U.PORTO

CICLO DE ESTUDOS MESTRADO

MEDICINA LEGAL

The genes from the pseudoautosomal region 1 (PAR1) of the mammalian sex chromosomes: synteny, phylogeny and selection

Carla Sofia Ribeiro dos Santos

M

2020

U.PORTO

Carla Sofia Ribeiro dos Santos The genes from the pseudoautosomal region 1 (PAR1) of the mammalian sex chromosomes: synteny, phylogeny and selection

The genes from the pseudoautosomal region 1 (PAR1) of the mammalian sex chromosomes: synteny, phylogeny and selection

Carla Sofia Ribeiro dos Santos

M.ICBAS **2020**

INSTITUTO DE CIÊNCIAS BIOMÉDICAS ABEL SALAZAR

Carla Sofia Ribeiro dos Santos

The genes from the pseudoautosomal region 1 (PAR1) of the mammalian sex chromosomes: synteny, phylogeny and selection.

Candidature dissertation to the Master degree in Legal Medicine submitted to Institute of Biomedical Sciences Abel Salazar (ICBAS) of University of Porto

Supervisor: Prof. Dr. Agostinho Antunes

Category: Principal Investigator & Professor

Affiliations: Interdisciplinary Centre of Marine and Environmental Research (CIIMAR/CIMAR) and Department of Biology, Faculty of Sciences, University of Porto

Co-Supervisor: Tito Mendes

Category: Investigator

Affiliation: Interdisciplinary Centre of Marine and Environmental Research (CIIMAR/CIMAR) and Department of Biology, Faculty of Sciences, University of Porto

Abstract

Sexual chromosomes recombination is restricted to a homologous area common to both, the pseudoautosomal region (PAR), composed by PAR1 and PAR2, which behaves like an autosome in both pairing and recombination. The PAR1 region, common to most of the eutherian mammals, is located at the terminus of the short arm of the sexual chromosomes and presents a recombination rate 20 times higher than the autosomes. In order to gain insight into the evolution and the interspecific differences of PAR1, 15 genes from the PAR1 region were collected from 41 mammalian genera (representing six orders) and were used to perform phylogenetic and selection analyses. The synteny of the PAR1 genes was also analysed revealing differences among mammalian species, especially in Rodentia, order in which chromosomic translocations to the autosomes were observed. Differences between the expected and the produced phylogenetic trees were found, including an increase in the branch length for rodents. Regarding the selection analyses performed, three genes (*ASMT*, *PLCXD1* and *ZBED1*) exhibited positive selection signatures in site models. Additionally, two orders, Rodentia and Primates displayed an inflated ω value in the branch models. Lastly, concerning the branch-site models, positively selected sites were obtained for the branches of the two assessed orders, specifically two genes in Primates (*ASMTL* and *IL3RA*) and five genes in Rodentia (*ASMT*, *CSF2RA*, *IL3RA*, *P2RY8* and *PPP2R3B*). The lack of strong positive selection may reinforce the evolutionary constraints imposed by the important function of the PAR1 genes. Mutations in these genes are associated with various diseases, including stature problems (Klinefelter Syndrome), leukaemia and mental diseases. Finally, we hypothesized that the Euarchontoglires superorder, in which Primates and Rodentia are included, may have a predisposition to positive selection for some of the PAR1 genes, as suggested by the positive selection evidences exclusively found for these two orders. Furthermore, when compared with other Eutherians, both these orders have distinctive PARs. Rodentia has the smallest PAR1 and simian primates/humans PAR1 has been reduced in 3-5 fold of the size. Additionally, based on the PAR1 genes translocation to autosomes in the Rodentia order and on the stronger evidences of positive selection when comparing to Primates, we suggest that such genome migration may have affected the selection pressures in the PAR1 genes.

Resumo

A recombinação nos cromossomas sexuais está restrita a uma zona comum a ambos, denominada por região pseudoautossomal (PAR), sendo composta pelas regiões PAR1 e PAR2, e que se comporta como um autossoma em termos de emparelhamento e recombinação. A região PAR1, comum à maioria dos mamíferos eutérios, localiza-se na extremidade do braço curto dos cromossomas sexuais e apresenta uma recombinação 20 vezes mais elevada que os autossomas. De forma a melhor compreender a evolução e diferenças interespecíficas ocorrentes no PAR1, 15 genes que o compõem foram recolhidos de 41 géneros de mamíferos e usados para executar análises filogenéticas e de seleção. A sintenia dos genes do PAR1 foi também analisada, revelando diferenças entre várias espécies de mamíferos, em particular nos Roedores, ordem na qual foram observadas translocações cromossómicas para os autossomas. Foram encontradas diferenças entre a árvore filogenética expectável e as filogenias produzidas, incluindo um aumento no comprimento dos ramos nos Rodentia. Ao executar as análises de seleção, três genes (*ASMT*, *PLCXD1* e *ZBED1*) exibiram sinais de seleção positiva nos *site models*. Para além disso, duas ordens, primatas e roedores, apresentaram um valor elevado de ω nos *branch models*. Por último nos *branch-site models*, foram detetados resíduos positivamente selecionados para os ramos das duas ordens analisadas, especificamente em cinco genes para os Rodentia (*ASMT*, *CSF2RA*, *IL3RA*, *P2RY8* e *PPP2R3B*) e em dois genes para os Primates (*ASMTL* e *IL3RA*). A escassez de seleção positiva pode estar relacionada com a função importante dos genes do PAR1. Mutações nestes genes estão associadas com inúmeras doenças, tal como problemas de estatura (síndrome de Klinefelter), leucemia e doenças mentais. Concluindo, foi colocada a hipótese que a superordem Euarchontoglires, à qual os primatas e os roedores pertencem, poderá ter uma predisposição para seleção positiva para alguns genes do PAR1, como sugerido pelas evidências de seleção positiva encontradas em exclusivo para estas duas ordens. Inclusive, quando comparado com outros eutéria, ambas as ordens têm PAR muito distintos, sendo que os Rodentia têm o PAR1 mais pequeno e os símios/humanos tiveram o tamanho do seu PAR1 reduzido em 3-5 vezes. Adicionalmente, com base na translocação dos genes do PAR1 para os autossomas nos Rodentia e nas evidências mais demarcadas de seleção positiva quando comparando com os Primatas, sugerimos que a migração genómica pode afetar as pressões seletivas nos genes do PAR1.

Keywords

Mammals, Eutherians, XY system, Sex determination, Evolution, Recombination, Homologous chromosomes, Legal Medicine, Sex chromosome diseases.

List of abbreviations

AICc - Corrected Akaike Information Criterion
BEB - Bayes Empirical Bayes
BI - Bayesian Inference
BUSTED - Branch-site Unrestricted Statistical Test for Episodic Diversification
CR - Chromosome
DAMBE5 - Data Analysis for Molecular Biology and Evolution 5
dN - Nonsynonymous substitution rate
dS - Synonymous substitution rates
E.g. - *Exempli gratia*
FEL - Fixed Effects Likelihood
FUBAR - Fast, Unconstrained Bayesian AppRoximation for Inferring Selection
GARD - Genetic Algorithm for Recombination Detection
Iss - Index of substitution saturation
IssC - Critical index of substitution saturation
LRT - Likelihood Ratio Test
MEME - Mixed Effects Model of Evolution
ML - Maximum-Likelihood
P - P-value
PAB - Pseudoautosomal boundary
PAML - Phylogenetic Analysis by Maximum Likelihood
PAR - Pseudoautosomal region
PP - Posterior probability
PSS - Positively selected sites
SLAC - Single-Likelihood Ancestor Counting
XCI - X chromosome inactivation
ω - Ratio of nonsynonymous to synonymous substitution rates

Index

Introduction

The process of sex determination can either be controlled by environmental or genotypic factors[1]. Environmental sex determination is a process in which the sex of a given individual is defined by environmental factors, such as temperature and social factors, regardless of the individual´s genetic information[2]. Environmental sex determination occurs mainly among unicellular eukaryotes, but it can also be found in multicellular species, like reptiles, amphibians, and some fish[2]. For example, the anemonefish *Amphiprion akallopisos* can be subjected to a sex change when the dominant female of its group dies. Here, the larger male becomes the new dominant female, which categorizes this species as sequential hermaphrodite[2 3 4 5]. On the other side, temperature-dependent sex determination is typical in some reptiles, such as *Sphenodon* sp., all crocodilians, and many species of turtles and lizards[6]. For this type of sex determination, the temperature during the egg development determines the sex of the embryo. For example in crocodilians, females are determined by cool or high temperature, while males are determined by warm temperatures[7 8]. On the other hand, the genotypic sex determination systems are based on the differences of either gene or allele content between males and females. For those cases involving a pair of morphologically distinguishable chromosomes (sex chromosomes), one of the genders is heterogametic and the other is homogametic. In the ZW sexual system present in birds and snakes, females are heterogametic[1]. In contrast, the females with a XY sexual system, represented in most of mammals, have two X chromosomes, while the males possess one X chromosome and one Y chromosome, thus the latter being the heterogametic gender[1].

Both X and Y chromosomes originated from a pair of autosomes, in which the eutherian X chromosome has evolved before the eutherian–marsupial split[1]. Then the sex chromosomes differentiated substantially from each other over time[9], mostly due to Y chromosome degradation[10]. The degradation of the Y chromosome (figure 1) is a result of recombination suppression that led to gene loss and, consequently, the degeneration of the

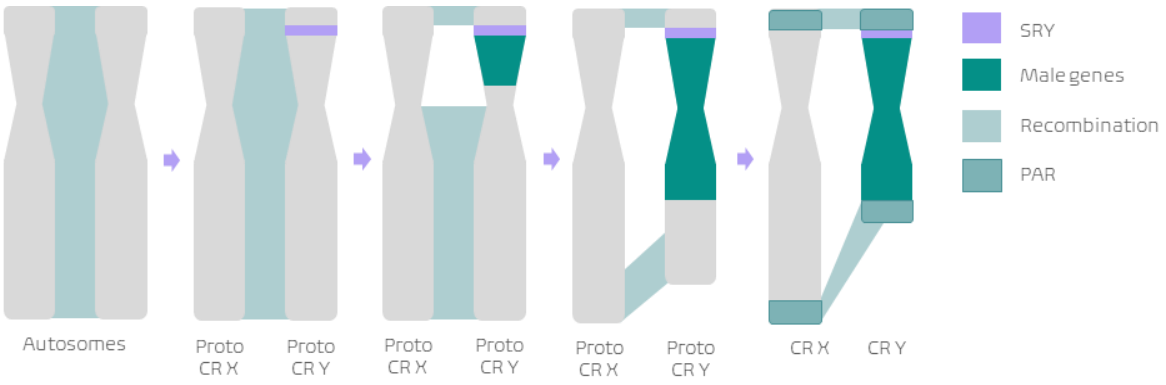

Figure 1 – Y chromosome degradation during the formation of the sexual chromosomes

chromosome. This degradation, due to a recombination suppression proves the importance of the recombination mechanism, which guarantees the proper pairing and segregation of the pairs of chromosomes during meiosis[11].

The recombination between the X and Y chromosomes is restricted to the region common to both chromosomes, the pseudoautosomal region (PAR, figure 2-A), which behaves as an autosome in pairing and recombination. Being the only region to be subjected to recombination in the sex chromosomes, the PAR exhibits a high degree of conservation among eutherian species[11]. In females, a process of inactivation of one of the X chromosomes (X chromosome inactivation - XCI) occurs in order to obtain a gene dosage equilibrium[12]. The PAR is the only region that consistently escapes XCI in all species and maintains activity in both X chromosomes, thus occurring pair recombination in both females and males. The PAR1, the most prominent region of the PAR, is located at the termini of the short arm of the sex chromosomes (in the first 2,7 Mb)[13]. This region, present in humans and common to most of the eutherian mammals, has an extraordinarily high rate of recombination compared to the autosomes (~20 times higher)[9]. The boundary of the human PAR1 (PAB - pseudoautosomal boundary) was originated by the most recent evolutionary strata, defined as “regions of the sex chromosomes which stopped

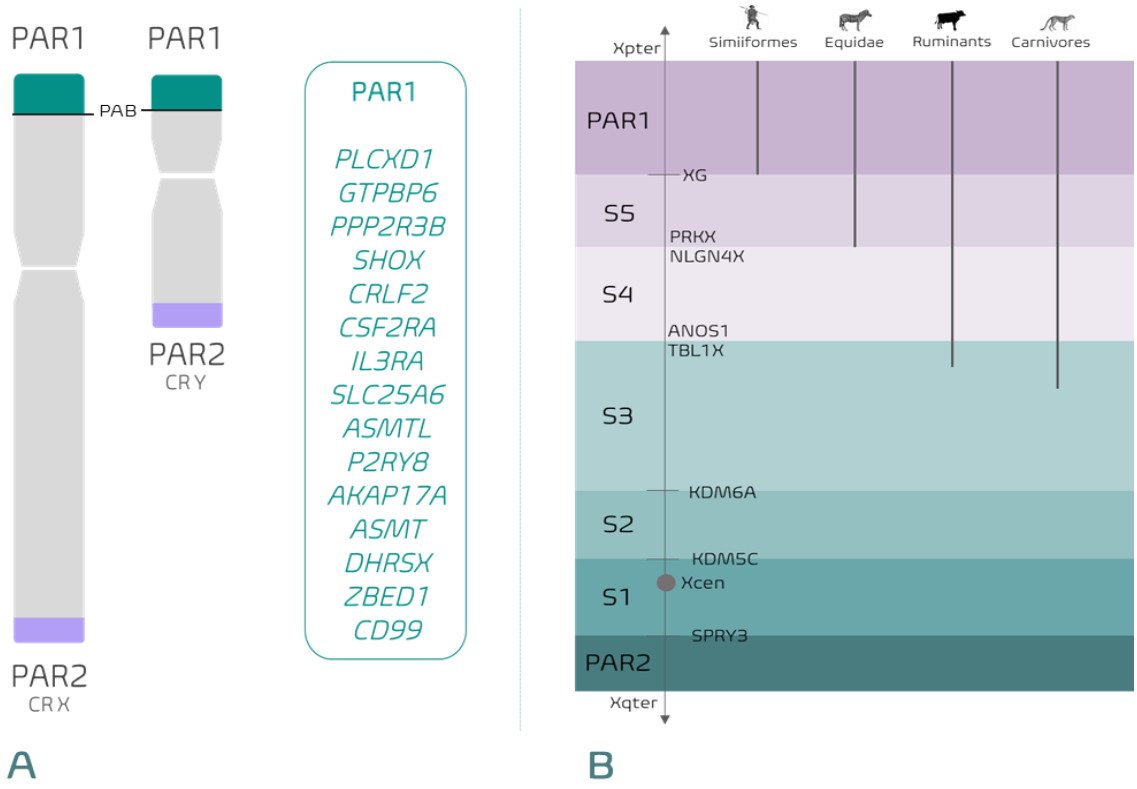

Figure 2 – A - Representation of both human PAR regions (PAR1 and PAR2) and PAB for PAR1 in the X and Y chromosomes. Names of the 15 genes of this region in the green box. B - Representation of the evolutionary strata (S1, S2, S3, S4 and S5) and PAR1 for some groups (Simiiformes, Equidae, Ruminants and Carnivores).

recombining with their homologues at different times in the past", according to Bergero & Charlesworth (2009) (figure 2-B). The PAR1 is one of the differences when comparing to the pre-existing sexual chromosomes shared with the marsupials, being this attested by the absence of this region in the species of this infraclass. Furthermore, the most basal groups, monotremes and marsupials, have quite distinct sexual chromosomes[14]. Monotremes exhibit multiple sex chromosomes[15], in contrast to marsupials, in which a XY pair is present, although these chromosomes share no homology and where recombination does not occur[13 16 17]. Moreover, the X-Y divergence time of gene pairs increases the further we get from PAR, which means that gene pairs closest to the PAR were recombining until more recently[1]. Consequently, more differences among species are found in the PAB, in both content and size, being PAR in humans delimited by the *XG* gene[1]. Although PAR1 is generally highly conserved among species, some interspecific differences can be found near PAB that affects the total length of this region. Nevertheless, there are some exceptions that exhibit a very divergent PAR1. According to Raudsepp & Chowdhary (2016), the Rodentia order has the shorter and most divergent PARs, with less genes, when compared to other eutherian species, possibly due to the rapid evolution of PAR in this group[18 19]. A considerable amount of changes also shaped this region in simian primates/humans and equids, reducing the size of PAR to about 3-5 fold when compared to other eutherians, such as ruminants, pigs, carnivores, camelids, cetaceans and prosimian primates. The groups referred previously display higher PAR homology among them and with the putative ancestral configuration than when compared with simian primates/humans and equids PAR1[11]. Even though there are some differences between species, especially around PAB, the human PAR1 sequence remains homologue, considering the gene synteny, to those of other Catarrhine primates (old world monkeys, great apes and gibbons)[10 20]. Humans have also another pseudoautosomal region, named PAR2, at the sex chromosome's terminus of the long arm[13], which is stated as being exclusive to humans[11]. Unlike PAR1, PAR2 has a recombination rate only ~5 times higher than the autosomes[9].

The main goal of this work is to better understand the evolution and the interspecific differences existing in the genes of the PAR1 region of the sex chromosomes, which are linked with serious diseases (e.g. male infertility, Klinefelter Syndrome, leukaemia, mental diseases[11]), by performing evolutionary genomics and bioinformatics analyses. A way of studying the evolution of a gene and its corresponding encoded protein is by performing selection analyses. Selection analyses can identify two types of sites: a) sites that are conserved, which indicate that they have an important function (e.g. sites needed for protein-protein interactions); and, b) highly variable sites, which may suggest the occurrence of positive selection[21 22]. Conserved sites are usually subjected to purifying

selection, which benefits the synonymous substitutions, indicating that substitutions which altered the coded amino acid have a negative effect on fitness. On the other side, highly variable sites can be the target of positive selection that benefits non-synonymous substitutions, due to an evolutionary advantage of the organism to such change[23]. Concluding, with the selection analyses we may be able to understand the changes that the genes of the PAR1 have gone through and correlate those changes with the evolutionary pattern of the assessed species.

Materials and Methods

Sequence collection, multiple sequence alignments and saturation assessment

Protein-coding nucleotide sequences of the 15 human PAR1 genes (*AKAP17A*, *ASMT*, *ASMTL*, *CD99*, *CRFL2*, *CSF2RA*, *DHRSX*, *GTPBP6*, *IL3RA*, *P2RY8*, *PLCXD1*, *PPP2R3B*, *SHOX*, *SLC25A6*, *ZBED1*) from 41 genera of the assessed mammalian orders (Primates, Artiodactyla, Perissodactyla, Carnivora, Rodentia and Lagomorpha) were retrieved from the GenBank[24] and Ensembl databases[25] (table S1). Resorting only to a single species of each genus was not possible to complete the database for further analyses (phylogenetic and selection analyses). Therefore, to overcome this problem various species of the same genus were retrieved. Hence, for the same operational taxonomic unit several species from the same genera were used and referred by their genus name. Additionally, for absent sequences, the option protein2genome of the Exonerate v2.2 software[26] was used, employing as a query a sequence already retrieved, from the same order as the required sequence. A codon-based multiple sequence alignment was performed for each gene (15 datasets with all the retrieved mammalian genera), using the GUIDANCE2 webserver[27], employing the MAFFT algorithm, 100 bootstrap repeats and the pre-defined level of confidence threshold (0.93). Afterwards, all alignments were manually refined. Saturation signatures were tested employing the Xia´s test[28] available in the Data Analysis for Molecular Biology and Evolution 5 (DAMBE5) software[29]. In order to calculate the saturation signatures of the genes, the values of Iss (index of substitution saturation) and IssC (critical index of substitution saturation) were compared. Whenever the Iss value was significantly lower than the IssC ($P < 0.05$), a low and not significant saturation was considered.

Synteny analyses

Mammalian genomes and annotated files of X chromosomes from several species were retrieved from the GenBank and Ensembl databases (table 1). In contrast to the remaining analyses, for this section of the work mammalian species were used instead of genera. The chromosomic location of the 15 selected genes from the PAR1 were retrieved from the annotation files. A query with the PAR1 genes was constructed to retrieve the remaining locations from the genomes for each assessed mammalian order (Primates, Artiodactyla, Perissodactyla, Rodentia, Carnivora and Lagomorpha) using the annotated genes from a model species (*Homo sapiens* for Primates, *Bos indicus x Bos Taurus* (Braford breed[30]) for Artiodactyla, *Mus* sp., *Rattus norvegicus* and *Nannospalax galili* for Rodentia, and *Canis lupus* for Carnivora). Due to a lack of species with annotated sequences (or even species with available genomes) the genes of *H. sapiens* were used as queries for the Perissodactyla and Lagomorpha orders. The queries were used in the option

protein2genome of the Exonerate v2.2 software [26] in order to retrieve non-annotated gene sequences and their chromosomic locations from both chromosome X and the whole genomes files (retrieved from NCBI databases) of several species from the above mentioned mammalian orders. Comparative synteny analyses were performed, using the synteny of the human PAR1 as reference.

Table 1 - Accession numbers of the different sequences used in the synteny analyses. X chromosomes were retrieved from the GenBank and Ensembl databases. Whole genomes files were retrieved from NCBI databases.

Order	Species	Accession number	Database
Primates	*Callithrix jacchus*	NC_013918.1	X chromosome GenBank
		WJHW01000052.1	whole genomes GenBank
		BJKT01000004.1	whole genomes GenBank
	Gorilla gorilla	NC_044625.1	X chromosome GenBank
		chromosome:gorGor4:X:1:156331669:1	X chromosome Ensemble
	Homo sapiens	NC_000023.11	X chromosome GenBank
	Macaca mulatta	NC_041774.1	X chromosome GenBank
	Microcebus murinus	NC_033692.1	X chromosome GenBank
	Pan troglodytes	NC_036902.1	X chromosome GenBank
	Theropithecus gelada	QGDE01000017.1	whole genomes GenBank
		Tgel_1.0:X:1:153451627:1	X chromosome Ensemble
Artiodactyla	*Balaenoptera musculus*	VNFC01000005.1	whole genomes GenBank
		VNFC01000001.1	whole genomes GenBank
	Bos indicus x Bos taurus	NC_040105.1	X chromosome GenBank
		primary_assembly:UOA_Brahman_1:X:1:146092946:1	X chromosome Ensemble
	Bubalus bubalis	NC_037569.1	X chromosome GenBank
	Camelus dromedarius	NC_044547.1	X chromosome GenBank
	Muntiacus reevesi	CM018500.1	X chromosome GenBank
		VCEB01000006.1	whole genomes GenBank
	Ovis aries	NC_040278.1	X chromosome GenBank
	Phocoena sinus	CM018177.1	X chromosome GenBank
		VOSU01000006.1	whole genomes GenBank
Carnivora	*Canis lupus familiaris*	NC_006621.3	X chromosome GenBank
	Felis catus	NC_018741.3	X chromosome GenBank
	Mustela erminea	WNLY01000049.1	whole genomes GenBank
	Neomonachus schauinslandi	NINY01007732.1	whole genomes GenBank
	Suricata suricatta	NC_043717.1	X chromosome GenBank
	Zalophus californianus	WPOA01000008.1	whole genomes GenBank
		UZVU01004191.1	whole genomes GenBank
Perissodactyla	*Diceros bicornis*	PVJY020067174.1	whole genomes GenBank
	Equus caballus	NC_009175.3	X chromosome GenBank
		primary_assembly:EquCab3.0:X:1:128206784:1	X chromosome Ensemble

Phylogenetic analyses

A concatenated alignment, encompassing all alignments performed for the 15 genes, was constructed using the FASconCAT v1.11 software[31].The concatenated alignment was submitted to the jModelTest2 software[32], which was used to calculate the likelihood of different nucleotide substitution models and to find the best fit model resorting to the corrected Akaike Information Criterion (AICc). The best fit model and the corresponding parameter adjustments were used for two distinct phylogenetic reconstructions. The first phylogeny was constructed based in the Maximum-Likelihood (ML) algorithm, using the IQ-TREE software[33] with 1000 ultrafast bootstrap replicates, while the second phylogeny was produced based in the Bayesian Inference (BI) algorithm, employing the MrBayes v3.2.6 software[34], performing 5,000,000 generations, a sample tree collection every 500 generations and a final burn-in corresponding to 25% of the sampled trees. These two phylogenies were produced aiming to get a phylogenetic point of view of the PAR1 genes evolution. Lastly, employing only the genera names, a third phylogenetic tree was produced resorting to the TimeTree webserver[35], in order to have the accepted mammalian topology, to be use as a comparison. The phylogenies were posteriorly edited in the Tree Of Life webserver[36].

Selection analyses

One of the processes that we used to examine the evolution of this genes was the employment of selection analyses, where three categories of evolution were tested. A pressure to a certain class of mutations is present in two of these categories, for positive selection the pressure is towards the promotion of new phenotypes and the fixation of beneficial genetic variations[37], in contrast to the purifying/negative selection in which the pressure is towards the conservation of the sequence, purging nonsynonymous mutations[37]. On the other side for the third category, the neutral evolution, the genetic changes are under a non-selective random genetic drift[38]. Moreover, all the different approaches applied in the selection analyses take into account the ω, the ratio of nonsynonymous to synonymous substitution rates[39]. In the case of $\omega = 1$, we are in the presence of neutral evolution. On other hand, when ω assumes a value > 1, positive selection is considered, while $\omega < 1$ is an indication of purifying/negative selection[39 40]. This concept was applied in different selection analyses, including site models, branch models and branch-site models, being the difference where ω is assumed to vary (either in the phylogeny branches or sequence sites)[39 41].

Selection analyses – Site models

Firstly, the Genetic Algorithm for Recombination Detection (GARD) analysis[42] was performed in the Datamonkey webserver[43], for the detection of putative recombination break points, which separate nonrecombinant fragments of the alignment with different evolutionary ratios, that may create discordant phylogenetic signals[42] [44]. Subsequently, each gene alignment was fractured at the obtained break points.

Thereafter, site models selection analyses were performed, in which the variation of the ω value along the sites of the sequence will be tested, pursuing the detection of sites under positive selection. The Datamonkey webserver was also used to perform adaptive selection analyses at the codon level. In this webserver, four tests were employed – Fixed Effects Likelihood (FEL) [45], Fast, Unconstrained Bayesian AppRoximation for Inferring Selection (FUBAR)[46], Mixed Effects Model of Evolution (MEME)[47] and Single-Likelihood Ancestor Counting (SLAC)[45]. These tests were performed aiming to detect positively selected sites (PSS). Only those PSS reported in three or more tests were considered.

Secondly, the CODEML program included in the Phylogenetic Analysis by Maximum Likelihood (PAML) v4 package[41] was used for site models analyses, also aiming to detect PSS. The likelihood of two nested site models was calculated, a null model that does not take in account codons under positive selection (M7), and an alternative model, in which positive selection is considered (M8). Based on the comparison between the likelihood of calculated models, it is possible to determine the most suitable model through a Likelihood Ratio Test (LRT = 2 x (lnL [Alternative model] - lnL [null model])). If the LRT is significant ($P < 0.05$) it indicates that the alternative (positive selection) model is the most suitable. In opposition, for the genes that the LRT is not significant ($P > 0.05$), the null model is accepted. For those cases in which the LRT is significant, the Bayes Empirical Bayes (BEB) method[48] was performed to identify potential PSS, being considered only those presenting a posterior probability (PP) value equal or higher than 0.95.

Selection analyses – Branch models

In order to test the possibility of significant differences between the ω of the several phylogeny branches, the CODEML program was also used for branch models analyses. The branches of different mammalian orders were analyzed employing two approaches: 1) in which the ω values were analyzed in the whole order clade and 2) in which the ω varies only in the ancestral branches of each mammalian order. Therefore, two types of labels were employed for each order present in each gene tree (Carnivora, Perissodactyla, Artiodactyla, Primates, Rodentia and Lagomorpha). In these analyses, two models were compared: a) the one-ratio model (M0), which admits that the ω is constant among all

branches of the phylogeny; and, b) the alternative model, a two-ratio model, which acknowledges the ω variation among different branches regarding the used label. In order to compare these two models, the LRTs were calculated and, for the genes with a significant LRT ($P < 0.05$), the alternative model was considered the most suitable. For the genes in which a ω value higher than 1 was achieved, the dN and dS were checked, and the inflated ω were only considered if none of these variables were equal to 0.

Selection analyses – Branch-Site models

Branch-site models are a combination of the two previous analyses (site and branch models), allowing the detection of ω variations along the tree branches and sites of the sequences. Two selection analyses were performed in the Datamonkey webserver, BUSTED (Branch-site Unrestricted Statistical Test for Episodic Diversification)[49] and FEL[45]. Firstly, BUSTED was used in each gene, in order to examine if the selected order, was subjected to positive selection. For the genes in which positive selection was considered, FEL was employed to determine the PSS in the selected order.

Additionally, similarly as for the previous selection analyses, the CODEML program was used for branch-site models selection analysis. A neutral branch-site model that fixes the ω value ($\omega = 1$) is compared to an alternative model, that allows ω variations. Finally, the LRT were calculated and for those genes in which the LRT was significant ($P < 0.05$), the alternative model that takes in account positive selection was considered. The BEB method was performed in the genes were positive selection was acknowledged, in order to identify potential PSS, although only those displaying a PP value equal or higher than 0.95 were considered.

Results

Sequence collection and saturation assessment

The multiple sequence alignments were performed using 526 sequences relative to 15 genes of 41 different genera, retrieved from the GenBank and Ensembl databases and publicly available genomes.

The results of the nucleotide saturation assessment are showed in table 2. Among the 15 analyzed genes, 11 presented low and not significant saturation signatures (*AKAP17A*, *ASMT*, *ASMTL*, *DHRSX*, *GTPBP6*, *P2RY8*, *PLCXD1*, *PPP2R3B*, *SHOX*, *SLC25A6*, *ZBED1*), while the *CD99*, *CRLF2* and *CSF2RA* genes showed high saturation signatures (significant for the last two). Also, in the case of the *IL3RA* gene, no saturation signature was achieved.

Table 2 - Saturation signatures. Results in bold are statistically significant. Underlined genes have low and not significant saturation signatures. Values from the test of substitution saturation (Xia´s test): Iss - index of substitution saturation, Iss c - critical index of substitution saturation, DF - degrees of freedom and *P* - *P*-value. NA – Non-applicable.

	P (inv)	Iss	Iss c	DF	*P*
AKAP17A	0.130	0.355	0.473	1114	**0.000**
ASMT	0.100	0.413	0.458	752	**0.019**
ASMTL	0.086	0.391	0.440	895	**0.005**
CD99	0.146	0.406	0.359	173	0.257
CRLF2	0.136	0.498	0.409	663	**0.000**
CSF2RA	0.051	0.501	0.426	838	**0.000**
DHRSX	0.111	0.354	0.424	773	**0.000**
GTPBP6	0.113	0.325	0.464	1047	**0.000**
IL3RA	NA	NA	NA	NA	NA
P2RY8	0.099	0.268	0.441	894	**0.000**
PLCXD1	0.114	0.308	0.427	789	**0.000**
PPP2R3B	0.111	0.309	0.459	1009	**0.000**
SHOX	0.211	0.259	0.409	605	**0.000**
SLC25A6	0.248	0.168	0.413	599	**0.000**
ZBED1	0.162	0.281	0.520	1689	**0.000**

Synteny analyses

The synteny analyses were performed in 18 mammalian species from four distinct orders (Primates, Artiodactyla, Carnivora and Perissodactyla), in which the synteny of the human PAR1 was used as the reference (figure 3, figure S1).

In Primates, the synteny was maintained in *Gorilla gorilla* and *Theropithecus gelada,* even though the PAR1 region of *T. gelada* was not located in the beginning of the X chromosome as expected. The start of the PAR1 region of *Macaca mulatta* was very distinct compared to that of the human. Here, five genes translocated into different positions (*SHOX*, *CRLF2*, *PPP2R3B*, *GTPBP6* and *PLCXD1*) and three were not found (*CSF2RA*, *IL3RA* and *SLC25A6*). This result was concordant with those of previous studies that found higher divergence in this species when compared to other primates, including the observed intrachromosomal rearrangements[50]. Another difference that was observed in this species refers to the overlapping of the genes *ZBED1* and *DHRSX*[51], an occurrence also displayed by other species, such as *Pan troglodytes, Bos indicus x Bos Taurus*, *Ovis aries, Balaenoptera musculus* and *Diceros bicornis*. Another noticeable result was that of *Microcebus murinus.* It presented the more dissimilar synteny, probably relatable to the fact that this species (from the Strepsirrhini suborder) is more phylogenetically distant in comparison to the rest of the Primates assessed in this study, which belong to the Haplorrhini suborder. A subject for future analyses may be the assessment whether these alterations are either a trait of all basal primates or if *M. murinus* is a divergent case, thus being this alteration specific to this species.

Secondly, considering the Artiodactyla order, apart from the *ZBED1/DHRSX* overlapping referred previously (exhibited by *Bos indicus x Bos Taurus* and *Ovis aries*), one gene seems to be absent (*CSF2RA*) in *Camelus dromedarius.* Additionally, all the genes of *Ovis aries* had their position altered, which could indicate that the region was being misread backwards. However, the dissimilar synteny is supported by two arguments: a) the position of the genes that were supposed to be after this region, but are located in the middle of PAR1 genes, and b) the presence of genes after *PLCXD1*, where if the PAR1 was in the typical position, no genes should be present. Due to these arguments, the dissimilar synteny was accepted.

Thirdly, regarding the Carnivora species, the more dissimilar synteny was found in *Canis lupus*, presenting seven genes with changed loci (*PPP2R3B*, *GTPBP6, PLCXD1, ASMT, AKAP17A* and *P2RY8)*. In this order, the only other species with genes with altered locations was the *Felis catus,* in which *PLCXD1* was separated from the other PAR1 genes, at the end of the X chromosome. Apart from that, only 2 species had gene absence, a)

SHOX in *Suricata suricatta* and b) *PLCXD1*, *GTPBP6* and *ASMTL* in *Neomonachus schauinslandi.*

Finally, concerning the Perissodactyla order, one gene in *Equus caballus* was absent (*IL3RA*) and two genes had altered locations (*ASMTL* and *CSF2RA*). Additionally, based on its position, the whole PAR1 of *Diceros bicornis* was similar to the human PAR1.

The synteny analyses were not performed for the Rodentia, due to the unusual chromosomic location of the PAR1 genes. According to the locations present in the NCBI database, the genes are spread in the autosomes, instead of a cluster in the sexual chromosomes, including *GTPBP6* (chromosome 5 in *Mus musculus*, chromosome 12 in *Rattus norvegicus* and chromosome 19 in *Microtus ochrogaster*), *CSF2RA* (chromosome 19 in *M. musculus* and chromosome 14 in *R. norvegicus*) or *PPP2R3B* (chromosome 14 in *R. norvegicus* and chromosome 6 in *M. ochrogaster*). Although some rare cases of genes are still located in the X chromosome, such as *ASMT* in the *M. musculus*, in *R. norvegicus* this gene is in the autosomes (chromosome 12). This dissimilarity indicates that high syntenic divergence occurs even within the same order, as suggested previously[11].

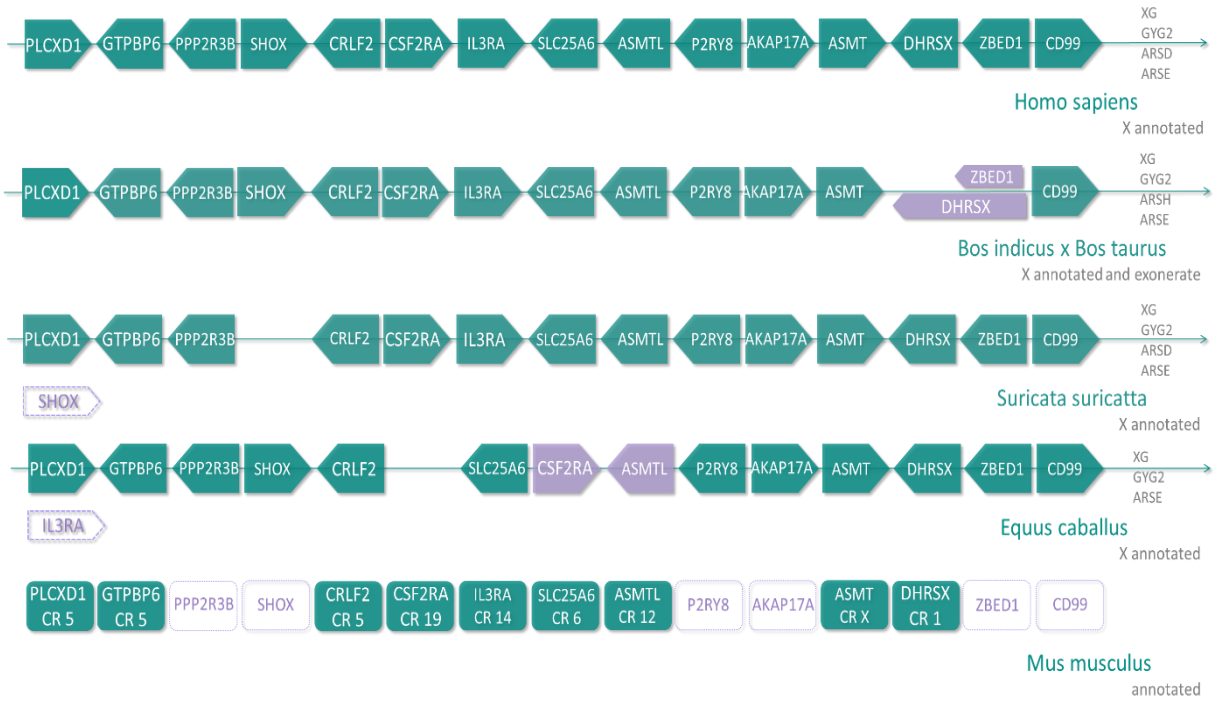

Figure 3 – Results of the synteny analyses. Genes in blue had maintained the position, genes in purple had their location changed. Genes in a blank background box under the arrow were missing from the assessed species. Genome position of the genes present in *Mus musculus* (in blue, chromosome number of each gene described under gene name).

Phylogenetic analyses

The best fit model was calculated using the concatenated alignment of the 15 genes. According to the AICc, the General Time Reversible + I + G was the most suitable model. Using the obtained parameters, two phylogenies were produced, employing both the ML and the BI algorithms. The produced phylogenies were very similar between them in terms of branch length, but with differences when referring to the obtained topology. Some incongruences were found. For example, the BI phylogeny presented *Microcebus* in the Rodentia clade instead of that of the Primates (figure S1). Additionally, both phylogenetic trees were very dissimilar from the accepted mammalian topology, *i.e.* the phylogenetic tree produced in the TimeTree webserver (figure 4). The two produced phylogenies (ML and BI) failed to present monophyletic clades for each mammalian order and to illustrate the accepted mammalian evolution, as seen in the Time tree phylogeny (figure 4B). Instead, they presented numerous clades for each order and the basal division was absent, in contrast to the two basal mammalian partitions, forming a clade with Primates, Rodentia and Lagomorpha, and another one with Artiodactyla, Perissodactyla and Carnivora. Also, regarding the branch length, all the Rodentia taxa presented higher branch lengths when compared to the rest of the taxa in the produced phylogenies.

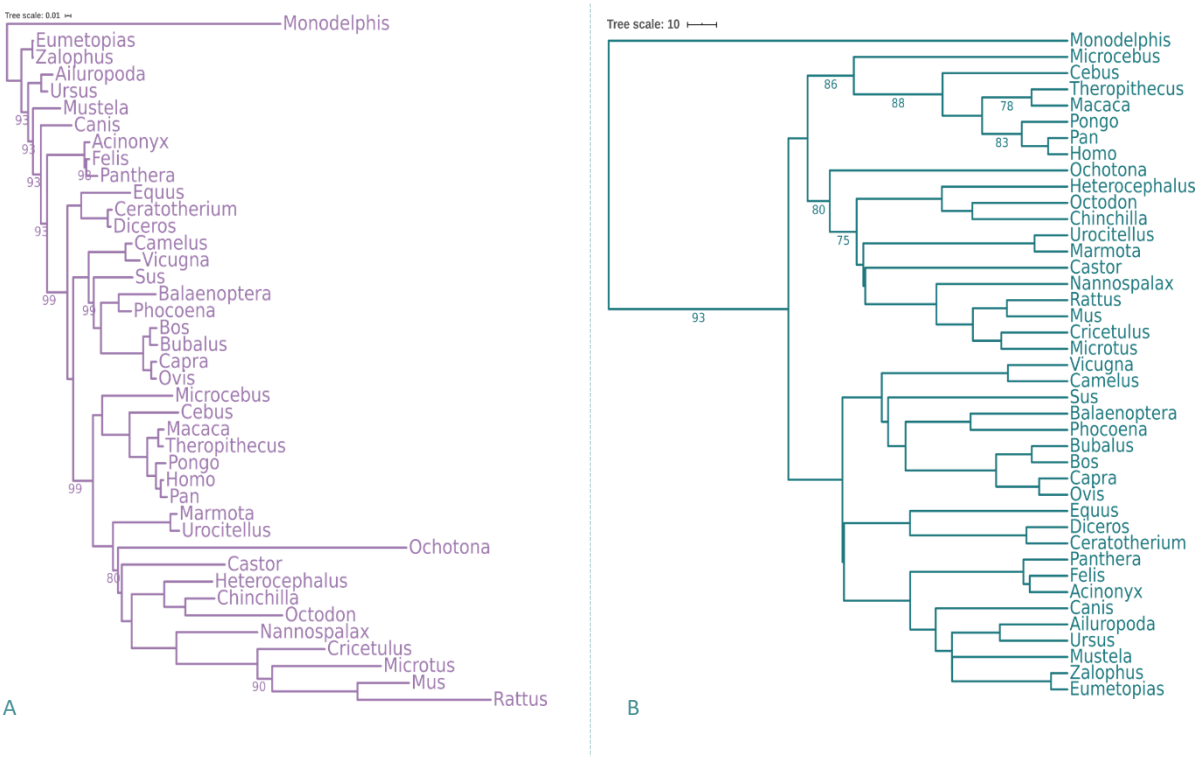

Figure 4 – Phylogenetic trees constructed employing: A) the ML algorithm; and, B) the TimeTree webserver. Values at each node correspond to branch support (only values >70 are shown).

Selection analyses – Site models

The alignment for each gene was fragmented, according to the results obtained in the GARD analysis (table S2) and then submitted to four selection tests incorporated in the Datamonkey webserver, as shown in table 3. Only were considered the PSS detected in, at least, three different tests. A total of 79 PSS distributed throughout 15 genes were detected along the four tests (table S2). However, only four of those sites fulfilled the previous requirement (table 3). One PSS was located in the *ASMT* gene (position 515), another was located in the *PLCXD1* gene (position 286), and two PSS were located in the *ZBED1* gene (positions 986 and 1142).

Additionally, selection analyses were performed employing the CODEML program, also aiming to ascertain the previous selection results. These analyses were done using both the ML phylogeny and the mammalian accepted phylogeny from the TimeTree webserver. The ML phylogeny was chosen instead of the BI phylogeny due to its higher similarity to the mammalian accepted phylogeny, considering the monophyletic order clades. In order to avoid the negative influence in the analyses of the differences between the ML and the accepted mammalian phylogeny (for example, lack of monophyletic clades and higher branch length in Rodentia, as referred before) only the results from the Time Tree phylogeny were taken into consideration, since only this topology truly reflects the phylogenetic relation between orders. Unlike what was expected, no PSS was detected employing the TimeTree phylogenetic tree in the CODEML program.

Table 3 – Results of adaptive selection analyses at the codon level performed at Datamonkey webserver. Sites in bold have a significant *P*/PP.

	FUBAR		SLAC		MEME		FEL	
	site	PP	site	*P*	site	*P*	site	*P*
ASMT	**515**	0.00	515	0.15	**515**	0.05	**515**	0.03
PLCXD1	**286**	0.01	**286**	0.02	**286**	0.03	**286**	0.02
ZBED1	**986**	0.00	**986**	0.01	**986**	0.00	**986**	0.02
	1142	0.01	**1142**	0.03	**1142**	0.04	**1142**	0.03

Selection analyses – Branch models

All orders in study were tested to ascertain the possibility of differential selection occurring in the distinct branches. These analyses were performed only using the mammalian accepted phylogeny from the TimeTree webserver. For the genes with a significant LRT, the alternative model was accepted, in which the ω for the whole tree is not the same but differs for each marked branch to branch.

Two approaches were employed in this phylogeny, as previously mentioned, one in which the whole clade is selected and another where the *ω* is able to vary only in the ancestral branches of each order. A significant LRT ($P < 0.05$) was detected in *AKAP17A*, *ASMT*, *ASMTL*, *CRLF2*, *CSF2RA, PPP2R3B* and *ZBED1,* when selecting the ancestral branches (table 4). Consequently, in these genes the alternative model is accepted as the most suitable, thus allowing the *ω* value to vary among the orders of the phylogenetic tree for the selected branches, in contrast to the genes with a not significant LRT, in which is considered that the *ω* value is equal in all branches of the phylogenetic tree. Additionally, as shown in the table 5, a higher number of genes are more suitable for the alternative model, 12 genes, when selecting the whole clade, instead of 7 genes selecting the ancestral branches. On other side, although in table 5 a higher number of genes has a significant *P*, only selecting the ancestral branches an *ω* value higher than 1 is accomplished, such as for Primates in *CRLF2* and *CSF2RA* and also for Rodentia in *ASMT* and *ZBED1*. In the orders with a *ω* value higher than 1, positive selection can be considered. Essentially, only two orders present a *ω* value superior to 1, Rodentia and Primates, that were selected for further analysis.

Table 4 – Branch models analysis results selecting the ancestral branches using the TimeTree phylogeny

	P	*ω*						
		Background	Carnivora	Perissodactyla	Artiodactyla	Primates	Rodentia	Lagomorpha
AKAP17A	**0.00**	0.04	0.02	0.01	0.00	0.01	0.00	0.00
ASMT	**0.00**	0.17	0.26	0.07	0.10	0.26	**21.43**	0.01
ASMTL	**0.00**	0.17	0.07	0.69	0.14	0.13	0.06	0.00
CD99	0.16	0.13						
CRLF2	**0.01**	0.23	0.17	0.09	0.27	**796.64**	0.03	0.13
CSF2RA	**0.03**	0.21	0.29	0.28	0.40	**999.00**	0.40	
DHRSX	0.42	0.12						
GTPBP6	0.60	0.11						
IL3RA	0.60	0.22						
P2RY8	0.11	0.07						
PLCXD1	0.45	0.09						
PPP2R3B	**0.00**	0.05	0.00	0.00	0.02	0.00	0.00	0.01
SHOX	0.33	0.04						
SLC25A6	0.40	0.01						
ZBED1	**0.03**	0.05	0.04	0.02	0.05	0.12	**38.78**	0.08

Table 5 – Branch models analysis results selecting the whole clade using the TimeTree phylogeny

	P	*ω*						
		Background	Carnivora	Perissodactyla	Artiodactyla	Primates	Rodentia	Lagomorpha
AKAP17A	**0.00**	0.02	0.02	0.02	0.05	0.02	0.04	0.00
ASMT	**0.00**	0.24	0.16	0.14	0.18	0.29	0.15	0.01
ASMTL	**0.00**	0.12	0.19	0.28	0.17	0.17	0.12	0.00
CD99	**0.01**	0.01	0.12	0.03	0.15	0.21	0.10	
CRLF2	**0.01**	0.13	0.24	0.15	0.27	0.23	0.22	0.16
CSF2RA	**0.00**	0.29	0.32	0.13	0.24	0.17	0.22	
DHRSX	0.42	0.07						
GTPBP6	**0.00**	0.09	0.16	0.07	0.12	0.08	0.11	
IL3RA	0.22	0.13						
P2RY8	**0.00**	0.21	0.07	0.02	0.08	0.06	0.09	0.10
PLCXD1	**0.05**	0.08	0.09	0.13	0.11	0.07	0.08	
PPP2R3B	**0.00**	0.03	0.03	0.04	0.05	0.07	0.06	0.01
SHOX	**0.04**	0.00	0.06	0.02	0.04	0.03	0.03	
SLC25A6	0.21	0.00						
ZBED1	**0.00**	0.13	0.02	0.03	0.06	0.03	0.08	0.08

Box 1 – Results of the branch models analyses. LRT in bold are significant *(P* < 0.05).

Selection analyses – Branch-Site models

Lastly, branch-site models analyses were performed aiming to detect positively selected sites along the pre-selected branches. Differently from the previous analyses, the hypothesis of positive selection is analyzed not in all branches of the given phylogenetic tree, but only in the foreground selected branches, which focus more the analyses and can make them more robust and powerful[52 53]. These analyses were performed employing two approaches, one by selecting the Rodentia order and another by selecting the Primates order, since these two orders had shown signals of positive selection in the branch models.

Once again, the CODEML and the Datamonkey webserver were employed. Yet again, using CODEML program no significant results were acquired. On contrast, employing the Datamonkey tests the results were different for both selected orders. Two analyses were employed, BUSTED and FEL, although only when positive selection was detected in both, the results were considered. From the 15 genes tested, only two genes presented signs of positive selection when selecting the Primates (table 6) and 5 genes when the Rodentia were selected (table 7).

Referring to the Primates analyses, only *ASMTL* and *IL3RA* had evidences of positive selection, not concordant with the branch and site models analyses.

Prespecifying Rodentia, from the 5 genes with evidence of positive selection, only *ASMT* is in concordance with the branch models analyses and *ASMT* is also in agreement with site models, although the site is not the same.

Table 6 – Branch-site models analyses results employing the Datamonkey webserver selecting the Primates order

		Site	Partition	alpha	beta	omega	alpha=beta	LRT	*P*	Total branch length
ASMTL	BUSTED	found evidence (LRT, $P = 0.000 \leq .05$)								
	FEL	262	1	0.165	1.723	10.446	0.768	4.105	0.043	1.726
IL3RA	BUSTED	found evidence (LRT, $P = 0.000 \leq .05$)								
	FEL	211	1	0.212	2.423	11.409	0.935	5.053	0.025	2.34

Table 7 – Branch-site models analyses results employing the DataMonkey webserver selecting the Rodentia order

		Site	Partition	alpha	beta	omega	alpha=beta	LRT	*P*	Total branch length
ASMT	BUSTED	found evidence (LRT, $P = 0.001 \leq .05$)								
	FEL	75	1	0.117	1.085	9.31	0.464	5.061	0.024	3.243
		104	1	0	0.7	Infinity	0.462	3.949	0.047	2.024
		219	1	0	1.92	Infinity	1.236	3.938	0.047	5.556
CSF2RA	BUSTED	found evidence (LRT, $P = 0.000 \leq .05$)								
	FEL	278	1	0.331	4.295	12.982	1.209	11.099	0.001	20.552
		122	1	0	1.67	Infinity	0.96	9.31	0.002	7.785
		282	1	0	5.818	Infinity	1.792	8.738	0.003	27.12
		290	1	0	0.59	Infinity	0.32	5.381	0.02	2.752
		336	1	0	0.95	Infinity	0.613	4.245	0.039	4.426
IL3RA	BUSTED	found evidence (LRT, $P = 0.016 \leq .05$)								
	FEL	49	1	0	1.073	Infinity	0.765	5.856	0.016	6.162
		103	1	0.204	2.208	10.797	0.8	4.402	0.036	13.038
		61	1	0	0.756	Infinity	0.522	4.065	0.044	4.342
P2RY8	BUSTED	found evidence (LRT, $P = 0.016 \leq .05$)								
	FEL	663	1	0.183	7.9	43.064	0.796	7.838	0.005	11.794
		776	1	0	0.787	Infinity	0.228	4.594	0.032	1.17
		718	1	0.189	1.934	10.232	0.507	4.184	0.041	2.928
PPP2R3B	BUSTED	found evidence (LRT, $P = 0.000 \leq .05$)								
	FEL	590	1	0	11.743	Infinity	0.743	8.369	0.004	47.748
		147	1	0	0.579	Infinity	0.313	5.563	0.018	2.354
		391	1	0	0.884	Infinity	0.329	5.465	0.019	3.596
		665	1	0	1.664	Infinity	1.172	3.95	0.047	6.764

Discussion

Molecular Evolution of the PAR1 genes in mammals

Bearing in mind the referred properties and significance, pseudoautosomal regions are of great importance, as their absence is related with a lack of recombination in the sexual chromosomes, which may result in several genetic problems and diseases. Therefore, in this work, evolutionary genomics and bioinformatics analyses were performed in the genes of the PAR1, pursuing a better understanding of the evolution and the interspecific differences present in this region.

Firstly, saturation signatures were assessed, in which three genes presented high levels of saturation. For the *IL3RA*, no saturation signature was achieved due to the lack of a common site among all sequences of the alignment, which may be a consequence of, for example, the poor sequence quality, the incorrect identification of homologous sites by sequence alignment or even the sequence divergence[54]. The high saturation can be reduced excluding taxa more phylogenetic distant or by performing a concatenated alignment. The saturation occurs when a nucleotide substitution happen multiple times that the phylogenetic algorithm underestimates the number of substitutions that actually occurred, leading to an underestimation of the phylogenetic distances[54 55]. Consequently, if the sequence or alignment are larger, the effect of the sites with high saturation is reduced, since in a coding sequence not all sites display the same variability. Some sites are quite conserved and not mutable reducing the effect of the high saturated sites, being this the purpose of producing a concatenated alignment. Therefore, the saturation signatures should be decreased and a sequence with more sites to be taken in consideration by the phylogenetic algorithm for the calculation of the phylogenetic distance is obtained, resulting in a more accurate and robust phylogeny.

Regarding the synteny analyses, the human PAR1 reference was overall maintained along the species analyzed. Although, some differences were observed, such as some absent genes, the *ZBED1/DHRSX* overlapping, or gene position alterations, as described previously. The genes absent in the PAR1 and X chromosome, could be due to an actual absence of the gene in the species genome or to a translocation to another chromosome, which was inspected through a search in the NCBI database. A chromosome translocation was not found for any of the absent genes, which is a subject for future analyses. None of the synteny alterations, seem to have a phylogenetic correlation. These modifications appear to be lineage specific. For example, in Primates, the *ZBED1/DHRSX* overlapping occurred in two species (*M. mulata* and *P. troglodytes*) phylogenetically distinct, corresponding to two different branches, Cercopithecidae and Hominidae respectively. Since it is not an ancestral trait, a scenario of convergent evolution may have occurred. The

order with the more dissimilar synteny was present in Rodentia, where the genes were spread along the autosomal and sexual chromosomes. There is a high intra order variation of PAR1 genes considering their chromosomic location, which suggests that PAR1 in rodents was subjected to a rapid evolution, like it was referred in previous studies[11].

Subsequently, the rapid evolution in Rodentia was again supported by the higher branch lengths obtained in the produced ML and BI phylogenies, which was observed exclusively in this order. Relevant differences between the gene trees (ML and BI) and the species tree (mammalian accepted topology from the TimeTree webserver) were found. Theoretically, phylogenetic trees produced based on gene sequences can differ from the species accepted topology due to nucleotide or amino acid substitution being subject to stochastic errors ("estimation error in a model that arises from the exclusion of an important explanatory variable or due to incorrect specification of the relationships being examined"[56]), being affected by sampling errors of polymorphic alleles that existed in the ancestral populations[57 58 59 60] or if there are two or more copies of the same gene in the genome[61 62]. Furthermore, this divergence between the gene tree and the species accepted topology, can be also due to the divergent gene evolution. For example, one thing that differs in the evolution of different genes is the rate of nucleotide substitution, which has a profound impact in the phylogenetic analyses[63]. Also, in particular to these sequences, the high saturation observed in some genes can affect the produced phylogenies. The high saturation levels are a signal of too divergent sequences, which reduces the phylogenetic information contained in the sequences[54], even though a concatenate alignment was used in order to avoid, or at least minimize the influence of saturation. Therefore, the differences between the produced phylogenies and the accepted mammalian tree are due to both the genetic data, but also due to the variables related with the used softwares.

Regarding the selection analyses, the results from the two used methods (CODEML and Datamonkey) were not concordant. The absence of PSS in the analyses performed with the CODEML program could be due to the episodic or transient nature of the natural selection being hard to be precisely identified[47 64], the positive selection on the evolution of protein-coding genes being not sufficiently strong to be detected[41 64], or the purifying selection in some lineages being masking the signal of positive selection in the others[47 64]. These possible explanations are a consequence of the lower sensitivity of the CODEML program in comparison with the Datamonkey tests, which uses different approaches[65].The four analyses from this webserver consist in: a) calculation of the evolution rate of each site without constrains based on an approximate hierarchical Bayesian method using a Markov chain Monte Carlo routine[46] (FUBAR); b) the Maximum Likelihood reconstructions of ancestral sequences are used to calculate the number of both categories of substitutions[64] (SLAC); c) usage of a LRT to detect individual sites subjected to episodic diversifying

selection and being able to identify sites in which not all branches are under selective pressure, the only Datamonkey software used capable of identifying both episodic and pervasive positive selection[43 65] (MEME) and, d) testing if the *ω* value is different from 1 of every site of the alignment[43] (FEL).

On the other side, concerning the branch models, higher *ω* values were obtained for Rodentia (*ASMT* and *ZBED1*) and Primates (*CRLF2* and *CSF2RA*), although only in the analyses where the ancestral nodes were selected. The explanation to this result can be in the episodic nature of selection, detected more easily in shorter periods of time, such as those considering only the ancestral nodes instead of including the whole clade. Furthermore, no significant differences between the orders were found considering the whole results.

On the other hand, genes displaying evidences of positive selection and presenting PSS were found in both Rodentia and Primates. Although, only a few genes with indications of positive selection were in concordance between the several selection analyses. The *ASMT* and *ZBED1* genes are common to site and branch models regarding the Rodentia order. However, when considering all the selection analyses performed, only the *ASMT* gene had evidences of positive selection in all of them, even though only concerning Rodentia.

Functional and Medical Importance of the PAR1 genes

PAR1 genes display several different functions, which are associated with various diseases (table 8), such as male infertility, increased probability of generating embryos with XY aneuploidy, stature problems (Klinefelter Syndrome), leukaemia, mental diseases[11], among others. Essentially, the lack of strong positive selection can also be an evidence of the functional constraints of these genes. According to some authors, essential or less dispensable proteins are more evolutionary conserved[66], thus evolving slower[67]. The stronger selective constraints, which force the evolutionary rates to decrease, in functionally important residues and sequences are define as a functional constraint, an event that can explain the absence of PSS in the analysed PAR1 genes[68 69]. There are some major factors that are thought to influence the occurrence of a functional constraint: a) the number of different proteins a given protein interacts with[69], meaning the more proteins it connects with, the more proteins that will be affected by the consequences of a possible mutation; b) the local recombination intensity[69 70], explained by the DNA repair function of recombination, so with more recombination comes a higher number of checkpoints to repair occurring mutations; c) ratio of amino acids that are critical to the protein function[67], therefore, the higher the proportion of the sequence that is critical to fulfil the protein role, the less proportion of the sequence that can be changed without damaging the protein function, and so the evolutionary rates are forced to decrease; and d) the protein abundance, meaning

how broadly is the protein expressed, since the more tissues in which the protein is expressed, the broader it would be the repercussions of a possible mutation, so the slower it evolves[69 70].

Positive selection evidences in Euarchontoglires

Overall, evidences of positive selection were found exclusively for Primates and Rodentia. These two orders are phylogenetic close, composing the superorder Euarchontoglires. Consequently, we can suspect that in this superorder a positive selection predisposition is present for some genes of the PAR1, explaining the evidences of positive selection being restricted to Primates and Rodentia, contrarily to any other assessed mammalian order. Moreover, these two orders have also in common their distinctive PARs, when compared to other Eutherians. The PAR1 of Simian primates/humans was substantially reduced (3-5 fold as referred previously[11]), being 2.7 Mb long in contrast to 5-9 Mb in Ruminants or 6.6 Mb in Carnivores. Furthermore, the Rodentia has the smallest PAR1 of the Eutherian, a 0.7 Mb long pseudoautosomal region. Notwithstanding, Rodentia species exhibited more evidences of divergence than Primates species, including synteny alteration (most of the PAR1 genes were translocated into the autosomes, a feature not found in any other mammalian order), increased branch length (suggesting an increased mutation rate) and more genes exhibiting positive selection signatures. Since the PAR1 is a conserved region, with a high rate of recombination, mechanism that promotes the DNA repair, when these genes are translocated to the autosomes, just as what happened with Rodentia PAR1 genes, the selective pressures may have changed. Subsequently, the possibility of nucleotide change and positive selection to occur can be increased, explaining the higher number of positive selection evidences in Rodentia, when comparing to the other order of Euarchontoglires. In contrast to *PLCXD1*, *GTPBP6*, *PP2R3B* and *DHRSX* which are located in the autosomes in *M. musculus* and *R. norvegicus*, the *ASMT*, showing positive selection evidences, is located in the X chromosome for *M. musculus*, and could be conflicting with our hypothesis. However, as stated previously, a high syntenic divergence is present in this order, therefore *ASMT* is in the autosomes for *R. norvegicus* (chromosome 12) and *Rattus rattus* (chromosome 16)[24], being in agreement with the proposed hypothesis.

Concluding, we can hypothesise that Euarchontoglires could have a positive selection predisposition in some PAR1 genes and that chromosomic migration towards non-PAR1 locations can have an effect in the selection pressures.

Table 8 – PAR1 genes function and associated diseases

	Category[71]	Definition	Function	Associated diseases
AKAP17A	Uncharacterized	protein kinase A anchoring protein[71]	highly expressed in T cells, B cells, and dendritic cells; regulator of alternative splicing[71]	Null-Cell Leukemia and Chronic Tic Disorder[72]
ASMT	Tumor suppressors	final reaction in the synthesis of melatonin[71]	circadian rhythms[73]	induces cytotoxicity in hemoresistant leukemia cell lines[71 74 75]; Pineoblastoma and Pineocytoma[72]
ASMTL	Uncharacterized	domain similar to bacterial protein and another one identical to *ASMT*[71 72]	dual role in cell division arrest and in preventing the incorporation of modified nucleotides into cellular nucleic acids[72]	supposedly related with the occurrence of acute lymphoid leukemia[71 76]; Melanotic Neurilemmoma and Chronic Tic disorder[72]
CD99	Context-Dependent	E2 antigen[71]	glycoprotein that is expressed across the hematopoietic system[71]	upregulated in acute lymphoblastic leukemias; haploinsufficiency could perturb normal T cell apoptosis[71]; Extraosseous Ewing Sarcoma and Ewing Sarcoma[72]
CRLF2	Proto-oncogenes	receptor for thymic stromal lymphopoietin[71]	pathways controlling cell proliferation and development of the hematopoietic system[72]	overexpression associated with acute lymphoblastic leukemia[71]
CSF2RA	Context-Dependent	subunit of a receptor associated to the colony stimulating factor 2[71]	controls the production, differentiation and function of granulocytes and macrophages[72]	overexpression associated chronic myelomonocytic leukemia and juvenile myelomonocytic leucemia; loss supposedly connected with leukemogenesis[71] Surfactant Metabolism Dysfunction Pulmonary 4 and Hereditary Pulmonary Alveolar Proteinosis[72]
DHRSX	Tumor suppressors	oxidoreductase enzyme family[71]	positive regulation of starvation-induced autophagy[72]	ectopic expression increased starvation-induced autophagy; deletion or loss reduce autophagy; Partington X-Linked Mental Retardation Syndrome[72]
GTPBP6	Uncharacterized	putative GTP-binding protein[71]	regulate hematopoietic stem cell functions and differentiation[71]	hematologic diseases; Mongolian Spot and Sengers Syndrome[72]
IL3RA	Proto-oncogenes	receptor[71]	highly expressed in monocytes, macrophages, alveolar macrophages, and dendritic cells[71 77]	hematological malignancies, such as myelodysplastic syndrome and Hodgkin's lymphoma[77] Hairy Cell Leukemia and Diphtheria[72]
P2RY8	Context-Dependent	the G-protein coupled receptor family[71]	highly expressed in hematopoietic cells[71]	overexpression associated to leukemias and if absent linked to Burkitt's and other lymphomas[71]; Childhood B-Cell Acute Lymphoblastic Leukemia and Mental Retardation[72]
PLCXD1	Uncharacterized	receptor-regulated phosphodiesterases[71]	regulating the cytosolic calcium and/or the activity of several protein kinases[78]	Mongolian Spot[72]
PPP2R3B	Tumor suppressors	regulatory subunit of a Ser/Thr phosphatase; protein substrate selectivity[72]	negative control of cell growth and division; wields control on the DNA replication[79]	depletion increase in proliferation of tumor cells[71] Alzheimer Disease 15 and Adjustment Disorder[72]
SHOX	Tumor suppressors	Short stature homeobox[71]	bone growth and maturation[71]	idiopathic growth retardation, Turner syndrome; Delections and mutations in Hodgkin's lymphoma, mantle cell lymphoma and Leri-Weill dyschondrosteosis[71]
SLC25A6	Tumor suppressors	encode a component of the mitochondrial permeability transition pore[71]	regulating cellular energy metabolism and apoptosis[71]	downregulated in leukemia cell lines[71]; Influenza and Bubonic Plague[72]
ZBED1	Uncharacterized	encodes regulatory proteins with diverse functions[79]	cell proliferation; stimulates transcription[80]	Laryngostenosis and Fibrosclerosis Of Breast[72]

Conclusions

Although PAR1 is considered to be very conserved, some differences were found in this region among mammals. Several syntenic differences were found, such as the overlapping of the genes ZBED1 and DHRSX in *Pan troglodytes* and *Bos indicus x Bos Taurus*, and the high divergence in *Macaca mulatta* from other primates, illustrating the interspecific divergence observed in this region. Additionally, a high divergence in chromosomic locations was observed in Rodentia, order in which the PAR1 genes were mostly located in the autosomes. Differences between the produced phylogenies (ML and BI) and the accepted mammalian topology were found. The most evident modification was the higher branch length in Rodentia, supposedly linked to the chromosomic translocation observed in this order. Regarding the selection analyses, evidences of positive selection were detected in all models, although exclusively for Primates and Rodentia, when regarding the orders analyzed. The lack of strong positive selection evidences can be associated to the high importance of the PAR1 genes, since essential genes are supposedly more evolutionary conserved. Based in the selection analyses results, we hypothesize that the superorder Euarchontoglires, in which Primates and Rodentia are included, may have a predisposition to positive selection for some genes of the PAR1, justifying the exclusivity of positive selection evidences for these orders. Moreover, based in the autosome location and on the stronger evidences of positive selection obtained for Rodentia, we can also hypothesise that genomic migration may modulate the evolution of PAR1 genes, possibly increasing the positive selection in this order when compared with Primates, explaining the different selection results for the two orders of the same superorder as Primates.

Acknowledgements

I would like to thank professor Agostinho Antunes for the opportunity and the whole EGB group for all the help and knowledge shared, especially Tito Mendes for all the help during the whole process. Also thankful for the support by the BYT+ program and my master's degree director, Dr. Maria José Pinto da Costa. This project was supported in part by the Strategic Funding UIDB/04423/2020 and UIDP/04423/2020 through national funds provided by FCT and the European Regional Development Fund (ERDF) in the framework of the program PT2020, by the European Structural and Investment Funds (ESIF) through the Competitiveness and Internationalization Operational Program - COMPETE 2020 and by National Funds through the FCT under the project PTDC/CTA-AMB/31774/2017 (POCI-01-0145-FEDER/031774/2017).

References

1. Bergero, R., and Charlesworth, D. (2009). The evolution of restricted recombination in sex chromosomes. Trends Ecol. Evol. *24*, 94–102.

2. Ross, L., and Blackmon, H. (2016). Sex Determination. In Encyclopedia of Evolutionary Biology, p.

3. Bachtrog, D., Mank, J.E., Peichel, C.L., Kirkpatrick, M., Otto, S.P., Ashman, T.L., Hahn, M.W., Kitano, J., Mayrose, I., Ming, R., et al. (2014). Sex Determination: Why So Many Ways of Doing It? PLoS Biol.

4. Schwanz, L.E. (2015). The Evolution of Sex Determination . By Leo W. Beukeboom and Nicolas Perrin. Oxford and New York: Oxford University Press. $89.95. xvi + 222 p. + 2 pl.; ill.; taxonomic, author, and subject indexes. ISBN: 978-0-19-965714-8. 2014. . Q. Rev. Biol.

5. Fricke, H.W. (1977). Community structure, social organization and ecological requirements of coral reef fish (Pomacentridae). Helgoländer Wissenschaftliche Meeresuntersuchungen.

6. Pough, F.H. (2013). Reptiles, Biodiversity of. In Encyclopedia of Biodiversity: Second Edition, p.

7. Pough, F.H. (2017). Reptiles, Biodiversity of ☆. In Reference Module in Life Sciences, p.

8. Hale, M.D., Cloy-McCoy, J.A., Doheny, B.M., and Parrott, B.B. (2018). Reproductive biology of crocodilians. In Encyclopedia of Reproduction, p.

9. Cotter, D.J., Brotman, S.M., and Wilson Sayres, M.A. (2015). Genetic diversity on the human X chromosome does not support a strict pseudoautosomal boundary. BioRxiv 33936.

10. Mensah, M.A., Hestand, M.S., Larmuseau, M.H.D., Isrie, M., Vanderheyden, N., Declercq, M., Souche, E.L., Van Houdt, J., Stoeva, R., Van Esch, H., et al. (2014). Pseudoautosomal Region 1 Length Polymorphism in the Human Population. PLOS Genet. *10*, e1004578.

11. Raudsepp, T., and Chowdhary, B. (2016). The Eutherian Pseudoautosomal Region. Cytogenet. Genome Res. *147*,.

12. Pinheiro, I., and Heard, E. (2017). X chromosome inactivation: new players in the initiation of gene silencing. F1000Research *6*, F1000 Faculty Rev-344.

13. Otto, S.P., Pannell, J.R., Peichel, C.L., Ashman, T.L., Charlesworth, D., Chippindale, A.K., Delph, L.F., Guerrero, R.F., Scarpino, S. V., and McAllister, B.F. (2011). About PAR: The distinct evolutionary dynamics of the pseudoautosomal region. Trends Genet.

14. Graves, J.A.M. (2016). Did sex chromosome turnover promote divergence of the major

mammal groups? BioEssays.

15. Deakin, J.E., Graves, J.A.M., and Rens, W. (2012). The evolution of marsupial and monotreme chromosomes. Cytogenet. Genome Res.

16. Page, J., Viera, A., Parra, M.T., De La Fuente, R., Suja, J.Á., Prieto, I., Barbero, J.L., Rufas, J.S., Berríos, S., and Ferréndez-Donoso, R. (2006). Involvement of synaptonemal complex proteins in sex chromosome segregation during marsupial male meiosis. PLoS Genet.

17. Page, J., Berríos, S., Parra, M.T., Viera, A., Suja, J.Á., Prieto, I., Barbero, J.L., Rufas, J.S., and Fernández-Donoso, R. (2005). The program of sex chromosome pairing in meiosis is highly conserved across marsupial species: Implications for sex chromosome evolution. Genetics.

18. White, M.A., Ikeda, A., and Payseur, B.A. (2012). A pronounced evolutionary shift of the pseudoautosomal region boundary in house mice. Mamm. Genome.

19. Perry, J., Short, K.M., Romer, J.T., Swift, S., Cox, T.C., and Ashworth, A. (1999). FXY2/MID2, a gene related to the X-linked Opitz syndrome gene FXY/MID1, maps to Xq22 and encodes a FNIII domain-containing protein that associates with microtubules. Genomics.

20. Shearn, R., Lecompte, E., Régis, C., Mousset, S., Penel, S., Douay, G., Crouau-Roy, B., and Marais, G.A.B. (2018). Contrasted sex chromosome evolution in primates with and without sexual dimorphism. BioRxiv 445072.

21. Wheatcroft, R. (2002). Fundamentals of Molecular Evolution . Second Edition. By Dan Graur and , Wen-Hsiung Li. Sunderland (Massachusetts): Sinauer Associates . $48.95 (paper). xiv + 481 p; ill.; subject and taxonomic indexes. ISBN: 0–87893–266–6. 2000. DNA Technology: The Aw. Q. Rev. Biol.

22. Miyata, T., and Yasunaga, T. (1980). Molecular evolution of mRNA: A method for estimating evolutionary rates of synonymous and amino acid substitutions from homologous nucleotide sequences and its application. J. Mol. Evol.

23. Stern, A., Doron-Faigenboim, A., Erez, E., Martz, E., Bacharach, E., and Pupko, T. (2007). Selecton 2007: Advanced models for detecting positive and purifying selection using a Bayesian inference approach. Nucleic Acids Res.

24. NCBI Resource Coordinators (2017). Database Resources of the National Center for Biotechnology Information. Nucleic Acids Res. *45*, D12–D17.

25. Flicek, P., Amode, M.R., Barrell, D., Beal, K., Brent, S., Chen, Y., Clapham, P., Coates, G., Fairley, S., Fitzgerald, S., et al. (2011). Ensembl 2011. Nucleic Acids Res.

26. Slater, G.S.C., and Birney, E. (2005). Automated generation of heuristics for biological

sequence comparison. BMC Bioinformatics *6*, 31.

27. Sela, I., Ashkenazy, H., Katoh, K., and Pupko, T. (2015). GUIDANCE2: Accurate detection of unreliable alignment regions accounting for the uncertainty of multiple parameters. Nucleic Acids Res.

28. Xia, X., Xie, Z., Salemi, M., Chen, L., and Wang, Y. (2003). An index of substitution saturation and its application. Mol. Phylogenet. Evol.

29. Xia, X. (2013). DAMBE5: A comprehensive software package for data analysis in molecular biology and evolution. Mol. Biol. Evol.

30. Ayerza, R. (2010). Bos indicus and bos indicus x bos taurus heifers' performance under two grazing systems in the arid chaco of Argentina. Livest. Res. Rural Dev.

31. Kück, P., and Meusemann, K. (2010). FASconCAT: Convenient handling of data matrices. Mol. Phylogenet. Evol.

32. Darriba, D., Taboada, G.L., Doallo, R., and Posada, D. (2012). JModelTest 2: More models, new heuristics and parallel computing. Nat. Methods.

33. Nguyen, L.T., Schmidt, H.A., Von Haeseler, A., and Minh, B.Q. (2015). IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol. Biol. Evol.

34. Ronquist, F., Teslenko, M., Van Der Mark, P., Ayres, D.L., Darling, A., Höhna, S., Larget, B., Liu, L., Suchard, M.A., and Huelsenbeck, J.P. (2012). Mrbayes 3.2: Efficient bayesian phylogenetic inference and model choice across a large model space. Syst. Biol.

35. Kumar, S., Stecher, G., Suleski, M., and Hedges, S.B. (2017). TimeTree: A Resource for Timelines, Timetrees, and Divergence Times. Mol. Biol. Evol.

36. Letunic, I., and Bork, P. (2019). Interactive Tree of Life (iTOL) v4: Recent updates and new developments. Nucleic Acids Res.

37. Choudhuri, S. (2014). Fundamentals of Molecular Evolution. In Bioinformatics for Beginners, p.

38. Nevo, E. (2001). Genetic Diversity. In Encyclopedia of Biodiversity: Second Edition, p.

39. Gao, F., Chen, C., Arab, D.A., Du, Z., He, Y., and Ho, S.Y.W. (2019). EasyCodeML: A visual tool for analysis of selection using CodeML. Ecol. Evol.

40. Nei, M., and Gojobori, T. (1986). Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol. Biol. Evol.

41. Yang, Z. (2007). PAML 4: Phylogenetic analysis by maximum likelihood. Mol. Biol. Evol.

42. Pond, S.L.K., Posada, D., Gravenor, M.B., Woelk, C.H., and Frost, S.D.W. (2006). Automated phylogenetic detection of recombination using a genetic algorithm. Mol. Biol.

Evol.

43. Weaver, S., Shank, S.D., Spielman, S.J., Li, M., Muse, S. V., and Kosakovsky Pond, S.L. (2018). Datamonkey 2.0: A modern web application for characterizing selective and other evolutionary processes. Mol. Biol. Evol.

44. Lanier, H.C., and Knowles, L.L. (2012). Is recombination a problem for species-tree analyses? Syst. Biol.

45. Kosakovsky Pond, S.L., and Frost, S.D.W. (2005). Not so different after all: A comparison of methods for detecting amino acid sites under selection. Mol. Biol. Evol.

46. Murrell, B., Moola, S., Mabona, A., Weighill, T., Sheward, D., Kosakovsky Pond, S.L., and Scheffler, K. (2013). FUBAR: A fast, unconstrained bayesian AppRoximation for inferring selection. Mol. Biol. Evol.

47. Murrell, B., Wertheim, J.O., Moola, S., Weighill, T., Scheffler, K., and Kosakovsky Pond, S.L. (2012). Detecting individual sites subject to episodic diversifying selection. PLoS Genet.

48. Yang, Z., Wong, W.S.W., and Nielsen, R. (2005). Bayes empirical Bayes inference of amino acid sites under positive selection. Mol. Biol. Evol.

49. Murrell, B., Weaver, S., Smith, M.D., Wertheim, J.O., Murrell, S., Aylward, A., Eren, K., Pollner, T., Martin, D.P., Smith, D.M., et al. (2015). Gene-wide identification of episodic selection. Mol. Biol. Evol.

50. Kim, J., Farré, M., Auvil, L., Capitanu, B., Larkin, D.M., Ma, J., and Lewin, H.A. (2017). Reconstruction and evolutionary history of eutherian chromosomes. Proc. Natl. Acad. Sci. *114*, E5379–E5388.

51. Gravholt, C. (2009). Turner - know your body. An information book on Turner syndrome. Editor Gravholt CH, Gothenburg, pp 1-241, 2009. http://np.netpublicator.com/netpublication/n75088268.

52. Yang, Z., and Dos Reis, M. (2011). Statistical properties of the branch-site test of positive selection. Mol. Biol. Evol.

53. Gharib, W.H., and Robinson-Rechavi, M. (2013). The branch-site test of positive selection is surprisingly robust but lacks power under synonymous substitution saturation and variation in GC. Mol. Biol. Evol.

54. Xia, X., and Lemey, P. (2012). Assessing substitution saturation with DAMBE. In The Phylogenetic Handbook, p.

55. Philippe, H., Brinkmann, H., Lavrov, D. V., Littlewood, D.T.J., Manuel, M., Wörheide, G., and Baurain, D. (2011). Resolving difficult phylogenetic questions: Why more sequences are not enough. PLoS Biol.

56. Moles, P., and Terry, N. (1997). The Handbook of International Financial Terms.

57. Nei, M., and Tajima, F. (1981). DNA polymorphism detectable by restriction endonucleases. Genetics.

58. Nei, M. (1986). Stochastic errors in DNA evolution and molecular phylogeny. Prog. Clin. Biol. Res.

59. Neigel, J.E., and Avise, J.C. (1986). Phylogenetic relationships of mitochondrial DNA under various demographic models of speciation. Evol. Process. Theory.

60. (1988). Relationships between gene trees and species trees. Mol. Biol. Evol.

61. Goodman, M., Romero-Herrera, A.E., Dene, H., Czelusniak, J., and Tashian, R.E. (1982). Amino Acid Sequence Evidence on the Phylogeny of Primates and Other Eutherians. In Macromolecular Sequences in Systematic and Evolutionary Biology, p.

62. Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M., and Kumar, S. (2011). MEGA5: Molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol. Biol. Evol.

63. Hilu, K.W., Black, C.M., and Oza, D. (2014). Impact of gene molecular evolution on phylogenetic reconstruction: A case study in the rosids (Superorder Rosanae, angiosperms). PLoS One.

64. Maldonado, E., Sunagar, K., Almeida, D., Vasconcelos, V., and Antunes, A. (2014). IMPACT_S: Integrated Multiprogram Platform to Analyze and Combine Tests of Selection. PLoS One.

65. Tomasco, I.H., Boullosa, N., Hoffmann, F.G., and Lessa, E.P. (2017). Molecular adaptive convergence in the α-globin in subterranean octodontid rodents. Gene.

66. King Jordan, I., Rogozin, I.B., Wolf, Y.I., and Koonin, E. V. (2002). Essential genes are more evolutionarily conserved than are nonessential genes in bacteria. Genome Res.

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

68. Knudsen, B., and Miyamoto, M.M. (2001). A likelihood ratio test for evolutionary rate shifts and functional divergence among proteins. Proc. Natl. Acad. Sci. U. S. A.

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

70. Pál, C., Papp, B., and Hurst, L.D. (2001). Highly expressed genes in yeast evolve slowly. Genetics.

71. Weng, S., Stoner, S.A., and Zhang, D.-E. (2016). Sex chromosome loss and the pseudoautosomal region genes in hematological malignancies. Oncotarget *7*, 72356–72372.

72. Stelzer, G., Rosen, R., Plaschkes, I., Zimmerman, S., Twik, M., Fishilevich, S., Iny Stein, T., Nudel, R., Liede, I., Mazor, Y., et al. (2016). GeneCards – the human gene databas. Curr. Protoc. Bioinforma.

73. Tordjman, S., Chokron, S., Delorme, R., Charrier, A., Bellissant, E., Jaafari, N., and Fougerou, C. (2017). Melatonin: Pharmacology, Functions and Therapeutic Benefits. Curr. Neuropharmacol. *15*, 434–443.

74. Yamanishi, M., Narazaki, H., and Asano, T. (2015). Melatonin overcomes resistance to clofarabine in two leukemic cell lines by increased expression of deoxycytidine kinase. Exp. Hematol. *43*, 207–214.

75. Büyükavci, M., Ozdemir, O., Buck, S., Stout, M., Ravindranath, Y., and Savaşan, S. (2006). Melatonin cytotoxicity in human leukemia cells: relation with its pro-oxidant effect. Fundam. Clin. Pharmacol. *20*, 73–79.

76. Zhang, J., Mullighan, C.G., Harvey, R.C., Wu, G., Chen, X., Edmonson, M., Buetow, K.H., Carroll, W.L., Chen, I.M., Devidas, M., et al. (2011). Key pathways are frequently mutated in high-risk childhood acute lymphoblastic leukemia: A report from the Children's Oncology Group. Blood.

77. Hruz, T., Laule, O., Szabo, G., Wessendorp, F., Bleuler, S., Oertle, L., Widmayer, P., Gruissem, W., and Zimmermann, P. (2008). Genevestigator V3: A Reference Expression Database for the Meta-Analysis of Transcriptomes. Adv. Bioinformatics.

78. Gellatly, S.A., Kalujnaia, S., and Cramb, G. (2012). Cloning, tissue distribution and sub-cellular localisation of phospholipase C X-domain containing protein (PLCXD) isoforms. Biochem. Biophys. Res. Commun.

79. Hamosh, A., Scott, A.F., Amberger, J., Valle, D., and McKusick, V.A. (2000). Online Mendelian Inheritance in Man (OMIM). Hum. Mutat.

80. Ohshima, N., Takahashi, M., and Hirose, F. (2003). Identification of a human homologue of the DREF transcription factor with a potential role in regulation of the histone H1 gene. J. Biol. Chem.

Supplementary material

Table S1 - Accession numbers of the different collected sequences used in phylogenetic and selection analyses. Underlined accession numbers are from genomes.

	AKAP17A	*ASMT*	*ASMTL*	*CD99*	*CRFL2*	*CSF2RA*	*DHRSX*	*GTPBP6*	*IL3RA*	*P2RY8*	*PLCXD1*	*PPP2R3B*	*SHOX*	*SLC25A6*	*ZBED1*
Acinonyx	XM_027052900.1	XM_027052913.1	QURD01	XM_027052873.1			QURD01	XM_027053177.1	XM_027052904.1	XM_027052907.1	XM_027052891.1	XM_027052881.1	XM_027052897.1	XM_027052909.1	XM_027052902.1
Ailuropoda	XM_002920807.3		ENSAMET00000006155.1	XM_019801331.1	XM_002920819.3	XM_019801337.1	XM_011226972.2	XM_011236677.2	XM_011226973.2	XM_002920808.3	XM_002928758.3	XM_002928756.3	XM_011236678.1	XM_019801319.1	XM_019801334.1
Balaenoptera	XM_007173851.1		XM_007173853.1		XM_007173848.1	XM_007173844.1	ATDI01		XM_007173841.1	XM_007173852.1		XM_028164441.1	XM_007173855.1	XM_007173854.1	XM_007173832.1
Bos	XM_027534005.1	XM_005896337.2	XM_027533780.1	XM_027533050.1	XM_027535058.1	XM_014478403.1	XM_027535275.1	XM_027533292.1	XM_027534332.1	XM_024990312.1	XM_005894014.1	XM_027533943.1	XM_027535045.1	XM_027533777.1	XM_027535269.1
Bubalus	XM_025275946.1		XM_025276755.1	XM_006049297.2	XM_006044499.2	XM_006044500.2	XM_025276659.1	XM_025275958.1	XM_025276753.1	XM_006042453.2	XM_006043291.2	XM_006043292.2	XM_025276541.1	XM_025276754.1	XM_025276656.1
Camelus	XM_031446031.1	XM_031446032.1	XM_032474422.1	XM_010968271.1	XM_031446048.1	XM_010968263.1	XM_031446028.1	XM_032474429.1	XM_032474446.1	XM_031446662.1	XM_010997360.2	XM_032474423.1	XM_032474458.1	XM_032474457.1	XM_032474419.1
Canis	XM_025435431.1	XM_022416042.1	XM_025435409.1	XM_025435493.1	XM_005641052.2	XM_005641059.3	XM_025435467.1	XM_005641048.3	XM_014111817.2	XM_022416539.1	XM_025420347.1	XM_003640273.4	NM_001025622.2	XM_025435396.1	XM_005641065.3
Capra	XM_018044444.1	NM_001285598.1	XM_018044443.1	XM_018043745.1	XM_005701362.3	XM_005701352.3	XM_018044442.1	XM_018044432.1	XM_018044558.1	XM_013976840.2	XM_018044434.1	XM_018044433.1	XM_018045110.1	XM_018044555.1	XM_018044438.1
Castor	XM_020167211.1		XM_020167239.1		XM_020167267.1	XM_020174840.1	XM_020172211.1		XM_020167319.1	XM_020167358.1		XM_020183676.1		XM_020167347.1	XM_020172210.1
Cebus	XM_017518846.1	XM_017518838.1	XM_017518847.1	XM_017518837.1			XM_017518853.1	XM_017518357.1	XM_017518839.1	XM_017518851.1	XM_017518355.1	XM_017518354.1	XM_017518356.1	XM_017518854.1	XM_017518841.1
Ceratotherium	XM_004440757.2	XM_014795513.1	XM_014795515.1		XM_014795518.1	XM_004440778.1		XM_014795519.1	XM_014795516.1	XM_014795514.1	XM_004440781.2	XM_014795527.1	XM_004440760.2		XM_004440756.2
Chinchilla	XM_013505618.1	XM_005406516.1	XM_013505602.1	XM_013505598.1	XM_013505605.1	XM_013505607.1	XM_013505592.1	XM_013505582.1	XM_013505601.1	XM_005406556.2	XM_013505586.1	XM_013505589.1	XM_005406505.2		XM_013505591.1
Cricetulus	XM_027434168.1					XM_027397973.1	XM_027434166.1	XM_027406636.1	XM_027434161.1		XM_027424746.1				
Diceros	PVJY02		PVJY02		PVJY02	PVJY02	PVJY02	PVJY02	PVJY02	PVJY02	PVJY02		PVJY02	PVJY02	PVJY02
Equus	XM_014826769.1	XM_023633485.1	XM_014859108.1	XM_023633871.1	XM_014859060.1	XM_014859040.1	XM_023633482.1	XM_023633894.1	ENSECAT00000048497.2	XM_023633951.1	XM_023633896.1	XM_023633853.1	XM_023633696.1	XM_023634414.1	XM_023634372.1
Eumetopias	XM_028115022.1		XM_028115035.1	XM_028115086.1	XM_028115006.1		XM_028114960.1	XM_028115024.1	XM_028114988.1	XM_028115037.1	XM_028114951.1	XM_028114982.1	XM_028115088.1	XM_028114990.1	XM_028114954.1
Felis	XM_023248928.1	XM_023249442.1	XM_023248897.1		XM_023249294.1	XM_019823499.2	XM_023249148.1	XM_011291581.2	XM_023248887.1	XM_023249097.1	XM_004000235.5	XM_023249514.1	XM_011291584.3	XM_023248888.1	XM_023249146.1
Heterocephalus	XM_021246398.1		XM_004867313.3	XM_004867320.3	XM_004867306.3	XM_004867307.2	XM_004867319.2	XM_004867368.3	XM_021246368.1	XM_004867372.3	XM_021246360.1	XM_021246363.1	XM_004867304.1	XM_004867311.3	XM_013066494.2
Homo	NM_005088.3	NM_001171038.2	NM_001173473.1	NM_001122898.3	NM_022148.4	NM_001161529.1	NM_145177.3	NM_012227.3	NM_002183.4	XM_011546179.2	NM_018390.4	BC063429.1	NM_000451.3	NM_001636.4	NM_004729.4
Macaca	XM_028841650.1	NM_001032940.1	XM_028843002.1	XM_015126884.2	XM_028841970.1		XM_028842026.1	XM_028842021.1	XM_028842960.1	XM_011735017.2	XM_015126854.2	XM_028842031.1	XM_028842081.1	XM_015126871.2	XM_028841621.1

Continuation of Table S1

	AKAP17A	*ASMT*	*ASMTL*	*CD99*	*CRFL2*	*CSF2RA*	*DHRSX*	*GTPBP6*	*IL3RA*	*P2RY8*	*PLCXD1*	*PPP2R3B*	*SHOX*	*SLC25A6*	*ZBED1*
Marmota	XM_027954071.1	XM_027954106.1	XM_027954072.1	XM_027954101.1	XM_027954082.1	XM_015506862.1	XM_015506839.1	XM_027954094.1	XM_027954086.1	XM_027954074.1			XM_027954085.1	XM_027954084.1	XM_027954093.1
Microcebus	XM_012763670.2		XM_020285217.1	ABDC03	XM_012763651.2	XM_012763658.2	XM_020285150.1	ABDC03	XM_012763662.2	XM_012763665.2	XM_020285240.1	ENSMICT00000066621.1	XM_012737274.1	XM_012763664.2	XM_012763681.2
Microtus	XM_013355630.2	XM_005372239.2		ENSMOCT00000023246.1	XM_026778156.1	XM_026778154.1	XM_005370397.3	XM_005356575.1	XM_013355591.2		XM_013354319.1	XM_005348711.2			
Monodelphis	XM_001362178.4	XM_001363802.4	XM_007501012.2	XM_007500996.2	XM_007501023.1		XM_007501007.1	XM_001365342.4	XM_007501016.2	XM_007501009.2	XM_001365413.3	XM_001365071.3	XM_007501027.2	XM_001362259.4	XM_007501004.2
Mus	XM_029473888.1	NM_001199212.1			XM_021154266.2	XM_021189491.2	BC138599.1	NM_145147.5	XM_029541845.1		XM_021189410.1				
Mustela	XM_032331710.1	XM_032331743.1	XM_032331714.1	XM_032331812.1	XM_032331731.1	XM_032331724.1	XM_032331736.1	XM_032331721.1	XM_032331727.1	XM_032331732.1	XM_032330062.1	XM_032331716.1	XM_032331741.1	XM_032331740.1	XM_032331712.1
Nannospalax	XM_029556107.1	XM_029556119.1	XM_029556106.1		XM_008853072.2	XM_008853052.3	XM_008853062.3	XM_008853047.1	XM_029556105.1		XM_029556126.1	XM_008853049.3	XM_008853051.3	XM_008853055.3	XM_008853061.2
Ochotona	XM_012931060.1	XM_004599452.1	XM_004582127.1		XM_012929304.1				XM_004586133.1	XM_004580877.1		XM_012925968.1			XM_004599848.1
Octodon	XM_023700953.1	XM_004646632.1		XM_023706283.1		XM_023706889.1		XM_004646623.2	XM_023706891.1	XM_004646630.2		XM_023706887.1			XM_023706912.1
Ovis	XM_027963311.1	NM_001306120.1	XM_027963314.1	XM_027963297.1	XM_015104672.2	XM_027963315.1	XM_027963301.1	XM_027962935.1	XM_027963322.1	XM_027963324.1	XM_027962963.1		XM_027963325.1	NM_001127280.1	XM_027963300.1
Pan	XM_016943792.2	XM_008967035.1	XM_008967020.1	XM_003317334.4	XM_024353223.1	XM_024353063.1	XM_024353111.1	XM_016943603.1	ENSPTRT00000108204.1	XM_016943399.2	XM_008958940.3	XM_024353389.1	XM_016944342.2	XM_016947683.2	XM_024927221.1
Panthera	XM_019431190.1	XM_019431122.1	XM_019431123.1		XM_019431185.1	XM_019431183.1	XM_019431146.1	XM_019431128.1	XM_019431125.1	XM_019431138.1	XM_019439230.1	XM_019431131.1	ENSPPRT00000013334.1	XM_019431124.1	XM_019431176.1
Phocoena	XM_032620240.1		VOSU01	XM_032619115.1	VOSU01	VOSU01	XM_032618762.1			VOSU01		VOSU01	XM_032620410.1	VOSU01	XM_032618741.1
Pongo	XM_009234560.2		NM_001135364.1	XM_024241550.1		XM_024241556.1	XM_024241546.1	XM_024241535.1	ENSPPYT00000024417.2	XM_024241561.1	XM_024241581.1	XM_024241572.1	XM_002831323.4	XM_002831328.4	XM_024241544.1
Rattus	XM_006248999.2	EU741665.1_cds			XM_032916581.1	XM_032916576.1	BC167053.1	NM_001135840.1	NM_139260.3		XM_032886801.1	NM_001139492.1			
Sus	XM_021081363.1	XM_021081387.1	XM_021081390.1		XM_021081385.1	XM_021081380.1	XM_021082457.1	XM_021082210.1		XM_021081375.1	XM_021082211.1	XM_021081366.1	XM_021081396.1	AY528245.2	XM_021082456.1
Theropithecus	XM_025373455.1		XM_025373342.1	XM_025372045.1	XM_025373340.1	XM_025372876.1	XM_025372286.1	XM_025373569.1	XM_025373341.1	XM_025372975.1	XM_025372603.1	OGDE01	XM_025373311.1	XM_025372079.1	XM_025372285.1
Urocitellus	XM_026382620.1	XM_026382618.1	XM_026382303.1	XM_026381501.1	XM_026381470.1	XM_026382301.1	XM_026381494.1	XM_026382525.1	XM_026382297.1	XM_026382304.1	XM_026382529.1		XM_026381471.1	XM_026382295.1	XM_026381493.1
Ursus	XM_008694338.1	XM_008694303.1	XM_008694302.1	XM_008694304.1	XM_008694299.1	XM_008694301.1	XM_026478202.1	XM_008694296.1	XM_008694335.1	XM_026478205.1	XM_008694295.1	XM_008694333.1		XM_008694336.1	XM_008694340.1
Vicugna	ABRR02	XM_006210170.2	ABRR02			XM_015244356.2	XM_031671220.1	ABRR02	XM_031671269.1	XM_031671276.1	XM_006210169.3	XM_015244334.2	ABRR02	XM_031671273.1	
Zalophus	XM_027609155.1	XM_027609158.1	XM_027608957.1	XM_027607922.1	XM_027608955.1	XM_027607268.1	XM_027608406.1	XM_027609040.1	XM_027609486.1	XM_027607313.1	XM_027607285.1	XM_027607963.1	XM_027609101.1	XM_027609054.1	XM_027608405.1

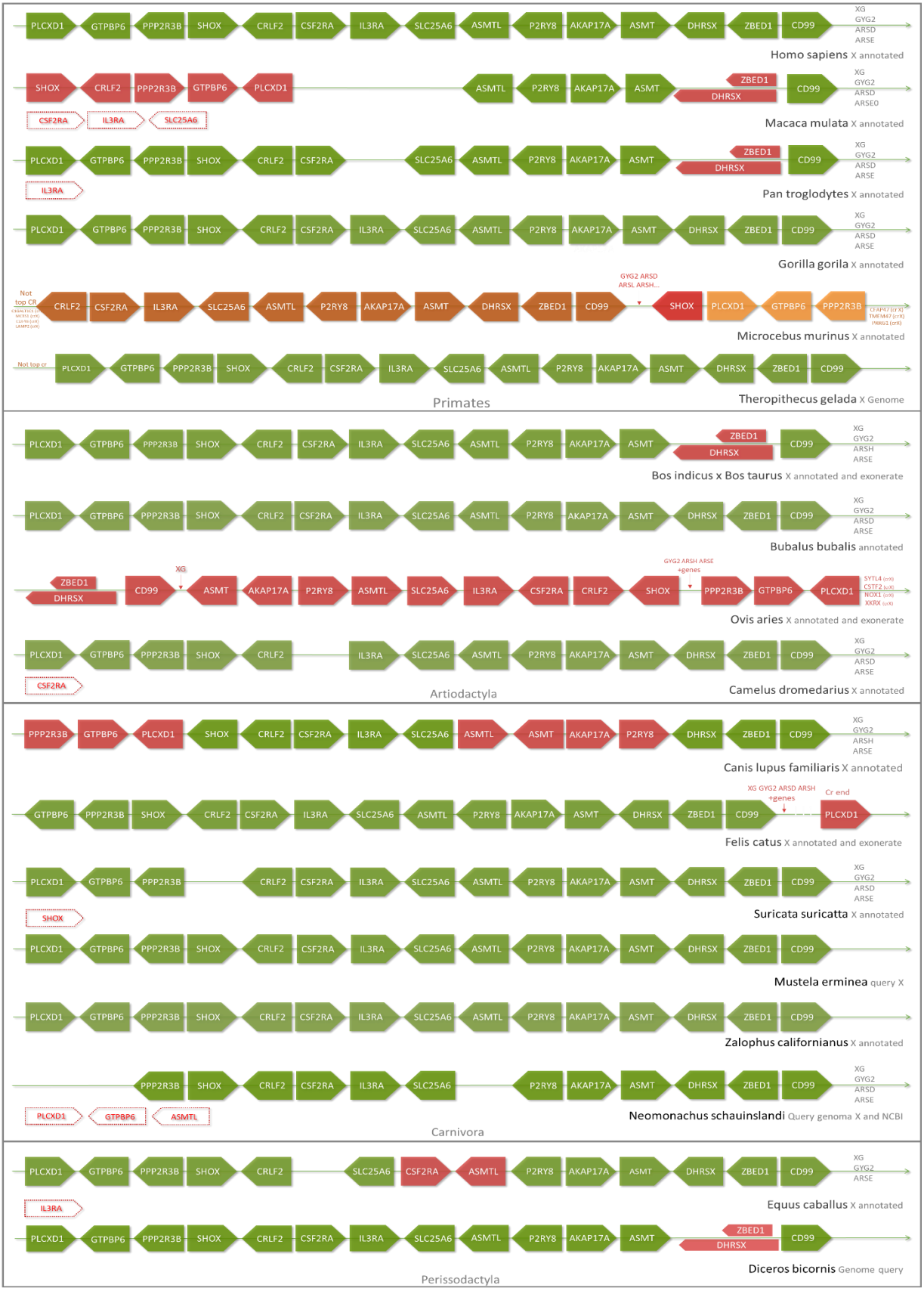

Figure S1 - Result of the synteny analyses. Genes in green had maintained the position, genes in red, orange and yellow had the location altered. Genes in a blank background box under the arrow were missing from the assessed species.

Figure S2 - Phylogenetic trees constructed employing the ML, BI algorithm and the TimeTree webserver. Values at each node correspond to branch support.

Table S2 - Result of the selection site models analyses performed in CodeML. Values in bold have a significant *P* (<0.05).

	model 7					model8											
	lnL	np	p	q	kappa	lnL	np	p0	p	q	p1	*ω*	kappa	BEB	LRT	DF	*P*
AKAP17A	-12770.227	80	0.606	15.234	2.339	-12770.232	82	1.000	0.606	15.234	0.000	2.682	2.339	NF	0.000	2	1.000
ASMT	-10711.074	56	0.544	2.111	2.131	-10711.057	58	0.995	0.551	2.185	0.005	1.000	2.132	NF	0.034	2	0.983
ASMTL	-12692.716	70	0.608	2.469	2.741	-12687.133	72	0.964	0.717	3.578	0.036	1.000	2.750	NF	11.165	2	**0.004**
CD99	-2141.884	54	0.453	2.538	3.106	-2141.884	56	1.000	0.453	2.539	0.000	1.000	3.106	NF	0.000	2	1.000
CRLF2	-12092.432	68	0.698	2.042	2.270	-12092.434	70	1.000	0.698	2.041	0.000	3.582	2.270	NF	0.000	2	1.000
CSF2RA	-15401.979	70	0.719	1.929	2.281	-15401.982	72	1.000	0.719	1.929	0.000	6.040	2.281	NF	0.000	2	1.000
DHRSX	-11359.011	74	0.470	2.780	2.265	-11357.262	76	0.986	0.505	3.314	0.014	1.000	2.273	NF	3.497	2	0.174
GTPBP6	-14470.730	72	0.603	4.013	3.259	-14470.734	74	1.000	0.603	4.013	0.000	26.726	3.259	NF	0.000	2	1.000
IL3RA	-6914.185	76	0.725	1.859	1.966	-6912.542	78	0.990	0.752	2.019	0.010	2.127	1.974	0.986* 102 A	3.287	2	0.193
P2RY8	-8344.698	70	0.306	3.320	1.886	-8344.701	72	1.000	0.306	3.320	0.000	15.553	1.886	NF	0.000	2	1.000
PLCXD1	-9919.111	68	0.477	4.085	2.971	-9914.661	70	0.997	0.493	4.432	0.003	1.769	2.986	NF	8.900	2	**0.012**
PPP2R3B	-12095.389	68	0.585	9.238	2.471	-12094.487	70	0.994	0.616	10.246	0.006	1.000	2.485	NF	1.806	2	0.405
SHOX	-4921.559	64	0.339	7.199	1.667	-4921.562	66	1.000	0.339	7.199	0.000	2.985	1.667	NF	0.000	2	1.000
SLC25A6	-4437.660	64	0.082	4.290	3.239	-4437.662	66	1.000	0.082	4.290	0.000	1.910	3.239	NF	0.000	2	1.000
ZBED1	-18863.598	70	0.500	8.578	2.648	-18863.605	72	1.000	0.500	8.578	0.000	2.441	2.648	NF	0.000	2	1.000

Table S3 – Results of the Branch – Site analyses performed in CodeML for Primates.

	Null								Alt											
	np	lnL	kappa	ω	0	1	2a	2b	np	lnL	kappa	ω	0	1	2a	2b	LRT	DF	*P*	BEB
AKAP17A	81	-13025.797	2.261	proportion	0.980	0.012	0.008	0.000	82	-13025.797	2.261	proportion	0.980	0.012	0.008	0.000				
				background w	0.032	1.000	0.032	1.000				background w	0.032	1.000	0.032	1.000				
				foreground w	0.032	1.000	1.000	1.000				foreground w	0.032	1.000	1.000	1.000	0.000	1	1.000	
ASMT	57	-10844.089	2.203	proportion	0.653	0.235	0.082	0.030	58	-10844.059	2.203	proportion	0.672	0.239	0.066	0.023				201 R 0.983*
				background w	0.104	1.000	0.104	1.000				background w	0.105	1.000	0.105	1.000				
				foreground w	0.104	1.000	1.000	1.000				foreground w	0.105	1.000	1.339	1.339	0.060	1	0.806	
ASMTL	71	-12810.457	2.842	proportion	0.717	0.169	0.093	0.022	72	-12810.457	2.842	proportion	0.717	0.169	0.093	0.022				
				background w	0.108	1.000	0.108	1.000				background w	0.108	1.000	0.108	1.000				
				foreground w	0.108	1.000	1.000	1.000				foreground w	0.108	1.000	1.000	1.000	0.000	1	1.000	
CD99	55	-2169.873	3.055	proportion	0.740	0.126	0.115	0.020	56	-2169.490	3.068	proportion	0.757	0.126	0.100	0.017				19 G 0.999**
				background w	0.076	1.000	0.076	1.000				background w	0.077	1.000	0.077	1.000				21 D 0.974*
				foreground w	0.076	1.000	1.000	1.000				foreground w	0.077	1.000	1.624	1.624	0.767	1	0.381	
CRLF2	69	-12234.296	2.438	proportion	0.666	0.222	0.084	0.028	70	-12234.272	2.437	proportion	0.674	0.225	0.076	0.025				251 N 0.994**
				background w	0.155	1.000	0.155	1.000				background w	0.155	1.000	0.155	1.000				
				foreground w	0.155	1.000	1.000	1.000				foreground w	0.155	1.000	1.107	1.107	0.048	1	0.826	
CSF2RA	71	-15566.142	2.473	proportion	0.656	0.344	0.000	0.000	72	-15566.142	2.473	proportion	0.656	0.344	0.000	0.000				
				background w	0.148	1.000	0.148	1.000				background w	0.148	1.000	0.148	1.000				
				foreground w	0.148	1.000	1.000	1.000				foreground w	0.148	1.000	13.966	13.966	0.000	1	1.000	
DHRSX	75	-11524.725	2.350	proportion	0.834	0.133	0.028	0.004	76	-11524.725	2.350	proportion	0.834	0.133	0.028	0.004				
				background w	0.086	1.000	0.086	1.000				background w	0.086	1.000	0.086	1.000				
				foreground w	0.086	1.000	1.000	1.000				foreground w	0.086	1.000	1.000	1.000	0.000	1	1.000	
GTPBP6	73	-14637.122	3.242	proportion	0.875	0.098	0.024	0.003	74	-14637.122	3.242	proportion	0.875	0.098	0.024	0.003				120 D 0.970*
				background w	0.091	1.000	0.091	1.000				background w	0.091	1.000	0.091	1.000				272 A 0.983*
				foreground w	0.091	1.000	1.000	1.000				foreground w	0.091	1.000	1.000	1.000	0.000	1	1.000	

Continuation of Table S3

	Null								Alt										
	np	lnL	kappa	ω	0	1	2a	2b	np	lnL	kappa	ω	0	1	2a	2b	LRT	DF	*P*	BEB
IL3RA	77	-6988.024	2.077	proportion	0.675	0.216	0.082	0.026	78	-6988.024	2.077	proportion	0.675	0.216	0.082	0.026				79 E 0.997**
				background w	0.163	1.000	0.163	1.000				background w	0.163	1.000	0.163	1.000				
				foreground w	0.163	1.000	1.000	1.000				foreground w	0.163	1.000	1.000	1.000	0.000	1	1.000	
P2RY8	71	-8525.937	1.583	proportion	0.913	0.083	0.003	0.000	72	-8525.948	1.570	proportion	0.914	0.084	0.002	0.000				
				background w	0.061	1.000	0.061	1.000				background w	0.061	1.000	0.061	1.000				
				foreground w	0.061	1.000	1.000	1.000				foreground w	0.061	1.000	1.000	1.000	0.000	1	1.000	
PLCXD1	69	-10114.976	2.803	proportion	0.813	0.134	0.045	0.007	70	-10114.976	2.803	proportion	0.813	0.134	0.045	0.007				162 V 0.983*
				background w	0.074	1.000	0.074	1.000				background w	0.074	1.000	0.074	1.000				
				foreground w	0.074	1.000	1.000	1.000				foreground w	0.074	1.000	1.000	1.000	0.000	1	1.000	
PPP2R3B	69	-12292.227	2.525	proportion	0.869	0.031	0.097	0.003	70	-12292.227	2.525	proportion	0.869	0.031	0.097	0.003				3 I 0.953* 39 T 0.970*
				background w	0.041	1.000	0.041	1.000				background w	0.041	1.000	0.041	1.000				10 R 0.972* 86 Q 0.956*
				foreground w	0.041	1.000	1.000	1.000				foreground w	0.041	1.000	1.000	1.000	0.000	1	1.000	322 K 0.966* 344 I 0.969*143 I 0.999**
SHOX	65	-4953.919	1.705	proportion	0.974	0.026	0.000	0.000	66	-4953.919	1.705	proportion	0.974	0.026	0.000	0.000				
				background w	0.031	1.000	0.031	1.000				background w	0.031	1.000	0.031	1.000				
				foreground w	0.031	1.000	1.000	1.000				foreground w	0.031	1.000	1.000	1.000	0.000	1	1.000	
SLC25A6	65	-4499.304	3.200	proportion	0.987	0.013	0.000	0.000	66	-4499.304	3.200	proportion	0.987	0.013	0.000	0.000				
				background w	0.009	1.000	0.009	1.000				background w	0.009	1.000	0.009	1.000				
				foreground w	0.009	1.000	1.000	1.000				foreground w	0.009	1.000	4.874	4.874	0.000	1	1.000	
ZBED1	71	-19338.993	2.565	proportion	0.977	0.019	0.003	0.000	72	-19338.993	2.565	proportion	0.977	0.019	0.003	0.000				379 A 0.995**
				background w	0.048	1.000	0.048	1.000				background w	0.048	1.000	0.048	1.000				
				foreground w	0.048	1.000	1.000	1.000				foreground w	0.048	1.000	1.000	1.000	0.000	1	1.000	

Table S4 – Results of the Branch – Site analyses performed in CodeML for Rodentia.

	Null								Alt											
	np	lnL	kappa	*ω*	0	1	2a	2b	np	lnL	kappa	*ω*	0	1	2a	2b	LRT	DF	*P*	BEB
AKAP17A	81	-13025.195	2.259	proportion	0.985	0.010	0.004	0.000	82	-13025.195	2.259	proportion	0.985	0.010	0.004	0.000				
				background w	0.032	1.000	0.032	1.000				background w	0.032	1.000	0.032	1.000				
				foreground w	0.032	1.000	1.000	1.000				foreground w	0.032	1.000	1.000	1.000	0.000	1	1.000	
ASMT	57	-10842.159	2.169	proportion	0.658	0.231	0.082	0.029	58	-10842.159	2.169	proportion	0.658	0.231	0.082	0.029				
				background w	0.099	1.000	0.099	1.000				background w	0.099	1.000	0.099	1.000				
				foreground w	0.099	1.000	1.000	1.000				foreground w	0.099	1.000	1.000	1.000	0.000	1	1.000	
ASMTL	71	-12816.235	2.821	proportion	0.756	0.189	0.044	0.011	72	-12816.235	2.821	proportion	0.756	0.189	0.044	0.011				206 E 0.961*
				background w	0.111	1.000	0.111	1.000				background w	0.111	1.000	0.111	1.000				
				foreground w	0.111	1.000	1.000	1.000				foreground w	0.111	1.000	1.000	1.000	0.000	1	1.000	
CD99	55	-2166.267	3.118	proportion	0.701	0.095	0.179	0.024	56	-2166.267	3.118	proportion	0.701	0.095	0.179	0.024				60 A 0.995**
				background w	0.082	1.000	0.082	1.000				background w	0.082	1.000	0.082	1.000				62 G 0.987*
				foreground w	0.082	1.000	1.000	1.000				foreground w	0.082	1.000	1.000	1.000	0.000	1	1.000	
CRLF2	69	-12203.731	2.392	proportion	0.613	0.183	0.157	0.047	70	-12203.731	2.392	proportion	0.613	0.183	0.157	0.047				25 E 0.993** 150 Y 0.982*110 F 0.999**
				background w	0.142	1.000	0.142	1.000				background w	0.142	1.000	0.142	1.000				46 D 0.969* 174 F 0.991** 129 E 0.993**
				foreground w	0.142	1.000	1.000	1.000				foreground w	0.142	1.000	1.000	1.000	0.000	1	1.000	87 A 0.990* 130 K 0.958*
CSF2RA	71	-15531.975	2.420	proportion	0.595	0.234	0.123	0.048	72	-15531.975	2.420	proportion	0.595	0.234	0.123	0.048				14 D 1.000**125 G 0.992**
				background w	0.137	1.000	0.137	1.000				background w	0.137	1.000	0.137	1.000				34 D 0.975* 201 N 0.956*
				foreground w	0.137	1.000	1.000	1.000				foreground w	0.137	1.000	1.000	1.000	0.000	1	1.000	59 V 0.998**
DHRSX	75	-11517.880	2.332	proportion	0.821	0.122	0.050	0.007	76	-11517.880	2.332	proportion	0.821	0.122	0.050	0.007				83 K 0.988*
				background w	0.086	1.000	0.086	1.000				background w	0.086	1.000	0.086	1.000				
				foreground w	0.086	1.000	1.000	1.000				foreground w	0.086	1.000	1.000	1.000	0.000	1	1.000	
GTPBP6	73	-14626.758	3.263	proportion	0.867	0.082	0.047	0.004	74	-14626.758	3.263	proportion	0.867	0.082	0.047	0.004				
				background w	0.089	1.000	0.089	1.000				background w	0.089	1.000	0.089	1.000				
				foreground w	0.089	1.000	1.000	1.000				foreground w	0.089	1.000	1.000	1.000	0.000	1	1.000	

Continuation of Table S4

	Null								Alt										
	np	lnL	kappa	ω	0	1	2a	2b	np	lnL	kappa	ω	0	1	2a	2b	LRT	DF	*P*	BEB
IL3RA	77	-6989.538	2.061	proportion	0.709	0.215	0.058	0.018	78	-6989.538	2.061	proportion	0.709	0.215	0.058	0.018				
				background w	0.167	1.000	0.167	1.000				background w	0.167	1.000	0.167	1.000				
				foreground w	0.167	1.000	1.000	1.000				foreground w	0.167	1.000	1.000	1.000	0.000	1	1.000	
P2RY8	71	-8494.444	1.609	proportion	0.859	0.065	0.071	0.005	72	-8494.444	1.609	proportion	0.859	0.065	0.071	0.005				108 S 0.956* 188 M 0.954*
				background w	0.055	1.000	0.055	1.000				background w	0.055	1.000	0.055	1.000				129 A 0.995** 217 L 1.000**
				foreground w	0.055	1.000	1.000	1.000				foreground w	0.055	1.000	1.000	1.000	0.000	1	1.000	226 H 0.993** 270 Y 0.999** 281 S 0.997**
PLCXD1	69	-10107.256	2.875	proportion	0.840	0.118	0.037	0.005	70	-10107.256	2.875	proportion	0.840	0.118	0.037	0.005				39 Q 0.974* 268 R 0.977*
				background w	0.073	1.000	0.073	1.000				background w	0.073	1.000	0.073	1.000				50 L 0.973*
				foreground w	0.073	1.000	1.000	1.000				foreground w	0.073	1.000	1.000	1.000	0.000	1	1.000	155 M 0.992**
PPP2R3B	69	-12291.203	2.502	proportion	0.901	0.033	0.063	0.002	70	-12291.203	2.502	proportion	0.901	0.033	0.063	0.002				161 A 0.961* 308 E 0.999**
				background w	0.044	1.000	0.044	1.000				background w	0.044	1.000	0.044	1.000				261 K 0.973*
				foreground w	0.044	1.000	1.000	1.000				foreground w	0.044	1.000	1.000	1.000	0.000	1	1.000	294 C 0.999**
SHOX	65	-4953.227	1.705	proportion	0.968	0.025	0.007	0.000	66	-4953.227	1.705	proportion	0.968	0.025	0.007	0.000				
				background w	0.030	1.000	0.030	1.000				background w	0.030	1.000	0.030	1.000				
				foreground w	0.030	1.000	1.000	1.000				foreground w	0.030	1.000	1.000	1.000	0.000	1	1.000	
SLC25A6	65	-4497.177	3.213	proportion	0.982	0.009	0.009	0.000	66	-4497.177	3.213	proportion	0.982	0.009	0.009	0.000				
				background w	0.009	1.000	0.009	1.000				background w	0.009	1.000	0.009	1.000				
				foreground w	0.009	1.000	1.000	1.000				foreground w	0.009	1.000	1.000	1.000	0.000	1	1.000	
ZBED1	71	-19260.665	2.498	proportion	0.893	0.014	0.091	0.001	72	-19260.665	2.498	proportion	0.893	0.014	0.091	0.001				7 D 0.986* 10 Q 1.000** 103 D 0.952* 321 M 0.977* 325 Y 0.983* 333 P 0.990**
				background w	0.049	1.000	0.049	1.000				background w	0.049	1.000	0.049	1.000				108 K 0.964* 111 H 0.959*161 K 0.965* 337 V 0.981* 425 N 0.990**
				foreground w	0.049	1.000	1.000	1.000				foreground w	0.049	1.000	1.000	1.000	0.000	1	1.000	195 M 0.997** 197 R 0.997** 221 G 0.995** 502 G 0.990* 503 Y 0.990** 508 K 0.965* 548 W 1.000**666 I 0.977*