different coding language (Perl), which then allowed the retrieval of gene sequences from the database in a couple of minutes, instead of weeks of repetitive and time-consuming tasks.
The application of PAML software to the sequences also entailed a huge automation process, since preparation of the sequences required several steps with great time frames in between. Automation was a crucial point, since various datasets had to be tested in order to reduce false positives. Functional analysis seems to validate the results obtained by application of the various substitution models in PAML.
However, the detection of positive selection remains a difficult task, since different genetic sites are subjected to different selective pressure intensity.
This type of analysis seems to be optimized to closely related species, which is a major pitfall, since it is hard to apply to a great number of species, without introducing artifacts due to alignment problems.

After surpassing technical difficulties, the three initially proposed goals, were successfully attained, and the analysis of these genes resulted in a pipeline that easily allows the discovery of positive selection in other genes of future interest.

5 References

1. Miller, M. B. & Bassler, B. L. Quorum sensing in bacteria. *Annu. Rev. Microbiol.* **55,** 165 (2001).

2. Stoka, a. Phylogeny and evolution of chemical communication: an endocrine approach. *J. Mol. Endocrinol.* **22,** 207–225 (1999).

3. Iwasaki, A. & Medzhitov, R. Regulation of adaptive immunity by the innate immune system. *Science* **327,** 291–5 (2010).

4. Bajoghli, B. *et al.* Evolution of genetic networks underlying the emergence of thymopoiesis in vertebrates. *Cell* **138,** 186–97 (2009).

5. Cooper, M. D. & Alder, M. N. The evolution of adaptive immune systems. *Cell* **124,** 815–22 (2006).

6. Seder, R. A. & Ahmed, R. Similarities and differences in CD4+ and CD8+ effector and memory T cell generation. *Nat Immunol* **4,** 835–842 (2003).

7. Itano, A. A. & Jenkins, M. K. Antigen presentation to naive CD4 T cells in the lymph node. *Nat Immunol* **4,** 733–739 (2003).

8. Théry, C. & Amigorena, S. The cell biology of antigen presentation in dendritic cells. *Curr. Opin. Immunol.* **13,** 45–51 (2001).

9. Weninger, W., Andrian, U. H. Von, Manjunath, N. & Von Andrian, U. H. Migration and differentiation of CD8+ T cells. *Immunol. Rev.* **186,** 221–233 (2002).

10. Murphy, K. M. & Reiner, S. L. The lineage decisions of helper T cells. *Nat Rev Immunol* **2,** 933–944 (2002).

11. O'Shea, J. J. & Paul, W. E. Mechanisms Underlying Lineage Commitment and Plasticity of Helper CD4+ T Cells. *Sci.* **327 ,** 1098–1102 (2010).

12. Schoenborn, J. R. & Wilson, C. B. Regulation of interferon-gamma during innate and adaptive immune responses. *Adv. Immunol.* **96,** 41–101 (2007).

13. Basler, C. F., Garci, A. & García-Sastre, A. Viruses and the type I interferon antiviral system: induction and evasion. *Int. Rev. Immunol.* **21,** 305–337 (2002).

14. Holländer, G. *et al.* Cellular and molecular events during early thymus development. *Immunol. Rev.* **209,** 28–46 (2006).

15. Radtke, F. *et al.* Deficient T Cell Fate Specification in Mice with an Induced Inactivation of Notch1. **10,** 547–558 (1999).

16. Jenkinson, E. J., Jenkinson, W. E., Rossi, S. W. & Anderson, G. The thymus and T-cell commitment: the right niche for Notch? *Nat. Rev. Immunol.* **6,** 551–5 (2006).

17. Robey, E. *et al.* An Activated Form of Notch Influences the Choice between CD4 and CD8 T Cell Lineages. **87,** 483–492 (1996).

18. Wilson, A., de Villartay, J.-P. & MacDonald, H. R. T Cell Receptor δ Gene Rearrangement and T Early α (TEA) Expression in Immature αβ LineageThymocytes: Implications for αβ/γδ Lineage Commitment. *Immunity* **4,** 37–45 (1996).

19. Pennington, D. *et al.* Early events in the thymus affect the balance of effector and regulatory T cells. *Nature* **444,** (2006).

20. Kisielow, P., Teh, H. S., Bluthmann, H. & von Boehmer, H. Positive selection of antigen-specific T cells in thymus by restricting MHC molecules. *Nature* **335,** 730–733 (1988).

21. Koyanagi, A., Sekine, C. & Yagita, H. Expression of Notch receptors and ligands on immature and mature T cells. *Biochem. Biophys. Res. Commun.* **418,** 799–805 (2012).

22. Weissman, Y. I. L. Thymus cell maturation - Studies on the origin of cortisone-resistant thymic lymphocytes. *J. Exp. Med.* **13,** 504–510 (1973).

23. Anderson, M. S. *et al.* The cellular mechanism of Aire control of T cell tolerance. *Immunity* **23,** 227–39 (2005).

24. Neves, H., Dupin, E., Parreira, L. & Le Douarin, N. M. Modulation of Bmp4 signalling in the epithelial-mesenchymal interactions that take place in early thymus and parathyroid development in avian embryos. *Dev. Biol.* **361,** 208–19 (2012).

25. Gordon, J. *et al.* Functional evidence for a single endodermal origin for the thymic epithelium. *Nat. Immunol.* **5,** 546–53 (2004).

26. Kameda, Y., Arai, Y., Nishimaki, T. & Chisaka, O. The role of Hoxa3 gene in parathyroid gland organogenesis of the mouse. *J. Histochem. Cytochem.* **52,** 641–51 (2004).

27. Gordon, J. & Manley, N. R. Mechanisms of thymus organogenesis and morphogenesis. *Development* **138,** 3865–78 (2011).

28. Grevellec, A., Graham, A. & Tucker, A. S. Shh signalling restricts the expression of Gcm2 and controls the position of the developing parathyroids. *Dev. Biol.* **353,** 194–205 (2011).

29. Gordon, J., Bennett, a R., Blackburn, C. C. & Manley, N. R. Gcm2 and Foxn1 mark early parathyroid- and thymus-specific domains in the developing third pharyngeal pouch. *Mech. Dev.* **103,** 141–3 (2001).

30. Blackburn, C. C. & Manley, N. R. Developing a new paradigm for thymus organogenesis. *Nat. Rev. Immunol.* **4,** 278–89 (2004).

31. Boehm, T. & Bleul, C. C. Thymus-homing precursors and the thymic microenvironment. *Trends Immunol.* **27,** 477–84 (2006).

32. Kyewski, B. Seeding of thymic microenvironments defined by distinct thymocyte- stromal cell interactions is developmentally controlled. *J. Exp. Med.* **166,** 520–538 (1987).

33. Liu, C. *et al.* Coordination between CCR7- and CCR9-mediated chemokine signals in prevascular fetal thymus colonization. **108,** 2531–2539 (2006).

34. Wilkinson, B., Owen, J. J. T., Jenkinson, E. J. & Alerts, E. Factors Regulating Stem Cell Recruitment to the Fetal Thymus. (2014).

35. Bleul, C. C. & Boehm, T. Chemokines define distinct microenvironments in the developing thymus. *Eur. J. Immunol.* **30,** 3371–9 (2000).

36. Ara, T. *et al.* A Role of CXC Chemokine Ligand 12/Stromal Cell-Derived Factor-1/Pre-B Cell Growth Stimulating Factor and Its Receptor CXCR4 in Fetal and Adult T Cell Development in Vivo. *J. Immunol.* **170,** 4649–4655 (2003).

37. Wurbel, M. *et al.* Mice lacking the CCR9 CC-chemokine receptor show a mild impairment of early T- and B-cell development and a reduction in T-cell receptor ⍰⍰ ς gut intraepithelial lymphocytes. 2626–2632 (2014). doi:10.1182/blood.V98.9.2626

38. Liu, C. *et al.* The role of CCL21 in recruitment of T-precursor cells to fetal thymi. *Blood* **105,** 31–9 (2005).

39. Barthlott, T., Keller, M. P., Krenger, W. & Holländer, G. a. A short primer on early molecular and cellular events in thymus organogenesis and replacement. *Swiss Med. Wkly.* **137 Suppl ,** 9S–13S (2007).

40. Kawakami, N. *et al.* Roles of Integrins and CD44 on the Adhesion and Migration of Fetal Liver Cells to the Fetal Thymus. (2014).

41. Bajoghli, B. *et al.* Evolution of genetic networks underlying the emergence of thymopoiesis in vertebrates. *Cell* **138,** 186–97 (2009).

42. Boehm, T. & Bleul, C. C. The evolutionary history of lymphoid organs. *Nat. Immunol.* **8,** 131–5 (2007).

43. Yoshino, N., Ami, Y. & Terao, K. Upgrading of Flow Cytometric Analysis for Absolute Counts , Cytokines and Other Antigenic Molecules of Cynomolgus Monkeys (Macaca fascicularis) by Using Anti-Human Cross-Reactive Antibodies. **49,** 97–110 (2000).

44. Chen, Z. W., Kou, Z., Shen, L., Reimann, A. & Letvin, N. L. Conserved T-cell Receptor Repertoire in Simian Immunodeficiency Virus-Infected Rhesus Monkeys'. **151,** 2177–2187 (1993).

45. Messaoudi, I., Estep, R., Robinson, B. & Wong, S. W. Nonhuman primate models of human immunology. *Antioxid. Redox Signal.* **14,** 261–73 (2011).

46. Benson, D. A. *et al.* GenBank. *Nucleic Acids Res.* **28 ,** 15–18 (2000).

47. Hubbard, T. *et al.* The Ensembl genome database project. *Nucleic Acids Res.* **30 ,** 38–41 (2002).

48. Boeckmann, B. *et al.* The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.* **31 ,** 365–370 (2003).

49. Kanehisa, M. *et al.* KEGG for linking genomes to life and the environment. *Nucleic Acids Res.* **36 ,** D480–D484 (2008).

50. Maglott, D., Ostell, J., Pruitt, K. D. & Tatusova, T. Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res.* **33 ,** D54–D58 (2005).

51. Krauthammer, M., Rzhetsky, A., Morozov, P. & Friedman, C. Using BLAST for identifying gene and protein names in journal articles. *Gene* **259,** 245–252 (2000).

52. Consortium, T. E. P. The ENCODE (ENCyclopedia Of DNA Elements) Project. *Sci.* **306 ,** 636–640 (2004).

53. Hedrick, P. W. Balancing selection. *Curr. Biol.* **17,** R230–R231 (2007).

54. Crow, J. F. Anectdotal , Historical and Critical Commentaries on Genetics 90 Years Ago : The Beginning of Hybrid Maize. *Genet. Soc. Am.* **148,** 923–928 (1998).

55. Ohta, T. The Nearly Neutral Theory Of Molecular Evolution. *Annu. Rev. Ecol. Syst.* **23,** 263–286 (1992).

56. Yang, Z. & Bielawski, J. Statistical methods for detecting molecular adaptation. *Trends Ecol. Evol.* **15,** 496–503 (2000).

57. MAYNARD, J. & HAIGH, J. The hitch-hiking effect of a favourable gene. *Genet. Res. (Camb).* **89,** 391–403 (2007).

58. McDonald, J. & Kreitman, M. Adaptive protein evolution at the Adh locus in Drosophila. *Nature* **351,** 652–654 (1991).

59. Charlesworth, J. & Eyre-Walker, A. The McDonald-Kreitman test and slightly deleterious mutations. *Mol. Biol. Evol.* **25,** 1007–15 (2008).

60. 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 (1985).

61. Vitti, J. J., Grossman, S. R. & Sabeti, P. C. Detecting natural selection in genomic data. *Annu. Rev. Genet.* **47,** 97–120 (2013).

62. Hurst, L. The K a/ K s ratio: diagnosing the form of sequence evolution. *Trends Genet.* **18,** 486–487 (2002).

63. Nielsen, R. & Yang, Z. Estimating the distribution of selection coefficients from phylogenetic data with applications to mitochondrial and viral DNA. *Mol. Biol. Evol.* **20,** 1231–9 (2003).

64. Group, N. P. In search of molecular darwinism. *Nature* **385,** 111–112 (1997).

65. Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24,** 1586–91 (2007).

66. Bielawski, J. P. & Yang, Z. A maximum likelihood method for detecting functional divergence at individual codon sites, with application to gene family evolution. *J. Mol. Evol.* **59,** 121–32 (2004).

67. Nielsen, R. & Yang, Z. Likelihood Models for Detecting Positively Selected Amino Acid Sites and Applications to the HIV-1 Envelope Gene. **936,** 929–936 (1998).

68. Yang, Z. Bayesian Inference In Molecular Phylogenetics. *Syst. Biol.* **54,** 455–470 (2005).

69. Zhang, J., Nielsen, R. & Yang, Z. Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. *Mol. Biol. Evol.* **22,** 2472–9 (2005).

70. Yang, Z. & Nielsen, R. Codon-Substitution Models for Detecting Molecular Adaptation at Individual Sites Along Specific Lineages. 908–917 (2001).

71. Ge, Q. & Zhao, Y. Evolution of thymus organogenesis. *Dev. Comp. Immunol.* **39,** 85–90 (2012).

72. Abascal, F., Zardoya, R. & Telford, M. J. TranslatorX: multiple alignment of nucleotide sequences guided by amino acid translations. *Nucleic Acids Res.* **38,** W7–13 (2010).

73. Katoh, K., Misawa, K., Kuma, K. & Miyata, T. MAFFT : a novel method for rapid multiple sequence alignment based on fast Fourier transform. **30,** 3059–3066 (2002).

74. Stamatakis, a, Ludwig, T. & Meier, H. RAxML-III: a fast program for maximum likelihood-based inference of large phylogenetic trees. *Bioinformatics* **21,** 456–63 (2005).

75. Ng, P. C. SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Res.* **31,** 3812–3814 (2003).

76. Bateman, A. *et al.* The Pfam protein families database. *Nucleic Acids Res.* **32,** D138–41 (2004).

77. Schwede, T., Kopp, J., Guex, N. & Peitsch, M. C. SWISS-MODEL: an automated protein homology-modeling server. *Nucleic Acids Res.* **31 ,** 3381–3385 (2003).

78. Ng, P. C. & Henikoff, S. Predicting the effects of amino acid substitutions on protein function. *Annu. Rev. Genomics Hum. Genet.* **7,** 61–80 (2006).

79. Landar, a *et al.* Design, characterization, and structure of a biologically active single-chain mutant of human IFN-gamma. *J. Mol. Biol.* **299,** 169–79 (2000).

80. Pang, S. S. *et al.* The structural basis for autonomous dimerization of the pre-T-cell antigen receptor. *Nature* **467,** 844–8 (2010).

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

82. Huang, B. D. W. & Lempicki, R. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* **4,** 44–49 (2008).

83. McKay, P. B. & Griswold, C. K. A comparative study indicates both positive and purifying selection within ryanodine receptor (RyR) genes, as well as correlated evolution. *J. Exp. Zool. A. Ecol. Genet. Physiol.* **321,** 151–63 (2014).

84. Yang, Z. Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. *Mol. Biol. Evol.* **15,** 568–73 (1998).

85. Levi-Acobas, F., Mars, L. T., Orth, a, Bureau, J.-F. & Bonhomme, F. Adaptive evolution of interferon-gamma in Glire lineage and evidence for a recent selective sweep in Mus. m. domesticus. *Genes Immun.* **10,** 297–308 (2009).

86. Coers, J. *et al.* Chlamydia muridarum Evades Growth Restriction by the IFN- -Inducible Host Resistance Factor Irgb10. *J. Immunol.* **180,** 6237–6245 (2008).

87. Caffrey, D. R., Somaroo, S., Hughes, J. D., Mintseris, J. & Huang, E. S. Are protein – protein interfaces more conserved in sequence than the rest of the protein surface ? 190–202 (2004). doi:10.1110/ps.03323604.Many

88. Perelman, P. *et al.* A molecular phylogeny of living primates. *PLoS Genet.* **7,** e1001342 (2011).

89. Pipes, L. *et al.* The non-human primate reference transcriptome resource (NHPRTR) for comparative functional genomics. **41,** 906–914 (2013).

90. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25,** 1754–60 (2009).

91. Sequencing, T. C. & Consortium, A. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* **437,** 69–87 (2005).

92. Lu, J. *et al.* The accumulation of deleterious mutations in rice genomes: a hypothesis on the cost of domestication. *Trends Genet.* **22,** 126–31 (2006).

93. Olson, M. V. MOLECULAR EVOLUTION ' 99 When Less Is More : Gene Loss as an Engine of Evolutionary Change. 18–23 (1999).

94. Kim, P. W., Sun, Z.-Y. J., Blacklow, S. C., Wagner, G. & Eck, M. J. A zinc clasp structure tethers Lck to T cell coreceptors CD4 and CD8. *Science* **301,** 1725–8 (2003).

95. Green, D. S., Center, D. M. & Cruikshank, W. W. Human immunodeficiency virus type 1 gp120 reprogramming of CD4+ T-cell migration provides a mechanism for lymphadenopathy. *J. Virol.* **83,** 5765–72 (2009).

96. Hvilsom, C. *et al.* Genetic subspecies diversity of the chimpanzee CD4 virus-receptor gene. *Genomics* **92,** 322–8 (2008).

97. Fontenot, D. *et al.* Critical role of Arg59 in the high-affinity gp120-binding region of CD4 for human immunodeficiency virus type 1 infection. *Virology* **363,** 69–78 (2007).

98. Pond, S. L. K. & Frost, S. D. W. Datamonkey: rapid detection of selective pressure on individual sites of codon alignments. *Bioinformatics* **21,** 2531–3 (2005).

99. Gilbert, S. F. *Developmental Biology*. (Sinauer Associates, 2010).

6 Appendix

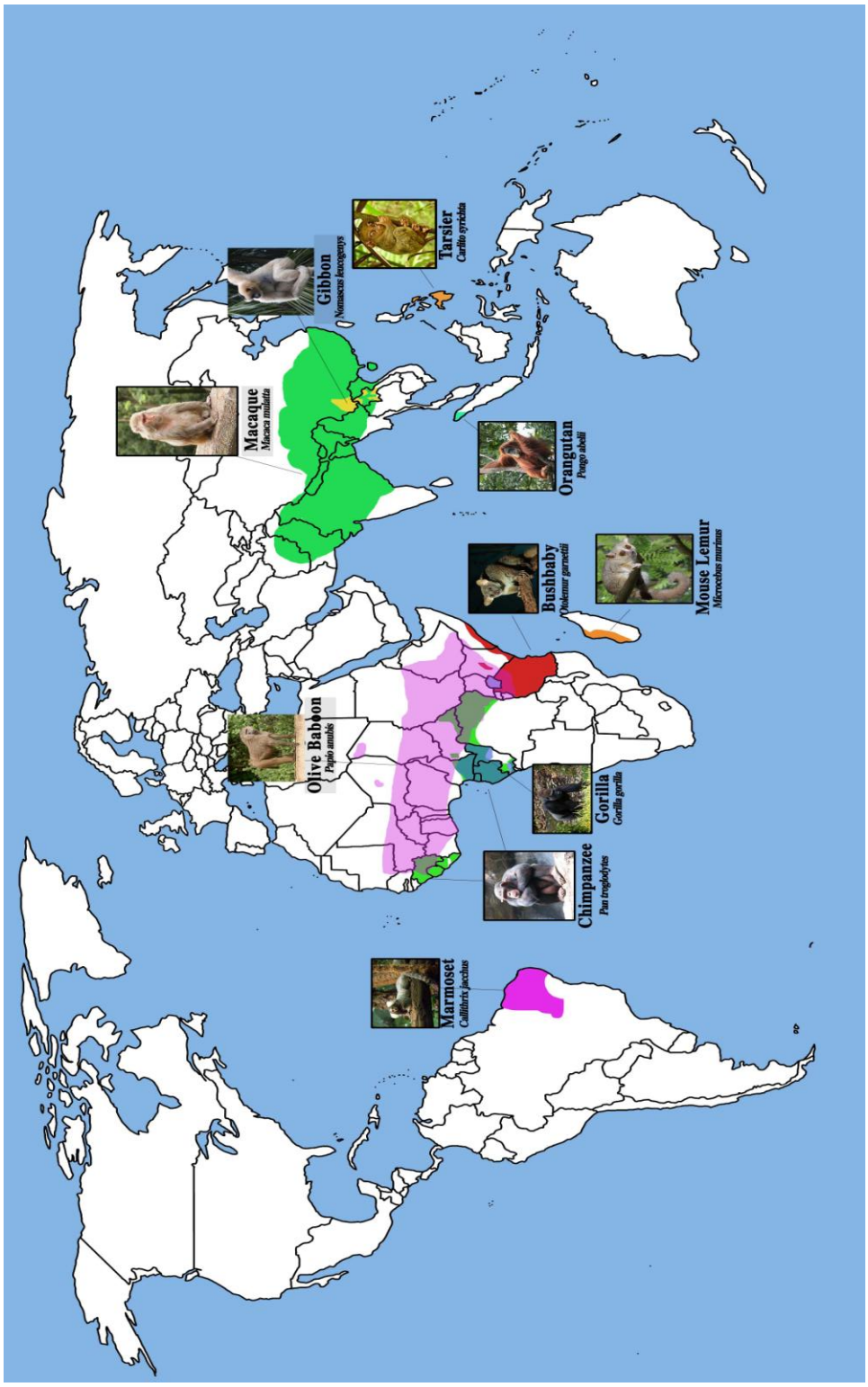


Figure 4.6.1 – Global distribution of primate species analyzed in the present study

Table 6.01 – Gene classification into biological categories.

Row Labels	Biological adhesion	Biological regulation	Cell killing	Cellular component organization or biogenesis	Cellular process	Developmental process	Growth	Immune system process	Localization	Locomotion	Metabolic process	Multi-organism process	Multicellular organismal process	Reproduction	Response to stimulus	Rhythmic process	Signaling	Single-organism process	Grand Total
ACKR2		x			x	x		x		x			x		x		x	x	9
ACKR3	x	x			x	x				x		x	x		x		x	x	10
ADAM10	x	x		x	x	x	x	x	x	x	x		x		x		x	x	14
ADAM17	x	x		x	x	x	x	x	x	x	x	x	x		x		x	x	15
AIRE		x			x			x			x				x				5
CCL25		x			x			x	x	x					x		x	x	8
CCR9		x			x			x		x					x		x	x	7
CD4	x	x			x	x		x	x	x	x	x	x		x		x	x	13
CD8A		x			x	x		x				x	x		x		x	x	9
CHUK		x			x	x		x	x		x	x	x	x	x		x	x	12
CTNNB1	x	x		x	x	x		x	x	x	x		x	x	x		x	x	14
CXCL12	x	x		x	x	x		x	x	x		x	x	x	x		x	x	14
CXCR4		x		x	x	x		x	x	x	x	x	x	x	x		x	x	14
DLL1	x	x			x	x		x			x		x		x		x	x	10
DTX1		x			x	x		x			x		x		x		x	x	9
FOXN1		x			x	x		x			x		x		x			x	8
HES1		x			x						x								3
HOXA3		x			x	x		x			x		x					x	7
DLL4		x			x	x			x	x	x		x		x		x	x	10
GCM2		x			x	x					x		x		x			x	7
HOXB4		x			x	x		x			x		x					x	7
IFNG	x	x	x	x	x	x	x	x	x	x	x	x	x		x		x	x	16
IL17B		x			x			x					x		x		x	x	7
IL2RA		x			x	x		x				x	x		x		x	x	9
IL6ST		x			x	x		x			x		x		x		x	x	9
JAG1		x			x	x		x	x	x	x		x		x		x	x	11
JAG2	x	x			x	x		x	x	x	x		x	x	x		x	x	13
LMO2		x			x	x		x			x		x		x			x	8
NFKB1		x			x	x		x	x		x	x	x		x		x	x	11
NOTCH2		x			x	x	x	x	x		x		x		x		x	x	11
NOTCH4		x		x	x	x		x			x		x		x		x	x	10
PSEN1	x	x		x	x	x	x	x	x	x	x		x		x		x	x	14
PSEN2		x			x	x	x	x	x	x	x		x		x		x	x	12
PTCRA		x			x													x	3
RUNX1		x			x	x		x	x		x		x	x	x	x	x	x	12
RUNX1T1		x			x	x					x							x	5
SHH		x		x	x	x	x	x	x	x	x		x	x	x		x	x	14
STAT6		x			x	x		x			x		x		x		x	x	9
Grand Total	10	38	1	9	38	32	7	32	18	17	29	10	32	7	33	1	29	36	379

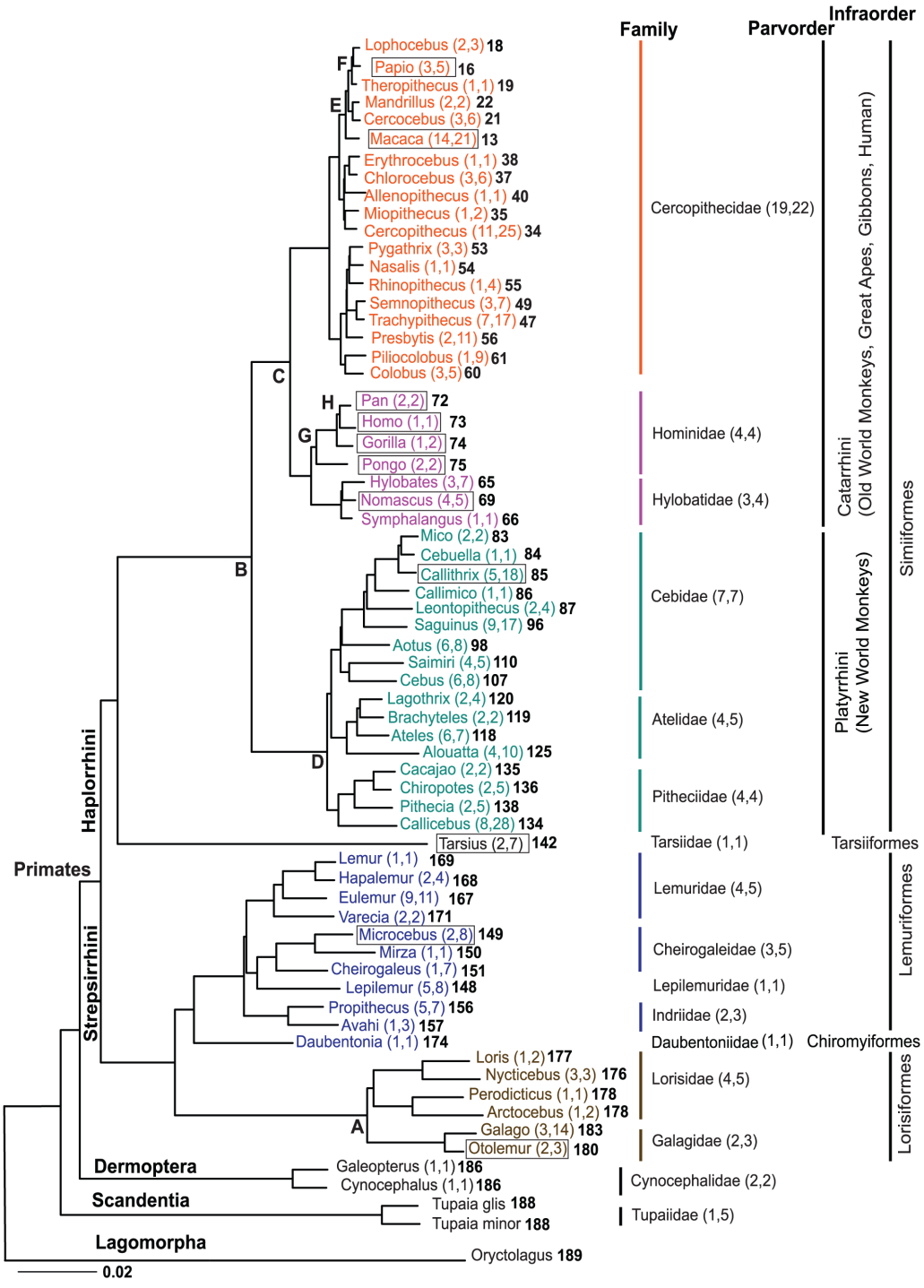


Figure 6.0.2 - Adapted from P.Perlman, W. Johnsom, et. al. ;2011; Plos genetics. A consensus phylogentic tree of living primates using genomic data. The species related to this study are highlighed with boxed around the species name.

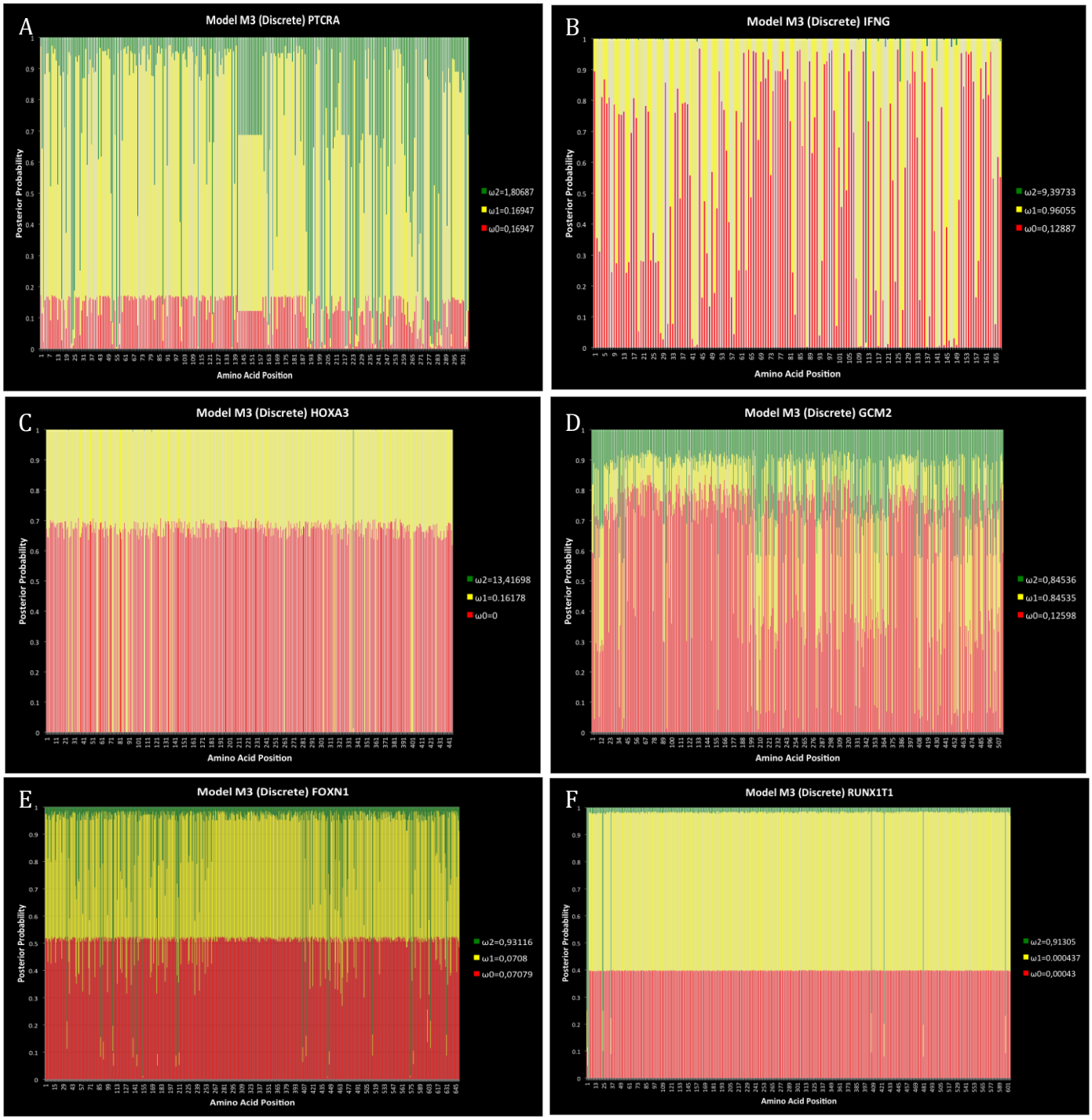


Figure 6.0.3 Stacked histograms representing posterior probabilities for the three site classes with different selective pressures identified by the CODEML model M3 (discrete). A PTCRA gene B IFNG gene C HOXA3 D GCM2 gene E FOXN1 gene F RUNX1T1 gene.