Detection of pathways affected by positive selection in primate lineages ancestral to humans

Daub, J.T. [1,2,4,*], Moretti, S. [2,3], Davidov, I. I. [2,3], Excoffier, L. [1,2], Robinson-Rechavi, M. [2,3,*]

[1] CMPG, Institute of Ecology and Evolution, University of Berne, Baltzerstrasse 6, 3012 Berne, Switzerland

[2] SIB Swiss Institute of Bioinformatics, 1015 Lausanne Switzerland

[3] Dept. of Ecology and Evolution, University of Lausanne, 1015 Lausanne, Switzerland

[4] Present address: Institute of Evolutionary Biology (UPF-CSIC), Barcelona, Spain

Corresponding author: E-mail: josephine.daub@upf.edu; marc.robinson-rechavi@unil.ch.

Abstract

Gene set enrichment approaches have been increasingly successful in finding signals of recent polygenic selection in the human genome. In this study, we aim at detecting biological pathways affected by positive selection in more ancient human evolutionary history, that is in four branches of the primate tree that lead to modern humans. We tested all available protein coding gene trees of the Primates clade for signals of adaptation in the four branches mentioned above, using the likelihood-based branch site test of positive selection. The results of these locus-specific tests were then used as input for a gene set enrichment test, where whole pathways are globally scored for a signal of positive selection, instead of focusing only on outlier "significant" genes. We identified several pathways enriched for signals of positive selection, which are mainly involved in immune response, sensory perception, metabolism, and energy production. These pathway-level results were highly significant, at odds with an absence of any functional enrichment when only focusing on top scoring genes. Interestingly, several gene sets are found significant at multiple levels in the phylogeny, but in such cases different genes are responsible for the selection signal in the different branches, suggesting that the same function has been optimized in different ways at different times in primate evolution.

Introduction

An important challenge in the study of protein evolution is the detection of substitutions fixed by positive selection on a background of genetic drift and purifying selection. The detection of such positive selection signal has progressed thanks to better codon models and statistical tests (Delport et al. 2009). Yet these tests suffer from low power (Anisimova and Yang 2007; Gharib and Robinson-Rechavi 2013), especially when they are applied to closely related species, where phylogenetic trees only have relatively short branches with few descending sequences. The situation is worse when positive selection is weaker, and thus harder to detect, e.g. in species with small population sizes. If cumulated, these effects make it notably difficult to reliably detect positive selection in recent primate evolution, such as on the phylogenetic branches directly leading to humans.

Despite these inherent limitations, there is much interest in detecting positive selection in humans and in the primate evolution which led to our lineage (Lachance and Tishkoff 2013). For example, as soon as the chimpanzee genome was available, genome-wide scans using codon models were performed, resulting in the detection of fast evolving genes related to immunity, host defense, or reproduction (Chimpanzee Sequencing and Analysis Consortium 2005; Nielsen et al. 2005). However, these genome scans often lacked power to distinguish positive selection from relaxed purifying selection. With more species (i.e., more data) positive selection was eventually detected, but very few genes remained significant after correcting for multiple tests (e.g. Bakewell et al. 2007; Gibbs et al. 2007), and thus only those genes with mutations of very large effect were identified.

While these limited results are interesting, a more comprehensive picture of positive selection in the evolution of the human lineage still remains to be uncovered. Here we propose to combine potentially moderate signals from several genes to gain statistical power, using biologically meaningful groupings of genes, such as known regulatory and metabolic pathways. Indeed, several genes with small effect mutations can altogether have a large impact on a biological pathway, even though these genes would have little chance to be identified by conventional genome scans. An increasing number of studies has thus shifted the focus from single gene approaches to the detection of polygenic selection (e.g. Serra et al. 2011; Daub et al. 2013; Fraser 2013; Berg and Coop 2014; Carneiro et al. 2014; Wellenreuther and Hansson 2016), taking advantage of existing databases of gene sets and pathways. For example, we have used a gene set enrichment test to detect gene sets with polygenic adaptive signals in human populations (Daub et al. 2013), showing that most significant pathways for positive selection were involved in defense against pathogens. This procedure has also been successfully applied to find signals of convergent adaptation in humans living at high altitudes (Foll et al. 2014) or in tropical forests (Amorim et al. 2015) and to detect positive selection in ant genomes using the results of the branch-site test (Roux et al. 2014). More recently, we have extended the classical McDonald-Kreitman test (McDonald and Kreitman 1991) to detect more ancient selection signals, i.e. outlier pathways affected by different modes of selection after the split of humans with chimpanzee (Daub et al. 2015).

In the present study, we investigate whether one can find traces of positive selection in older periods of human evolution. We use the branch-site likelihood-based test, contrasting codon models with and without positive selection (Zhang et al. 2005). This testing procedure can detect episodic positive selection, and it has been widely used to detect ancient positive selection in various species, including primates and other mammals (Zhang et al. 2005; Bakewell et al. 2007; Kosiol et al. 2008; Studer et al. 2008). We propose here to combine the gene-specific likelihoods of branch-site tests over all members of a gene set, to infer which biological systems have been under positive selection in the primate ancestors of humans.

Using this approach we detect signals of adaptation in pathways involved in immune response, sensory perception, metabolism, and energy production. Furthermore, we demonstrate that in many candidate pathways different genes are responsible for the selective signal in different time periods in primate evolution.

Results

We have performed a gene set enrichment analysis to detect positive selection at the pathway level in the inner branches of a phylogenetic tree leading to African great apes (Homininae), great apes (Hominidae), apes (Hominoidae), and old world monkeys and apes (Catarrhini) (Figure 1). We first ran separate branch-site tests of positive selection on inner branches of 15,738 protein coding gene trees of the Primates clade. The output of the procedure is a log-likelihood ratio test (ΔlnL) statistic, comparing the likelihood of a model without positive selection to that of a model with one additional parameter for positive selection. If a branch in a gene tree has a high ΔlnL statistic, it means that there is support for the corresponding gene to have a subset of positively selected codons (i.e. dN/dS>1) on this branch.

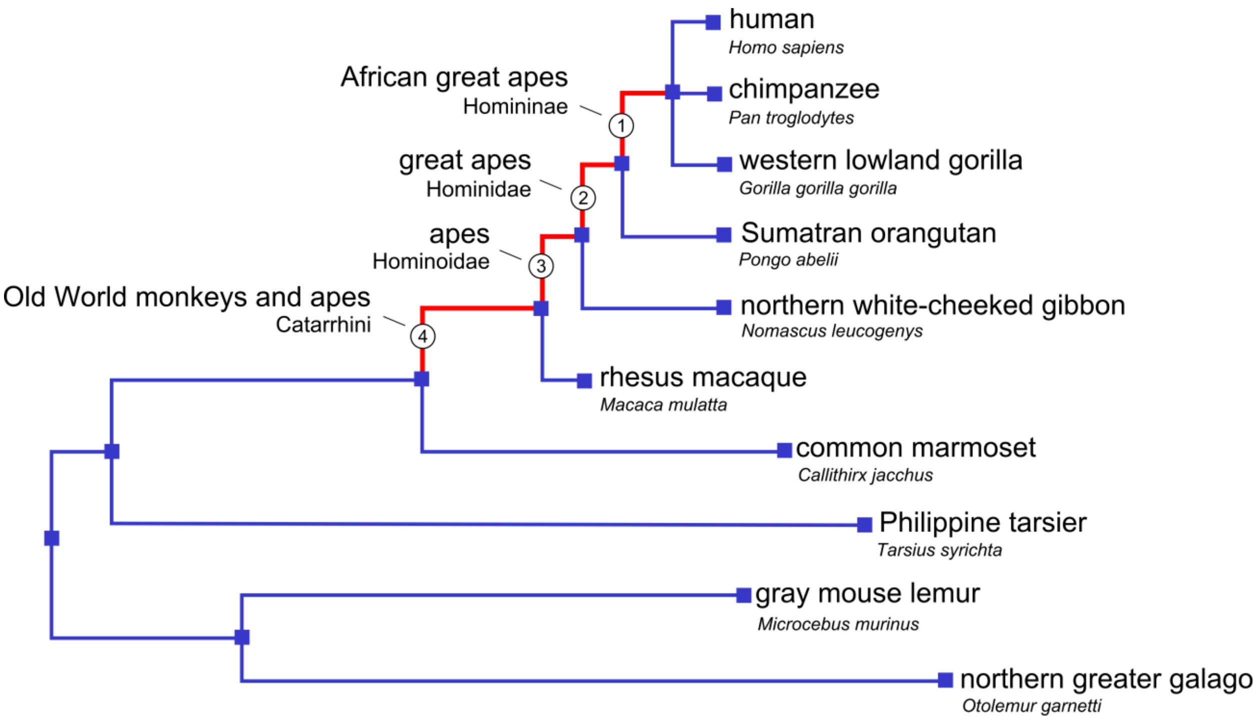


Figure 1 The Primates clade with the species used in the branch-site test. The four tested branches (Homoninae, Hominidae, Hominoidae, and Catarrhini) are numbered (used to identify branch specific lists of genes or gene sets, e.g. G_1, G_2, G_3, G_4) and marked in red (modified from http://selectome.unil.ch).

We tested over 1,400 pathways from the Biosystems database (Geer et al. 2010) for episodic positive selection (Table 1). For each pathway and each branch, we calculated the sum of the ΔlnL4 values of the genes in this set (where ΔlnL4 is the fourth root of ΔlnL, see Materials and Methods), and we inferred the significance of this 'SUMSTAT' score (Tintle et al. 2009) against a null distribution of random gene sets of the same size. As shown in a previous study on the human specific branch (Daub et al. 2015), we find that genes in gene sets tend to be more conserved in internal branches of the Primates tree than genes that are not included in any gene set. This pattern is most pronounced in the two youngest branches leading to the Homininae and the Hominidae (Supplemental Figure S1). To account for the fact that some genes in sets are more conserved, we created null distributions that explicitly reflect the expected genomic background by randomly sampling from the group of genes that are part of at least one gene set.

As we are essentially interested in finding gene sets that show a global shift in selection scores and not those gene sets that include a single or a few extremely significant genes, we repeated the procedure after excluding in each set the highest scoring gene. Only pathways that scored a q-value below 0.2 (thus allowing a 20% FDR) both before and after exclusion of the top significant gene were considered candidates for polygenic selection.

Table 1 Number of gene sets and genes part of sets in the four tested branches. For each branch the number of significant sets (q < 0.2) in the SUMSTAT gene set enrichment test is reported, both before and after removing overlapping genes ('pruning'), as well as the number of significant sets before pruning that remain significant after removal of their highest scoring gene.

					# significant sets		
	branch	**leading to[a]**	**# sets**	**# genes in sets**	**before pruning**	**without top scoring gene**	**after pruning**
1	Homininae	African Apes (Hu, Ch, Go)	1,415	7,600	8	6	2
2	Hominidae	Great Apes (Hu, Ch, Go, Or)	1,424	7,849	34	32	7
3	Hominoidae	Apes (Hu, Ch, Go, Gi)	1,441	8,016	43	42	6
4	Catarrhini	Apes & Old World Monkeys (Hu, Ch, Go, Gi, Ma)	1,441	8,058	95	93	9

[a]: Hu: human, Ch: chimpanzee, Go: gorilla, Or: orangutan, Gi: gibbon, Ma: macaque

We found 6, 32, 42, and 93 such significant pathways in the Homininae, Hominidae, Hominoidae and Catarrhini branches respectively (Table 1 and Supplemental Table S1). The fact that we find more candidate pathways when we go further back in evolutionary history is not necessarily caused by a change of selective pressures, but could be due to an increased power to detect selection in more ancient branches (Studer et al. 2008; Gharib and Robinson-Rechavi 2013). In all four tested branches, we found clusters of candidate pathways that share a considerable proportion of their genes and often have similar biological functions, i.e. which are partially redundant (Supplemental Figures S2 to S5). We removed this redundancy with a 'pruning' method, by iteratively removing the genes of the top scoring set from the remaining sets and rerunning the testing procedure on these remaining sets. The pruning procedure considerably reduced the number of significant sets and yielded 2, 7, 6, and 9 independent candidate sets for the Homininae, Hominidae, Hominoidae and Catarrhini branches respectively (Table 2).

The strongest candidate for positive selection in the Homininae branch is the *GPCR downstream signaling* pathway, which is also a top scoring candidate in the Hominoidae branch. G protein-coupled receptors (GPCRs) are membrane proteins that regulate the cellular response to external signals such as neurotransmitters and hormones, and they play an important role in vision, taste and smell (Rosenbaum et al. 2009). The second candidate is the *Immunoregulatory interactions between a Lymphoid and a non-Lymphoid cell* pathway, which contains receptors and cell adhesion molecules that are important in immune response regulation of lymphocytes. This pathway is also significant after pruning in the Hominidae and Hominoidae branches, and it scores high before pruning in the Catarrhini branch.

Table 2 Results of the SUMSTAT gene set enrichment test. For each branch, only the pathways that score significant (q < 0.2) both before and after pruning (removal of overlapping genes) are listed. The SUMSTAT scores and gene set sizes that changed after pruning are shown in parentheses. Pathways which score significant on more than one branch are highlighted by the symbol '§'.

	SUMSTAT	*size*	*q*
Homininae	*(post-pruning)*	*(post-pruning)*	*post-pruning*
GPCR downstream signaling §	125.10	645	0.0032
Immunoregulatory interactions between a Lymphoid and a non-Lymphoid cell §	19.20 (18.61)	54 (53)	0.0047
			q
Hominidae	*score*	*size*	*post-pruning*
Olfactory Signaling Pathway	69.34	230	$<10^{-5}$
Immunoregulatory interactions between a Lymphoid and a non-Lymphoid cell §	22.34	57	$<10^{-5}$
Metabolism of xenobiotics by cytochrome P450 §	16.81	57	0.1422
Oxidative phosphorylation (WikiPathways)	14.43	46	0.1422
Intestinal immune network for IgA production	12.68 (11.68)	40 (36)	0.1932
Fatty acid metabolism §	13.84	47	0.1932
Synthesis of bile acids and bile salts via 7alpha-hydroxycholesterol	8.08 (7.16)	20 (17)	0.1932
			q
Hominoidae	*score*	*size*	*post-pruning*
Staphylococcus aureus infection §	23.88	44	$<10^{-5}$
GPCR downstream signaling §	155.59 (150.17)	687 (680)	$<10^{-5}$
Electron Transport Chain	30.13	84	0.0010
Complement cascade	16.36 (9.30)	29 (16)	0.0319
Metabolism of xenobiotics by cytochrome P450 §	20.11	59	0.0319
Immunoregulatory interactions between a Lymphoid and a non-Lymphoid cell §	22.34 (16.92)	58 (52)	0.1226
			q
Catarrhini	*score*	*size*	*post-pruning*
Hematopoietic cell lineage	40.12	79	$<10^{-5}$
Non-alcoholic fatty liver disease (NAFLD)	48.94 (47.02)	129 (124)	$<10^{-5}$
Cytokine-cytokine receptor interaction	95.90 (73.29)	236 (199)	$<10^{-5}$
Staphylococcus aureus infection §	24.11 (20.48)	43 (37)	0.0025
Chemical carcinogenesis	28.27 (26.87)	63 (62)	0.0066
Defensins	17.02 (16.47)	40 (37)	0.0896
Pancreatic secretion	29.93 (28.45)	85 (83)	0.1177
Fatty acid metabolism §	18.41	48	0.1814
NF-kB activation through FADD/RIP-1 pathway mediated by caspase-8 and -10	7.98 (6.62)	12 (10)	0.1898

Interestingly, the genes in this pathway that contribute most to the polygenic selection signal differ among branches (Figure 2), which suggests that while the same pathway has been under selection over a long time period, different genes underwent adaptive changes at different times. This is a pattern that is shared by many of the significant pathways (Supplemental Figure S6).

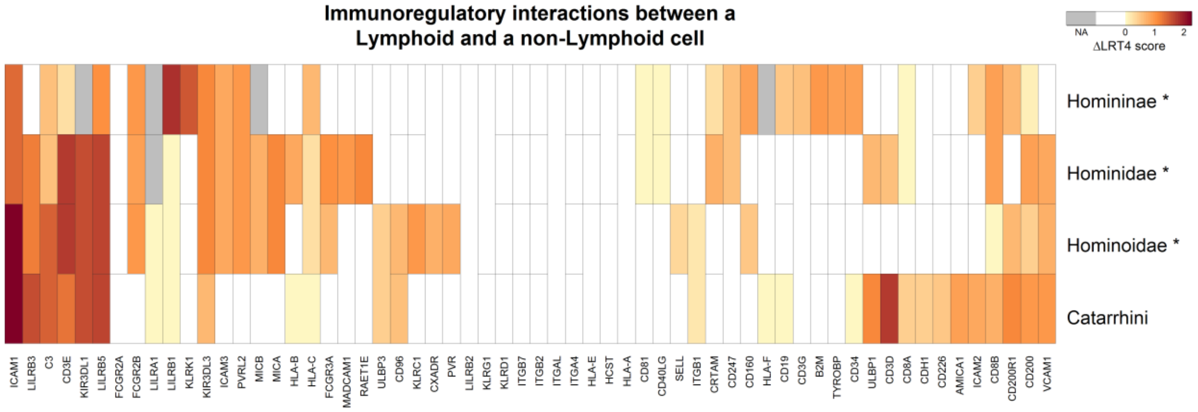


Figure 2: Heat map showing ΔlnL4 scores of genes in the *Immunoregulatory interactions between a Lymphoid and a non-Lymphoid cell* pathway for the four tested inner branches of the Primates tree. Branches where the pathway scores significant after pruning are marked with a '*'. The genes are grouped by hierarchical clustering to visualize blocks with similar signals within and among branches. Genes for which ΔlnL4 scores were not available (NA) in a certain branch are depicted in grey. Genes are merged (horizontally) with their paralog(s) into an 'ancestral gene' in the branches preceding a duplication and their scores were included only once in the calculation of the SUMSTAT score for these branches. Genes with (vertically) merged branches represent cases where the sequence of one or more species is missing or excluded, resulting in a single 'average' ΔlnL4 score over multiple branches. We use this score when testing each branch separately. The ΔlnL4 score is computed as the fourth root of log-likelihood ratio in the branch site test for positive selection.

The highest scoring candidate in the Hominidae branch is the *Olfactory Signaling Pathway*, i.e. smell perception. The pathway *Metabolism of xenobiotics by cytochrome P450*, which ranked third in Hominidae and fifth in Hominoidae, encodes detoxifying proteins which play a role in the metabolism of xenobiotics such as drugs and toxins; they include cytochrome P450 enzymes or glutathione S-transferases. Other candidate pathways are also involved in metabolism, such as the *Synthesis of bile acids and bile salts via 7alpha-hydroxycholesterol* and *Fatty acid metabolism*. Note that the latter pathway is also a candidate for positive selection in the Catarrhini branch. The remaining two top scoring pathways in the Hominidae branch are related to immune response (*Intestinal immune network for IgA production)* and electron transport (*Oxidative phosphorylation*).

In addition to the *Immunoregulatory interactions between a Lymphoid and a non-Lymphoid cell* pathway, two other immune response pathways remain significant after pruning in the Hominoidae branch. The highest scoring candidate, *Staphylococcus aureus infection*, includes genes coding for proteins used by the bacterium *S. aureus* to infect a host cell (Foster 2009). It is also a top candidate in the Catarrhini branch. The *Complement cascade* pathway plays an important role in immune surveillance and homeostasis (Ricklin et al. 2010). Finally, the last candidate in the Hominoidae branch, *Electron Transport Chain*, plays a role in energy production.

Five out of the nine pathways that remain significant after pruning in the Catarrhini branch have a function in immune response. Apart from the above mentioned *Staphylococcus aureus infection*, these are *Cytokine-cytokine receptor interaction*, *Hematopoietic cell lineage* (giving rise to various types of blood cells including leukocytes), the host defense peptides *Defensins*, and *NF-kB activation through FADD/RIP-1 pathway mediated by caspase-8 and -10*. The fourth candidate is *Non-alcoholic fatty liver disease (NAFLD)*, a disease linked with metabolic syndrome (Dowman et al. 2010). Note that before pruning, this latter pathway was part of a cluster of gene sets related to electron transport, a process which – when disturbed – can lead to oxidative stress, which is a key feature of NAFLD pathology. The remaining candidate pathways have functions in the metabolism of toxins (*Chemical carcinogenesis*) and nutrients (*Pancreatic secretion*).

Discussion

In this study we have combined the strength of two approaches to detect positive selection in the ancestral lineages of humans. The branch-site test identifies episodes of selection at specific evolutionary times and sites in a protein, and gene set enrichment combines the signal of multiple genes to find selection at the pathway level. We identified many significant pathways related to immune response, sensory perception, metabolism, and electron transport in different branches of the primate tree (Supplemental Figures S2 to S5). These pathways were often organized in clusters that share many genes and have similar biological functions. After removal of the overlap between these pathways with a pruning procedure, two to nine pathways remained significant per branch. These pathways are our prime candidates for having been shaped by positive selection in primate evolution, and correspond to four broad biological processes.

First, in all four tested branches, immune response related pathways were among the top candidates, both before and after pruning. This strong and enduring signal suggests that the ongoing challenge of adaptation to changing pathogens has been one of the major selective pressures in primate evolution. This is in line with similar findings in previous reports of selection or fast evolution in ancient (Chimpanzee Sequencing and Analysis Consortium 2005; Nielsen et al. 2005) and recent human evolution (Daub et al. 2013).

Second, several significant pathways are involved in sensory perception, with *GPCR downstream signaling* and the *Olfactory Signaling Pathway* remaining after pruning in Homininae and Hominidae respectively. Sensory perception pathways have many functions, from sensing environmental signals to internal signals such as hormones and neurotransmitters. Earlier studies have reported genes involved in sensory perception to be evolving rapidly in humans and other primates (Chimpanzee Sequencing and Analysis Consortium 2005; Nielsen et al. 2005; Arbiza et al. 2006; Kosiol et al. 2008), with at least some evidence for positive selection (Chimpanzee Sequencing and Analysis Consortium 2005; Nielsen et al. 2005; Arbiza et al. 2006; Kosiol et al. 2008), while our gene set study of human-specific evolution found them rather under relaxed selection (Daub et al. 2015). These genes could have been affected by positive selection because of changes in environment, behavior or diet.

Third, clusters of significant pathways involved in the metabolism of lipids and other nutrients were detected in Hominidae, Hominoidae and Catarrhini. This selective signal could be explained by changes in diet. Of note, the selective signal could also be due in part to the involvement of some of these pathways (*Metabolism of xenobiotics by cytochrome P450* and *Chemical carcinogenesis*) in the metabolism of potentially toxic xenobiotics. These results are specifically interesting, since adaptations in metabolism are expected with the changes in lifestyle that have marked hominid evolution, but yet have been rarely detected (although see (Mathieson et al. 2015)). The gene set approach thus allows us here to capture a subtle but biologically important signal of adaptation.

Fourth, there are several significant pathways involved in energy production in the Hominidae, Hominoidae and Catarrhini branches, with *Oxidative phosphorylation*, *Electron Transport Chain*, and *Non-alcoholic fatty liver disease (NAFLD)* remaining significant after pruning in these branches. Earlier candidate gene studies have also reported genes involved in electron transport to be under positive selection in anthropoid primates (reviewed in Grossman et al. 2004). It has been suggested that this could be related to anthropoid specific evolutionary changes, such as extended lifespan, prolonged fetal development and enlarged neocortex, as these traits require increased aerobic energy production (Grossman et al. 2004).

Several pathways reported here were also detected in an earlier study, in which we applied a gene set enrichment analysis on a more recent period of human evolutionary history (Daub et al. 2013). The *Cytokine-cytokine receptor interaction* pathway for example was detected as a potential candidate for recent polygenic human adaptation. Another candidate in that study, *Fatty Acid Beta oxidation*, scored significant here before pruning in Catarrhini, as did the *Malaria* pathway in all branches except Homininae. These findings suggest that certain functions have been under selection in multiple or ongoing episodes of positive selection, and might still be under selection now.

The fact that some pathways are significant over successive evolutionary periods (neighboring branches) could indicate that similar selective pressures occurred over a long time frame in primates. Interestingly, we found that in such cases the highest scoring genes often differed among branches (Figure 2 and Supplemental Figure S6), which suggests that biological pathways under long-term selective pressures have adapted by means of changes in different genes over time, allowing to fine tune the function of the pathway without altering previous adaptations. However, we cannot exclude that we might lack power to detect more continuity in selective pressure in the same gene over long evolutionary periods.

We have also identified several candidate pathways that are branch-specific. These isolated adaptive signals may be due to unique changes in selective pressure; they might also be due, at least partly, to methodological reasons. As noted above, we find more significant pathways in more ancient lineages, which is probably due to an increased power to evidence selection in older branches, a known property of the branch-site test (Studer et al. 2008; Gharib and Robinson-Rechavi 2013). A potential limitation of our study is that some genes were not present in the Selectome database, as only gene trees with at least six leaves and passing stringent alignment quality filters were included (Moretti et al. 2014). Therefore about 21% of genes were ignored in our enrichment test and we might have a deficit of trees with fast evolving genes that are difficult to align, as well as a potential excess of gene trees with duplications. The latter category contains several paralogs resulting in more leaves, and thus pass the criterion of six leaves even when a few sequences are missing from genomes or eliminated by alignment quality filters. To estimate the effect of this exclusion on our results, we tested for each gene set whether they had a significant excess of excluded genes (q-value < 0.2 with Fisher Exact test). We found 62 gene sets with such an excess (Supplemental Figure S7). These excluded gene sets are mainly involved in protein metabolism (15 sets) and the cell cycle (37 sets). Interestingly, we also find a few gene sets that contain many excluded genes, but still score high in our likelihood-based test, namely the *Olfactory signaling*, *Olfactory transduction* and *GPCR downstream signaling* pathways. Some of the pathways with an excess of excluded genes were candidates for being under selection in studies of recent human evolution, such as *Pathogenic Escherichia coli infection* (Daub et al. 2013), and we could thus not fully check if these pathways have also been under selection in more ancient primate evolution.

We also investigated other non-adaptive factors that might have affected our results, like a very small but significant correlation between ΔlnL scores and gene tree size (number of branches; Pearson's r = 0.05, p < 2.2e-16) or sequence length (r = -0.01, p = 8e-10). However, we found a more noticeable correlation between ΔlnL scores and branch length (r = 0.24, p < 2.2e-16). This makes sense since the branch length is inferred from the number of mutations, and with increasing numbers of mutations we have more power to detect positive selection, and positive selection can also lead to even more mutations being fixed. So our candidate pathways could be enriched in fast evolving genes, partly because positively selected genes evolve fast, and partly because there is more power to detect selection in fast evolving genes.

We want to emphasize that our gene set enrichment analysis is related to but differs from classical Gene Ontology (GO) enrichment tests, as we include all genes in our analysis instead of only considering top scoring genes after setting an arbitrary significance threshold. For comparison we also performed a GO enrichment test on the significant genes in each branch, but this procedure did not result in any significantly enriched GO terms. This negative result underlines the power of a gene set enrichment test as performed in this study, as we are able to detect pathways that contain many genes with small to moderate effect mutations.

In conclusion, by combining the specificity of the branch-site test and the power of the gene set approach, we have been able to uncover for the first time strong signals of polygenic positive selection in several biological processes during long-term primate evolution. In addition to immune response, we find evidence for adaptive evolution on sensory perception, as well as on metabolism and energy production. The fact that different genes are involved in pathways showing signals of positive selection in several branches, suggests that the fine tuning of biological functions can change over time during primate evolution. Our results allow us to bridge the gap between studies of selection in deep mammalian evolution, and recent adaptation in the human lineage.

Material and Methods

Data collection

In this study we aim at detecting biological pathways affected by episodic positive selection in primate evolutionary history, specifically in the four inner branches of the Primates tree that lead to modern humans (Figure 1).

Branch-site likelihood test

The data used in this study was produced as part of release 6 of Selectome (http://selectome.unil.ch/, Proux et al. 2009; Moretti et al. 2014), which is a database that provides results of the branch-site likelihood test for positive selection (Zhang et al. 2005) on internal branches of several clades. The branch-site test can detect codon sites on specific phylogenetic branches that are affected by episodic positive selection. In short, it estimates the rate of non-synonymous (dN) and synonymous (dS) nucleotide substitutions to assess differences in selective pressure (dN/dS ratio) among branches and over sites. Usually, tested branches that have a class of sites with a dN/dS ratio $\omega_2 > 1$ are candidates for positive selection. While the strength of positive selection can in principle be estimated by the ratio ω_2 or by the proportion p_2 of sites in this class, we have found the likelihood ratio to be a good estimator of the evidence for positive selection (Studer et al. 2008; Roux et al. 2014). In more detail, for each branch, the maximum likelihood of the data is estimated under two models: one that allows for positive selection (H1), and one that only allows negative selection and neutral evolution (H0), and a log-likelihood ratio statistic $\Delta\ln L = 2(\ln LH1-\ln LH0)$ is computed. To determine their significance, the $\Delta\ln L$ values are usually compared to a chi-square (χ^2) distribution with one degree of freedom; here we use all $\Delta\ln L$ values without applying any *a priori* significance cut-off. Only the internal branches of gene families with at least six sequences (leaves in the tree) were computed in Selectome.

For our gene set analysis, we thus used the ΔlnL values obtained by testing 15,738 gene trees from the Primates clade as defined in version 70 of Ensembl Compara (Vilella et al. 2009). We only kept test scores for the four branches mentioned above (and shown in Figure 1) if they led to at least one human gene. We could thus create for each branch *i* (Figure 1), a list G_i with human genes and their corresponding ΔlnL scores. Because of duplication events and the lack of resolution of the Homininae label, these initial lists often contained several rows per gene, whereas we need at most one ΔlnL score per gene per branch for our enrichment test. On the other hand, missing or excluded sequences can result in branches lacking ΔlnL scores in a number of genes. We describe below how we dealt with these situations and how we handled computational issues with likelihood estimation.

Multiple Homininae branches

In the Ensembl species tree, the Homininae branch is followed by a trifurcation leading to the human, chimp and gorilla leaves (Figure 1), lacking a label for the Hominini branch that leads to human and chimp. Since the gene trees in Ensembl are always bifurcating, there are many trees that have multiple branches annotated as Homininae. For example, the *DMXL1* gene family (Supplemental Figure S8) has a Homininae branch leading to the ancestor of human, chimp and gorilla followed by a branch leading to the ancestor of human and chimp which is labeled "Homininae", although it is properly an Hominini branch. In about 42% of the gene trees we found such multiple Homininae labeled branches. In these cases we took the test scores of the oldest branch, which in 95% of the cases is also the longest branch.

Missing branches

For several genes the sequence of one or more species is absent or excluded due to low quality (Moretti et al. 2014). The corresponding branch in the gene tree is then merged with its downstream branch, and the ΔlnL score is assigned to this lower branch, while it actually represents an 'average' score over both branches. We therefore use this ΔlnL score for both branches as input in the gene set enrichment test. Supplemental Figure S9 shows an example of the C3 gene tree with a missing macaque sequence. Supplemental Table S2 shows results of the gene set enrichment test where in the case of a missing sequence, we assign the ΔlnL score only to the lower branch.

Gene duplications

Gene duplications will result in some species having paralogous genes. For our gene set enrichment analysis we removed the branches in the Primates tree that led to a duplication event (about 3% of all branches). Our procedure is further detailed in Supplemental Figure S10. Briefly, for branches predating the duplication, the values in the gene list corresponding to the duplicated genes, which are redundant (same ancestral branches reported for each paralog), were merged and replaced by one value, while for the branches after the duplication both values were kept, as each represents a different paralog.

Non-converging likelihoods

The numerical optimization (BFGS algorithm, see PAML documentation) does not always converge to the maximum likelihood estimates of H0 or H1. This can result in negative ΔlnL values or in false positives having extreme high ΔlnLs. In order to reduce the number of events of non-convergence, we ran the branch-site test two more times on the whole Primates tree, thus yielding a total of three likelihood scores for each of H0 and H1. We then selected for each branch the highest log-likelihoods for H0 and H1 among the three runs ($\mathrm{lnLH0}_g$ and $\mathrm{lnLH1}_g$) and constructed the log-likelihood ratio score for a gene ($\Delta\mathrm{lnL}_g$) as follows:

$$\Delta\mathrm{lnL}_g = 2\left(\mathrm{lnLH1}_g - \mathrm{lnLH0}_\mathrm{g}\right) = 2(\max_{i=1,2,3}(\mathrm{lnLH1}_i) - \max_{i=1,2,3}(\mathrm{lnLH0}_i)) \qquad (1)$$

However, three runs could still allow for some non-convergence. If we obtained a negative $\Delta\mathrm{lnL}_g$ score (indication of non-convergence of all lnLH1 scores) we set $\Delta\mathrm{lnL}_g$ to zero (about 8% of the cases, with less than 0.06% having a $\Delta\mathrm{lnL}_g < -0.1$).

The first branch-site test was run with CODEML version 4.6 from PAML (Yang 2007). Most jobs of the first run were performed on the Swiss multi-scientific computing grid (SMSCG, http://lsds.hesge.ch/smscg/), while the longer jobs were submitted to the Vital-IT computer cluster (http://www.vital-it.ch) and to the Ubelix computer cluster of the university of Berne (http://www.ubelix.unibe.ch/). The second run was run with SlimCodeml (Schabauer et al. 2012), and the third with FastCodeML (Valle et al. 2014) starting with M1-estimated (κ and ω_0) starting parameters, both on the Vital-IT computer cluster.

Ensembl Gene ID to Entrez Gene ID conversion

We use gene sets from NCBI Biosystems (Geer et al. 2010) (see next section). Since these sets are annotated with Entrez gene IDs, whereas Selectome uses Ensembl gene IDs, we created a one-to-one Ensembl–Entrez conversion table, to map the Ensembl gene IDs in the branch specific gene tables (G_i) to Entrez gene IDs. First, we started with a gene list (G_{entrez}) containing 20,016 protein coding human genes located either on the autosomal, X or Y chromosomes, downloaded from the NCBI Entrez Gene (Maglott et al. 2011) website (http://www.ncbi.nlm.nih.gov/gene, accessed on July 16, 2014). We further collected conversion tables (often containing one-to-many or many-to-many mappings) from HGNC (http://www.genenames.org/biomart/, accessed on July 16, 2014), NCBI (ftp://ftp.ncbi.nih.gov/gene/DATA, accessed on July 16, 2014), and Ensembl (version 70, http://jan2013.archive.ensembl.org/biomart/martview). We only kept the rows in these conversion tables that contained genes from G_i and from G_{Entrez}. Next, we merged the three tables to one, and only uniquely mapped genes (Ensembl ID – Entrez ID) were used further. For each Ensembl ID we kept the Ensembl-Entrez mapping with the highest count (in case of multiple candidates we chose randomly one) and we then repeated this procedure for each Entrez ID. With this final list of unique one-to-one Ensembl-Entrez ID mappings we translated the Ensembl genes in the G_i tables to Entrez IDs, and unmapped genes were removed. The resulting

14,574, 15,026, 15,375 and 15,450 genes in the tables G_1, G_2, G_3 and G_4 respectively were used for further analyses.

Gene sets

We downloaded a list of 2,609 human gene sets of type 'pathway' from the NCBI Biosystems (Geer et al. 2010) database (http://www.ncbi.nlm.nih.gov/biosystems, accessed on July 16, 2014). The Biosystems database is a repository of gene sets collected from other pathway databases, such as BioCyc (Caspi et al. 2014), KEGG (Kanehisa and Goto 2000; Kanehisa et al. 2014), The National Cancer Institute Pathway Interaction Database (Schaefer et al. 2009), Reactome (Croft et al. 2014) and Wikipathways (Kelder et al. 2012).

For each primate branch of interest (Homininae, Hominidae, Hominoidae and Catarrhini) (Figure 1), we excluded genes that could not be mapped to the corresponding gene list G_i (see previous section), then removed gene sets with less than 10 genes, because the gene set enrichment test has low power to detect selection in small sets. We merged groups of nearly identical gene sets (i.e. sets that share 95% or more of their genes) into single gene sets, i.e. the union of all gene sets in these groups. In the text, the name if these union sets is followed by an asterisk ('*'). To distinguish gene sets with identical names, their source database is added to their name. After the filtering process, we obtained S_1=1,415, S_2=1,424, S_3=1,441 and S_4=1,441 gene sets for the four branches to be tested for selection with the gene set enrichment analysis (Table 1; numbering according to Figure 1). Note that for each branch we use a different gene list (G_i), leading to a different number of gene sets, as we condition on a minimum of ten genes per set for each branch.

Data analysis

Test for polygenic selection

We used a gene set enrichment approach to test for polygenic signals of positive selection on the four primate branches Catarrhini, Hominoidae, Hominidae and Homininae. We first calculated for each gene set its SUMSTAT score, which is the sum of selection scores of genes in the set of interest (Tintle et al. 2009; Daub et al. 2013). As selection score we took the fourth-root of the $\Delta \ln L_g$ values (called ΔlnL4 hereafter) to ensure that the distribution of non-zero ΔlnLs is approximately Normal (Hawkins and Wixley 1986; Roux et al. 2014). This procedure also prevents extreme scoring genes from getting too much weight in the test, which would otherwise result in significant pathways mostly due to a few outlier genes. The SUMSTAT score of a gene set *s* is then simply calculated as:

$$\mathrm{SUMSTAT}_s = \sum_{g \in s} \sqrt[4]{\Delta \ln \mathrm{L}_g} \qquad (1)$$

The R code to run the gene set enrichment pipeline together with detailed examples and data sets is freely available from https://github.com/CMPG/polysel.

Empirical null distribution

The significance of the SUSMTAT score of a pathway was inferred by creating a null distribution of random gene sets of identical size and calculating SUMSTAT on these random sets. The null distribution was built by sampling genes at random from all genes in G_i that belonged to at least one gene set. To improve computation time, we created the null distribution with a sequential random sampling method (Ahrens and Dieter 1985), which avoids the burden of high precision p-value estimation for low scoring, and thus for us uninteresting, gene sets. For this we first tested all sets against a small null distribution with 10,000 random sets and estimated their p-value. For those sets with a p-value < 0.5 we expanded the null distribution with another 10,000 randomizations. This process was continued with decreasing p-value thresholds ($p < \frac{1}{2i}$ at the *i*th iteration) until we reached a maximum of 1,000,000 randomizations.

We have considered using a parametric distribution, as we know from theory and simulation studies (Zhang et al. 2005) that under the null hypothesis the log-likelihood ratios in the branch-site test should be distributed as a mixture of 50% zeros and 50% a χ^2 distribution with one degree of freedom. Assuming such a distribution, we can infer the expected null distribution of a log-likelihood ratio score at the gene set level as well. However, in our real data the proportion of zeros is much larger (70%-80%), and we observed that a parametric null distribution produces a skewed distribution of p-values, leading to an overestimation of p-values (see also Supplemental Text S1) and thus an under-estimation of the signal for positive selection. Therefore we have favored the use of an empirical distribution.

Removal of sets with outlier genes

Gene sets can potentially have a high SUMSTAT score due solely to one gene with an extremely high ΔlnL value. However, we are interested here in gene sets affected by polygenic selection, where multiple genes have moderately high selection scores. Therefore, we also tested the gene sets after removing their highest scoring gene, and contrasted their SUMSTAT score against random sets that also had their top scoring gene removed. Those gene sets that were not scoring significant anymore after this test were not included in further analyses.

Removing redundancy in overlapping gene sets ('pruning')

The gene set enrichment test can result in partially redundant gene sets being called significant, because they share high scoring genes, and BioSystems includes overlapping or redundant sets. We therefore removed the overlap between gene sets with a 'pruning' method similar to one we described in a previous study (Daub et al. 2013). In short, we removed for each branch the genes of the most significant pathway from all the other pathways, and ran the enrichment test on these updated gene sets. We repeated this pruning procedure until no sets were left to be tested.

We estimated the False Discovery Rate (FDR) in our results empirically, since the tests in the pruning procedure are not independent and the results are biased toward low p-values (only the high scoring sets will remain after pruning). To estimate the FDR, we repeatedly (N=300) permuted ΔlnL4 scores among genes that are part of a set, and tested the gene sets including pruning. For each observed p-value $p*$ in our original results, we can estimate the FDR (if we would reject all hypotheses with a p-value ≤ $p*$) with:

$$F\hat{D}R(p^*) = \frac{\pi_0 \cdot \hat{V}(p^*)}{R(p^*)}\,, \tag{2}$$

where π_0 is the proportion of true null hypotheses, $\hat{V}(p^*)$ is the estimated number of rejected true null hypotheses if all hypotheses are true nulls and $R(p*)$ is the total number of rejected hypotheses. We conservatively set π_0 = 1, and estimated $\hat{V}(p^*)$ from the mean proportion of gene sets in the randomized data sets with p-value ≤ $p*$. The q-value was finally determined by taking the lowest estimated FDR among all observed p-values ≥ $p*$. We reported the gene sets that scored significant (q-value < 0.2) both before and after pruning.

Classical GO enrichment test

To contrast our results with a classical Gene Ontology (GO) enrichment test, we defined for each branch a list of top scoring genes, namely those genes with a ΔlnL value resulting in a q-value < 0.2 when compared against a null distribution consisting of 50% zeros and 50% χ^2 with one degree of freedom (see paragraph 'Empirical null distribution' and Zhang et al. 2005). Note that this is a much less conservative score compared to the branches reported on the Selectome website, as the latter are based on q < 0.10 and were computed over all branches and trees using a more conservative null distribution (χ^2 with one degree of freedom). The top genes thus defined were used as input for the online tool Fatigo (http://babelomics.bioinfo.cipf.es, (Al-Shahrour et al. 2004)). We tested for enrichment in (level 3-9) GO biological processes and molecular functions using as background the gene lists G_i. The resulting p-values were corrected for multiple testing by calculating the q-value per branch and GO category (biological process and molecular function), where q-values < 0.2 were considered significant.

Test for bias in genes filtered from Selectome

Since all gene trees with less than six leaves after removing unreliably aligned sequences were excluded in Selectome, we tested which categories of gene sets where enriched with these excluded genes, and were thus unlikely to show significant results in our gene set enrichment test. For this we performed for each gene set a Fisher's exact test on a contingency table with the counts of included and excluded genes in the set contrasted against the same counts for the rest of the genes in G_{entrez}, the list of genes downloaded from the NCBI Entrez website. The resulting p-values were corrected for multiple tests, and gene sets with a q-value < 0.2 were reported.

Correcting for multiple testing

For all tests (except when inferring significance after pruning), we calculated the q-value (Storey J. D. and Tibshirani 2003; Storey John D. et al. 2004) as a measure of the false discovery rate using the R package *qvalue* (with parameter *pi0.method* set to "bootstrap") and reported those gene sets with a q-value < 0.2.

Enrichment maps

The enrichment maps (Supplemental Figures S2 to S5 and S7) were created in Cytoscape v. 2.8.3 with the *Enrichmentmap* plugin (Merico et al. 2010).

Acknowledgements

This work was supported by the Swiss National Science Foundation (grant numbers PDFMP3-130309 to LE and 31003A_153341 and CR32I3_143768 to MRR). The computations were performed at the Vital-IT (http://www.vital-it.ch) Center for high-performance computing of the SIB Swiss Institute of Bioinformatics, on the Ubelix HPC cluster of the University of Bern and on the Swiss Multi-Science Computing Grid.

References

Ahrens JH, Dieter U. 1985. Sequential random sampling. *ACM Trans Math Softw* 11:157-169.

Al-Shahrour F, Diaz-Uriarte R, Dopazo J. 2004. FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes. *Bioinformatics* 20:578-580.

Amorim CE, Daub JT, Salzano FM, Foll M, Excoffier L. 2015. Detection of convergent genome-wide signals of adaptation to tropical forests in humans. *PLoS One* 10:e0121557.

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

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

Bakewell MA, Shi P, Zhang J. 2007. More genes underwent positive selection in chimpanzee evolution than in human evolution. *Proc Natl Acad Sci U S A* 104:7489-7494.

Berg JJ, Coop G. 2014. A population genetic signal of polygenic adaptation. *PLoS Genet* 10:e1004412.

Carneiro M, Rubin CJ, Di Palma F, Albert FW, Alfoldi J, Barrio AM, Pielberg G, Rafati N, Sayyab S, Turner-Maier J et al. 2014. Rabbit genome analysis reveals a polygenic basis for phenotypic change during domestication. *Science* 345:1074-1079.

Caspi R, Altman T, Billington R, Dreher K, Foerster H, Fulcher CA, Holland TA, Keseler IM, Kothari A, Kubo A et al. 2014. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. *Nucleic Acids Res* 42:D459-471.

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

Croft D, Mundo AF, Haw R, Milacic M, Weiser J, Wu G, Caudy M, Garapati P, Gillespie M, Kamdar MR et al. 2014. The Reactome pathway knowledgebase. *Nucleic Acids Res* 42:D472-477.

Daub JT, Dupanloup I, Robinson-Rechavi M, Excoffier L. 2015. Inference of Evolutionary Forces Acting on Human Biological Pathways. *Genome Biol Evol* 7:1546-1558.

Daub JT, Hofer T, Cutivet E, Dupanloup I, Quintana-Murci L, Robinson-Rechavi M, Excoffier L. 2013. Evidence for Polygenic Adaptation to Pathogens in the Human Genome. *Molecular Biology and Evolution* 30:1544-1558.

Delport W, Scheffler K, Seoighe C. 2009. Models of coding sequence evolution. *Brief Bioinform* 10:97-109.

Dowman JK, Tomlinson JW, Newsome PN. 2010. Pathogenesis of non-alcoholic fatty liver disease. *QJM* 103:71-83.

Foll M, Gaggiotti OE, Daub JT, Vatsiou A, Excoffier L. 2014. Widespread signals of convergent adaptation to high altitude in Asia and america. *Am J Hum Genet* 95:394-407.

Foster TJ. 2009. Colonization and infection of the human host by staphylococci: adhesion, survival and immune evasion. *Vet Dermatol* 20:456-470.

Fraser HB. 2013. Gene expression drives local adaptation in humans. *Genome Res* 23:1089-1096.

Geer LY, Marchler-Bauer A, Geer RC, Han L, He J, He S, Liu C, Shi W, Bryant SH. 2010. The NCBI BioSystems database. *Nucleic Acids Res* 38:D492-496.

Gharib WH, 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. *Molecular Biology and Evolution* 30:1675-1686.

Gibbs RA, Rogers J, Katze MG, Bumgarner R, Weinstock GM, Mardis ER, Remington KA, Strausberg RL, Venter JC, Wilson RK et al. 2007. Evolutionary and biomedical insights from the rhesus macaque genome. *Science* 316:222-234.

Grossman LI, Wildman DE, Schmidt TR, Goodman M. 2004. Accelerated evolution of the electron transport chain in anthropoid primates. *Trends Genet* 20:578-585.

Hawkins DM, Wixley RAJ. 1986. A Note on the Transformation of Chi-Squared Variables to Normality. *American Statistician* 40:296-298.

Kanehisa M, Goto S. 2000. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* 28:27-30.

Kanehisa M, Goto S, Sato Y, Kawashima M, Furumichi M, Tanabe M. 2014. Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic Acids Res* 42:D199-205.

Kelder T, van Iersel MP, Hanspers K, Kutmon M, Conklin BR, Evelo CT, Pico AR. 2012. WikiPathways: building research communities on biological pathways. *Nucleic Acids Res* 40:D1301-1307.

Kosiol C, Vinar T, da Fonseca RR, Hubisz MJ, Bustamante CD, Nielsen R, Siepel A. 2008. Patterns of positive selection in six Mammalian genomes. *PLoS Genet* 4:e1000144.

Lachance J, Tishkoff SA. 2013. Population Genomics of Human Adaptation. *Annu Rev Ecol Evol Syst* 44:123-143.

Maglott D, Ostell J, Pruitt KD, Tatusova T. 2011. Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res* 39:D52-57.

Mathieson I, Lazaridis I, Rohland N, Mallick S, Patterson N, Roodenberg SA, Harney E, Stewardson K, Fernandes D, Novak M et al. 2015. Genome-wide patterns of selection in 230 ancient Eurasians. *Nature* 528:499-503.

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

Merico D, Isserlin R, Stueker O, Emili A, Bader GD. 2010. Enrichment map: a network-based method for gene-set enrichment visualization and interpretation. *PLoS One* 5:e13984.

Moretti S, Laurenczy B, Gharib WH, Castella B, Kuzniar A, Schabauer H, Studer RA, Valle M, Salamin N, Stockinger H et al. 2014. Selectome update: quality control and computational improvements to a database of positive selection. *Nucleic Acids Res* 42:D917-921.

Nielsen R, Bustamante C, Clark AG, Glanowski S, Sackton TB, Hubisz MJ, Fledel-Alon A, Tanenbaum DM, Civello D, White TJ et al. 2005. A scan for positively selected genes in the genomes of humans and chimpanzees. *PLoS Biol* 3:e170.

Proux E, Studer RA, Moretti S, Robinson-Rechavi M. 2009. Selectome: a database of positive selection. *Nucleic Acids Res* 37:D404-407.

Ricklin D, Hajishengallis G, Yang K, Lambris JD. 2010. Complement: a key system for immune surveillance and homeostasis. *Nat Immunol* 11:785-797.

Rosenbaum DM, Rasmussen SG, Kobilka BK. 2009. The structure and function of G-protein-coupled receptors. *Nature* 459:356-363.

Roux J, Privman E, Moretti S, Daub JT, Robinson-Rechavi M, Keller L. 2014. Patterns of positive selection in seven ant genomes. *Mol Biol Evol* 31:1661-1685.

Schabauer H, Valle M, Pacher C, Stockinger H, Stamatakis A, Robinson-Rechavi M, Ziheng Y, Salamin N. 2012. SlimCodeML: An Optimized Version of CodeML for the Branch-Site Model. Pp. 706-714. Parallel and Distributed Processing Symposium Workshops & PhD Forum (IPDPSW), 2012 IEEE 26th International.

Schaefer CF, Anthony K, Krupa S, Buchoff J, Day M, Hannay T, Buetow KH. 2009. PID: the Pathway Interaction Database. *Nucleic Acids Res* 37:D674-679.

Serra F, Arbiza L, Dopazo J, Dopazo H. 2011. Natural selection on functional modules, a genome-wide analysis. *PLoS Comput Biol* 7:e1001093.

Storey JD, Taylor JE, Siegmund D. 2004. Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 66:187-205.

Storey JD, Tibshirani R. 2003. Statistical significance for genomewide studies. *Proc Natl Acad Sci U S A* 100:9440-9445.

Studer RA, Penel S, Duret L, Robinson-Rechavi M. 2008. Pervasive positive selection on duplicated and nonduplicated vertebrate protein coding genes. *Genome Res* 18:1393-1402.

Tintle NL, Borchers B, Brown M, Bekmetjev A. 2009. Comparing gene set analysis methods on single-nucleotide polymorphism data from Genetic Analysis Workshop 16. *BMC Proceedings* 3 Suppl 7:S96.

Valle M, Schabauer H, Pacher C, Stockinger H, Stamatakis A, Robinson-Rechavi M, Salamin N. 2014. Optimization strategies for fast detection of positive selection on phylogenetic trees. *Bioinformatics* 30:1129–1137.

Vilella AJ, Severin J, Ureta-Vidal A, Heng L, Durbin R, Birney E. 2009. EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. *Genome Res* 19:327-335.

Wellenreuther M, Hansson B. 2016. Detecting Polygenic Evolution: Problems, Pitfalls, and Promises. *Trends in Genetics* 32:155–164.

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

Zhang J, Nielsen R, Yang Z. 2005. Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. *Mol Biol Evol* 22:2472-2479.

Detection of pathways affected by positive selection in primate lineages ancestral to humans

Daub, J.T., Moretti, S., Davidov, I. I., Excoffier, L. Robinson-Rechavi, M.

Supporting Information

Supplemental Text

Supplemental Tables

Supplemental Figures

Text S1

Using a parametric null distribution

If we assume independence between genes in a set, we can define the gene set level likelihoods, $\mathrm{LH0}_s$ and $\mathrm{LH1}_s$, as the product of the gene level likelihoods:

$$\mathrm{LH0}_s = \prod_{g \in set} \mathrm{LH0}_g, \quad \mathrm{LH1}_s = \prod_{g \in set} \mathrm{LH1}_g$$

We can then calculate a gene set level likelihood ratio test score ($\Delta\mathrm{lnL}_s$), which is equal to the sum of $\Delta\mathrm{lnL}_g$ scores of genes in the gene set:

$$\Delta \ln \mathrm{L}_s = 2\left(\sum_{g \in s} \ln \mathrm{LH1}_g - \sum_{g \in s} \ln \mathrm{LH0}_g \right) = \sum_{g \in s} \Delta \ln \mathrm{L}_g$$

This $\Delta\mathrm{lnL}_s$ score can then be compared against a parametric null distribution to infer its significance.

In the branch-site test for positive selection, the dN/dS ratio for the class of sites under positive selection (ω_2) is constrained by $\omega_2 \geq 1$ and the null model H_0 is nested in H_1 with ω_2 fixed on the boundary of the parameter space ($\omega_2 = 1$). Therefore, in case of no positive selection (H_0 is true), we would expect that the ΔlnL values at the gene level ($\Delta\mathrm{lnL}_g$) are distributed as a mixture of 50% zeros and 50% (Self and Liang 1987; Zhang et al. 2005). For the gene set level ΔlnL values ($\Delta\mathrm{lnL}_s$), we have *N* free parameters for a gene set of size *N*. The expected null distribution of $\Delta\mathrm{lnL}_s$ would be a mixture of χ^2 distributions with 0, 1, ... , *N* degrees of freedom, where the mixture weights are the binomial coefficients (Ota et al. 2000). For example, with gene sets containing three genes, the distribution would be:

$$\tfrac{1}{8}\chi_0^2 + \tfrac{3}{8}\chi_1^2 + \tfrac{3}{8}\chi_2^2 + \tfrac{1}{8}\chi_3^2$$

The probability for observing a ΔlnL$_s$ of x or higher in gene sets of size three can then be calculated as:

$$\Pr(\Delta \ln \mathrm{L}_s > x) = \tfrac{1}{8}\Pr(\chi_0^2 > x) + \tfrac{3}{8}\Pr(\chi_1^2 > x) + \tfrac{3}{8}\Pr(\chi_2^2 > x) + \tfrac{1}{8}\Pr(\chi_3^2 > x)$$

The ΔlnLs scores that we use have a higher proportion of zero's, namely around 70-80%. Testing our gene sets against a parametric null distribution, results therefore in unacceptable p-value distributions, which is shown by the following qq-plots comparing the p-values from a parametric test with an expected uniform distribution:

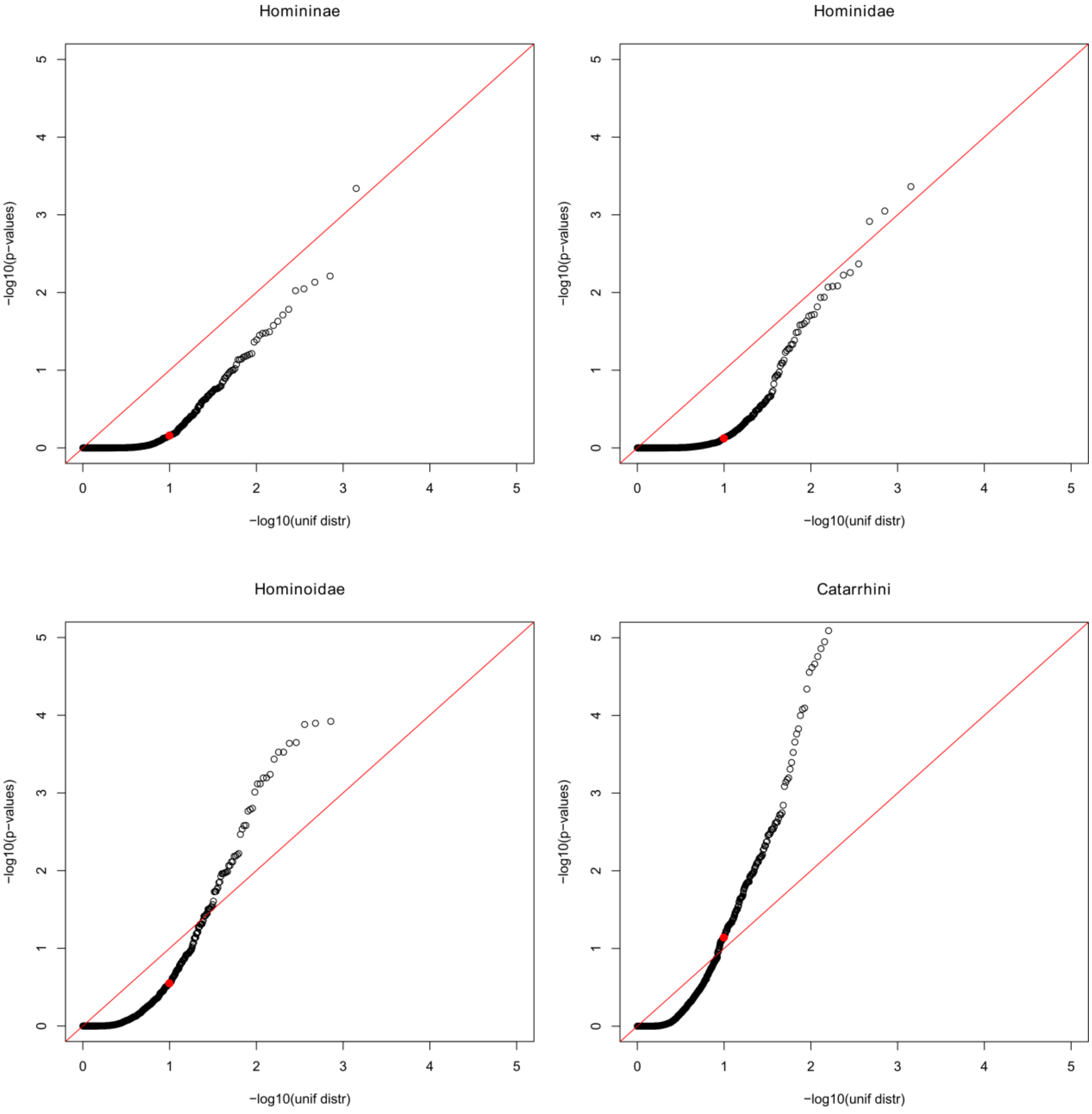

Homininae
Hominidae
Hominoidae
Catarrhini
−log10(p−values)
−log10(unif distr)
0
1
2
3
4
5

Table S1

This table is a separate excel file (TableS1.xlsx) which contains three tables: (A) Top scoring gene sets in the gene set enrichment test before pruning, their rank and score and p- and q-value (also after leaving out the highest scoring gene). All gene sets with $q<0.2$ are reported. (B) All significantly scoring pathways and their rank (before pruning) in all four branches. An '-' indicates that the set was not defined in that particular branch. Gene sets that scored significant after pruning are marked in bold. (C) Genes in top scoring gene sets and their ΔlnL4 and ΔlnL scores.

Table S2

This table is a separate excel file (TableS2.xlsx) presenting results of the gene set enrichment test where missing branches were not given the ΔlnL4 score of their child branch. It contains three tables: The number of available genes and gene sets for each branch excluding or including missing branches (using the ΔlnL4 score of their child branch) (A), and two tables with top scoring gene sets in the enrichment test before (B) and after (C) pruning, their rank and score and p- and q-value. All gene sets with q-value<0.2 are reported.

Figure S1

Proportion (A) and distribution (B) of non-zero ΔlnL4 scores in the genes in the four tested branches, split between genes that are not part of a gene set and genes that are part of at least one gene set. Especially the Homininae and Hominidae branches show a lower proportion and lower values of non-zero scores for genes in sets. The significance of the difference in distributions of non-zero ΔlnL4 scores is tested with a Wilcoxon signed-rank test.

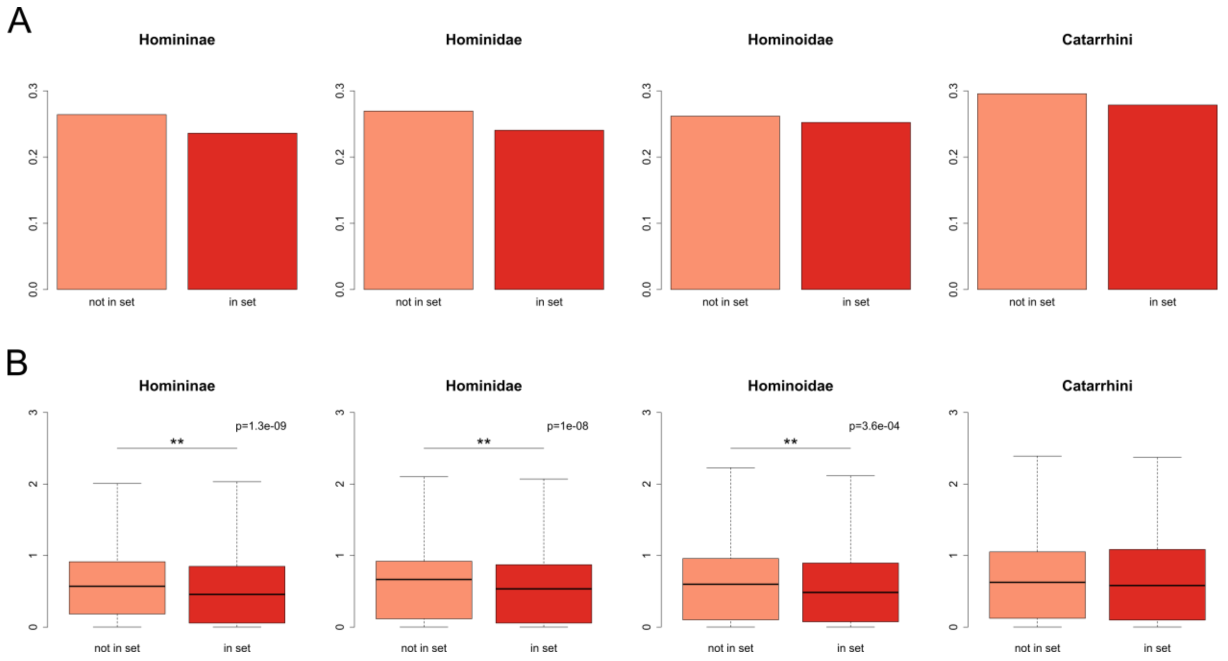


Figure S2

Enrichment map of the gene sets enriched for signals of positive selection in the Homininae branch. The nodes represent the 6 pathways with q-values < 0.2 before pruning (removal of overlapping genes). The node color scale represents gene set p-values. Edges represent mutual overlap; nodes are connected if one of the sets has at least 33% of its genes in common with the other gene set. The widths of the edges scale with the similarity between nodes. The two pathways that remained significant after pruning are marked with a black circle. A full-size version of this figure is available at https://dx.doi.org/10.6084/m9.figshare.3119029.

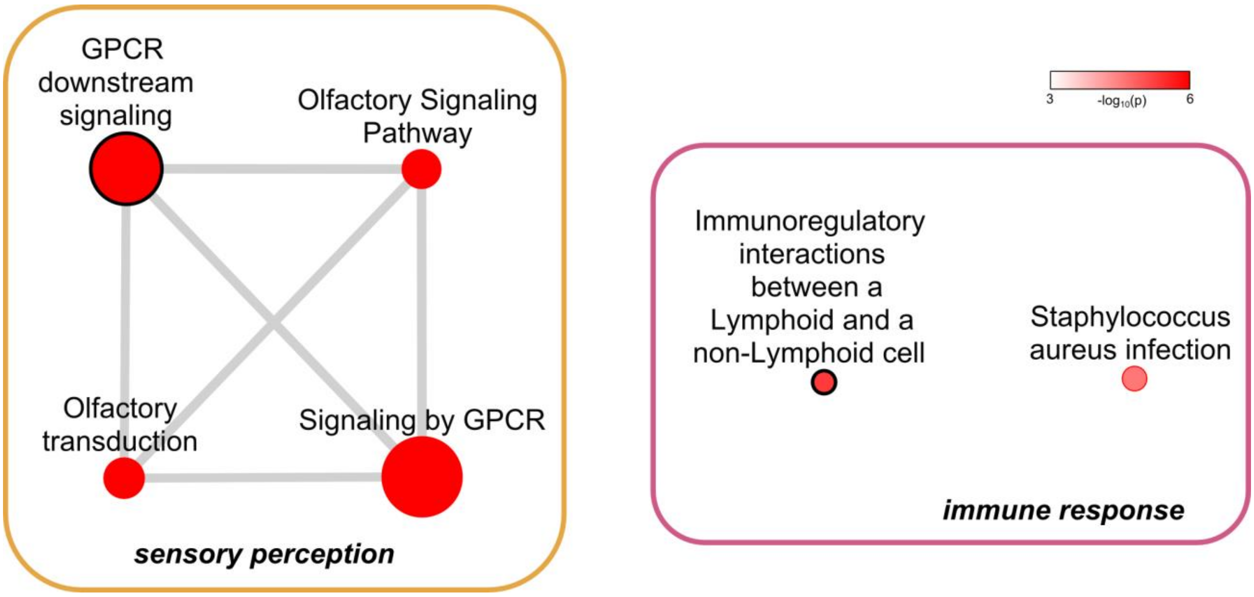
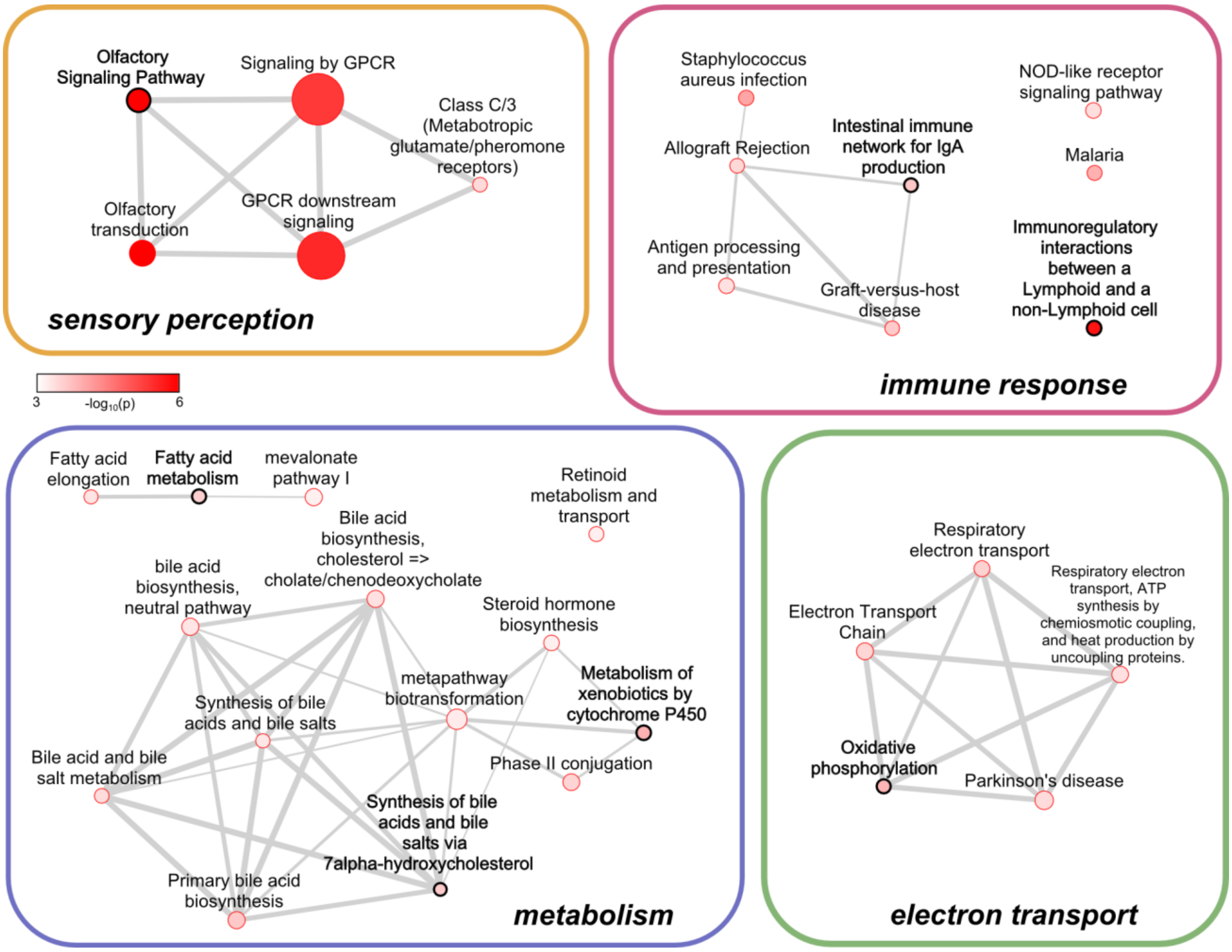

GPCR downstream signaling
Olfactory Signaling Pathway
Olfactory transduction
Signaling by GPCR
sensory perception
3
$-\log_{10}(p)$
6
Immunoregulatory interactions between a Lymphoid and a non-Lymphoid cell
Staphylococcus aureus infection
immune response

Figure S3

Enrichment map of the gene sets enriched for signals of positive selection in the Hominidae branch. The nodes represent the 32 pathways with q-values < 0.2 before pruning (removal of overlapping genes). The seven pathways that remained significant after pruning are marked with a black circle. See Figure S2 for a more detailed explanation of the enrichment map. Nodes marked with * represent unions of pathways that share more than 95% of their genes. A full-size version of this figure is available at https://dx.doi.org/10.6084/m9.figshare.3119029.



Olfactory Signaling Pathway
Signaling by GPCR
Class C/3 (Metabotropic glutamate/pheromone receptors)
Olfactory transduction
GPCR downstream signaling
sensory perception
3
-log10(p)
6
Staphylococcus aureus infection
NOD-like receptor signaling pathway
Intestinal immune network for IgA production
Allograft Rejection
Malaria
Immunoregulatory interactions between a Lymphoid and a non-Lymphoid cell
Antigen processing and presentation
Graft-versus-host disease
immune response
Fatty acid elongation
Fatty acid metabolism
mevalonate pathway I
Retinoid metabolism and transport
Bile acid biosynthesis, cholesterol => cholate/chenodeoxycholate
bile acid biosynthesis, neutral pathway
Steroid hormone biosynthesis
metapathway biotransformation
Metabolism of xenobiotics by cytochrome P450
Synthesis of bile acids and bile salts
Bile acid and bile salt metabolism
Phase II conjugation
Synthesis of bile acids and bile salts via 7alpha-hydroxycholesterol
Primary bile acid biosynthesis
metabolism
Respiratory electron transport
Respiratory electron transport, ATP synthesis by chemiosmotic coupling, and heat production by uncoupling proteins.
Electron Transport Chain
Oxidative phosphorylation
Parkinson's disease
electron transport

Figure S4

Enrichment map of the gene sets enriched for signals of positive selection in the Hominoidae branch. The nodes represent the 42 pathways with q-values < 0.2 before pruning (removal of overlapping genes). The six pathways that remained significant after pruning are marked with a black circle. See Figure S2 for a more detailed explanation of the enrichment map. A full-size version of this figure is available at https://dx.doi.org/10.6084/m9.figshare.3119029.

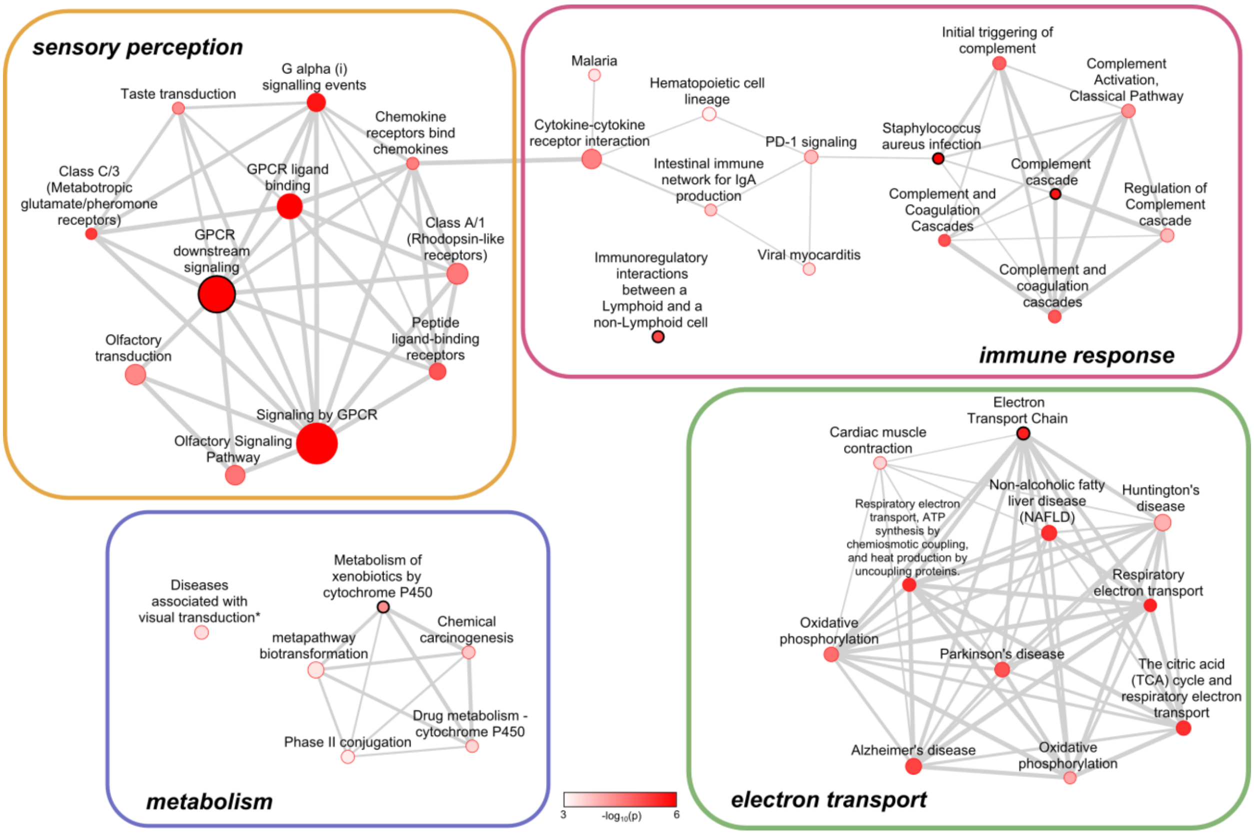

sensory perception
Taste transduction
G alpha (i) signalling events
Chemokine receptors bind chemokines
Class C/3 (Metabotropic glutamate/pheromone receptors)
GPCR ligand binding
GPCR downstream signaling
Class A/1 (Rhodopsin-like receptors)
Olfactory transduction
Peptide ligand-binding receptors
Signaling by GPCR
Olfactory Signaling Pathway
Malaria
Hematopoietic cell lineage
Cytokine-cytokine receptor interaction
Intestinal immune network for IgA production
PD-1 signaling
Immunoregulatory interactions between a Lymphoid and a non-Lymphoid cell
Viral myocarditis
Initial triggering of complement
Complement Activation, Classical Pathway
Staphylococcus aureus infection
Complement cascade
Complement and Coagulation Cascades
Regulation of Complement cascade
Complement and coagulation cascades
immune response
Diseases associated with visual transduction*
Metabolism of xenobiotics by cytochrome P450
Chemical carcinogenesis
metapathway biotransformation
Drug metabolism - cytochrome P450
Phase II conjugation
metabolism
3
-log10(p)
6
Electron Transport Chain
Cardiac muscle contraction
Non-alcoholic fatty liver disease (NAFLD)
Huntington's disease
Respiratory electron transport, ATP synthesis by chemiosmotic coupling, and heat production by uncoupling proteins.
Respiratory electron transport
Oxidative phosphorylation
Parkinson's disease
The citric acid (TCA) cycle and respiratory electron transport
Alzheimer's disease
Oxidative phosphorylation
electron transport

Figure S5

Enrichment map of the gene sets enriched for signals of positive selection in the Catarrhini branch. The nodes represent the 93 pathways with q-values < 0.2 before pruning (removal of overlapping genes). The nine pathways that remained significant after pruning are marked with a black circle. See Figure S2 for a more detailed explanation of the enrichment map. Nodes marked with * represent unions of pathways that share more than 95% of their genes. A full-size version of this figure is available at https://dx.doi.org/10.6084/m9.figshare.3119029.

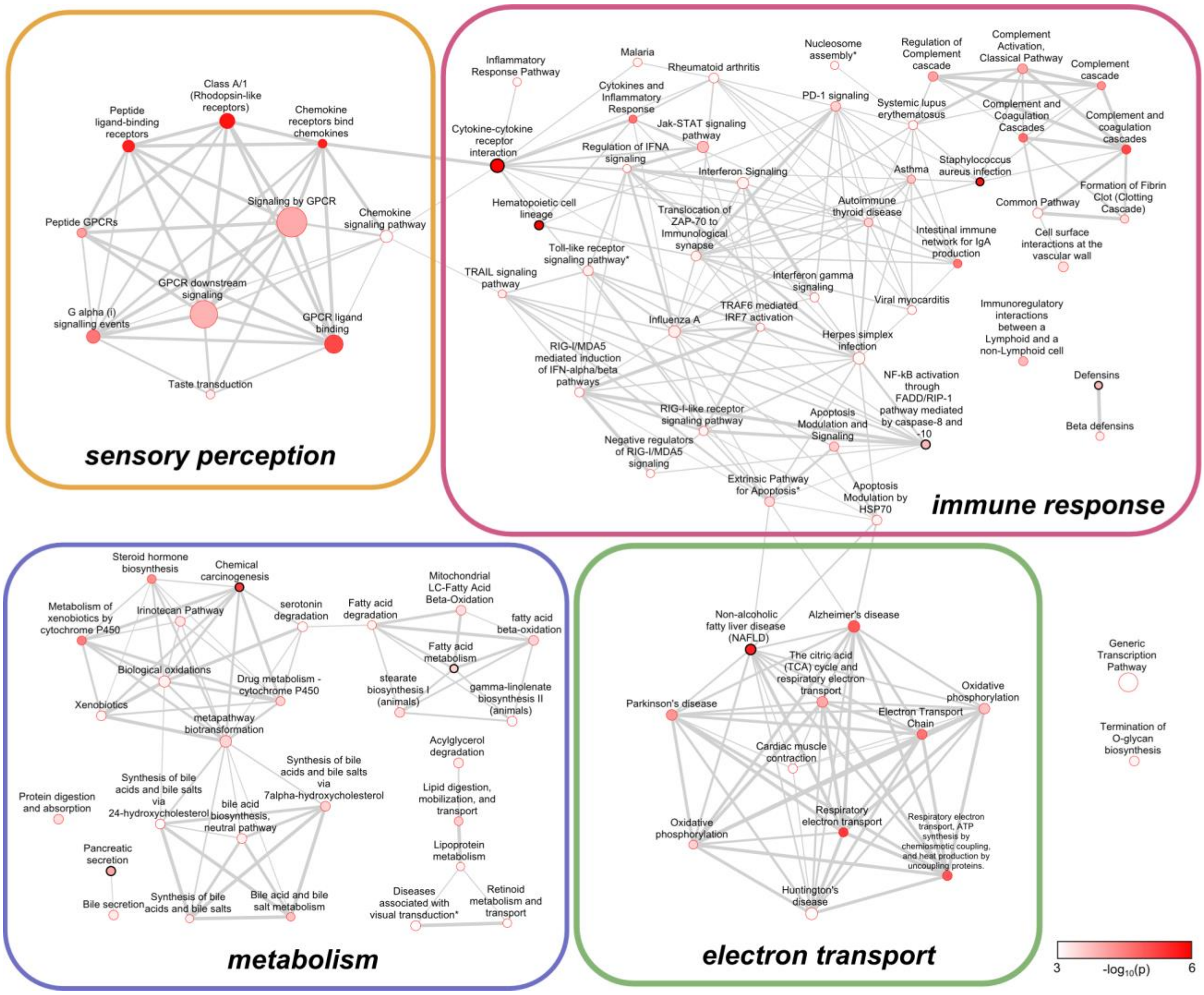
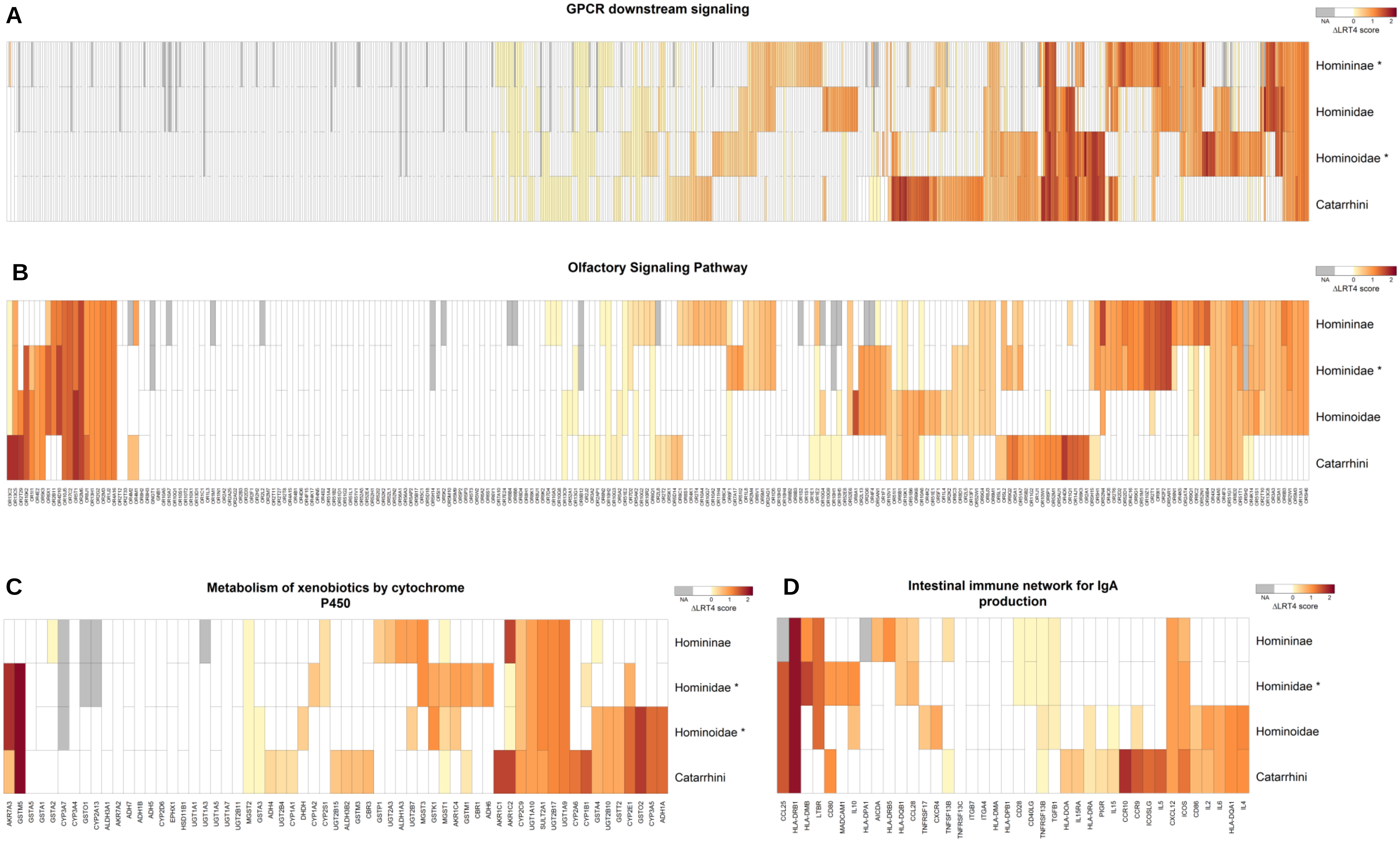
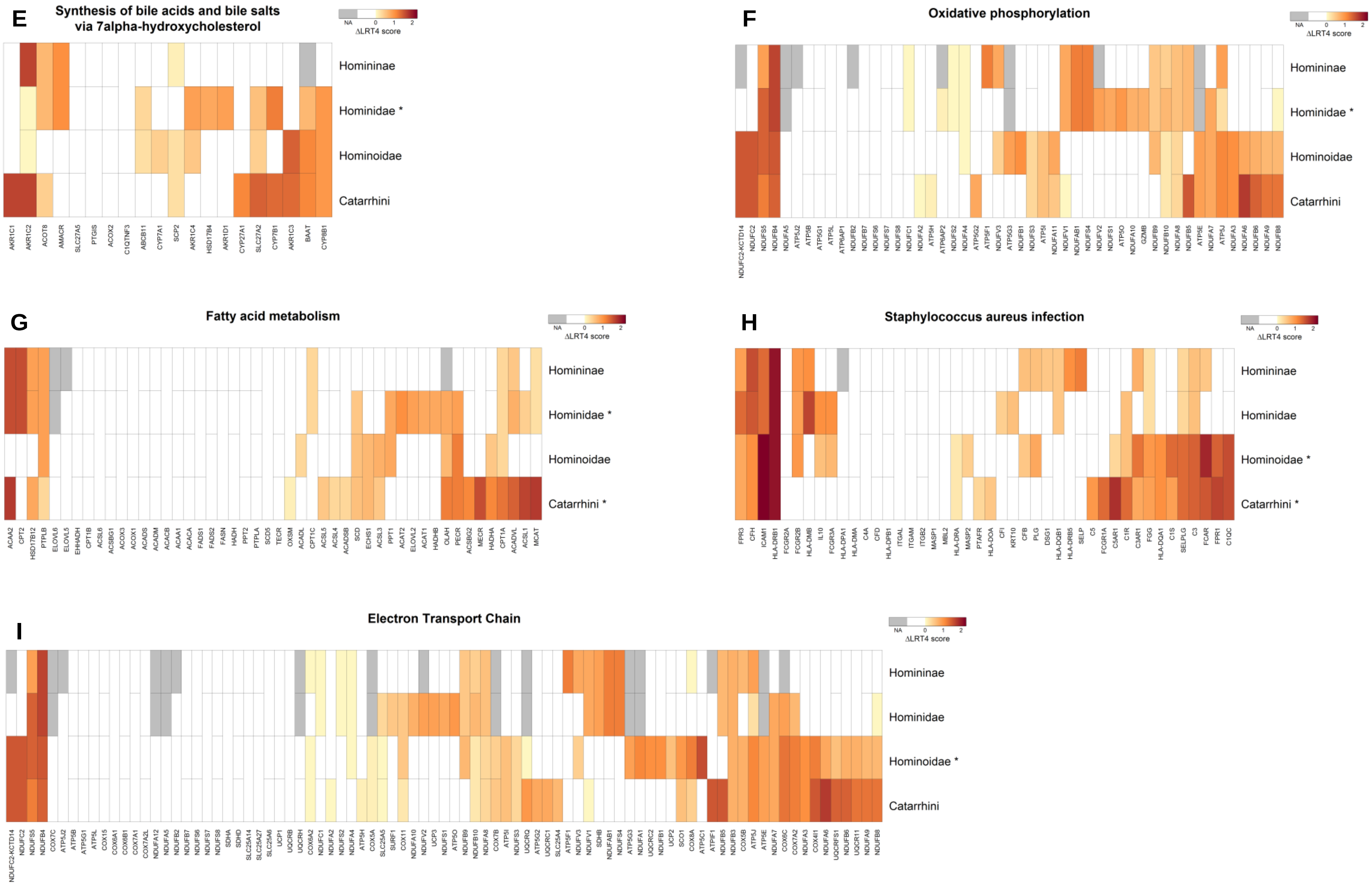
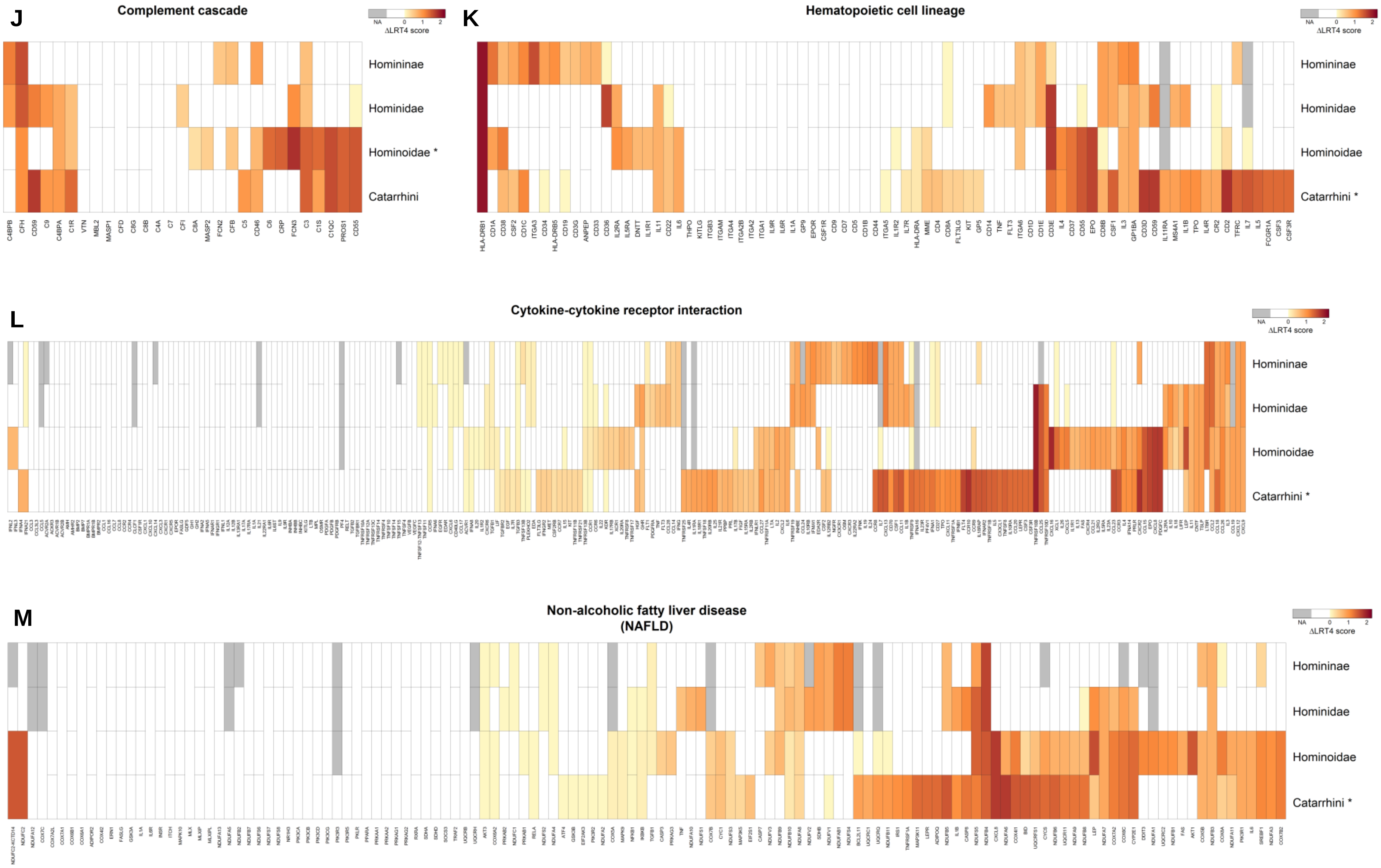

Class A/1 (Rhodopsin-like receptors)
Peptide ligand-binding receptors
Chemokine receptors bind chemokines
Signaling by GPCR
Chemokine signaling pathway
Peptide GPCRs
GPCR downstream signaling
G alpha (i) signalling events
GPCR ligand binding
Taste transduction
sensory perception
Inflammatory Response Pathway
Malaria
Rheumatoid arthritis
Nucleosome assembly*
Regulation of Complement cascade
Complement Activation, Classical Pathway
Complement cascade
Cytokines and Inflammatory Response
PD-1 signaling
Systemic lupus erythematosus
Complement and Coagulation Cascades
Complement and coagulation cascades
Cytokine-cytokine receptor interaction
Jak-STAT signaling pathway
Regulation of IFNA signaling
Interferon Signaling
Asthma
Staphylococcus aureus infection
Formation of Fibrin Clot (Clotting Cascade)
Common Pathway
Hematopoietic cell lineage
Translocation of ZAP-70 to Immunological synapse
Autoimmune thyroid disease
Intestinal immune network for IgA production
Cell surface interactions at the vascular wall
Toll-like receptor signaling pathway*
TRAIL signaling pathway
Interferon gamma signaling
Viral myocarditis
Immunoregulatory interactions between a Lymphoid and a non-Lymphoid cell
Influenza A
TRAF6 mediated IRF7 activation
Herpes simplex infection
RIG-I/MDA5 mediated induction of IFN-alpha/beta pathways
NF-kB activation through FADD/RIP-1 pathway mediated by caspase-8 and -10
Defensins
Beta defensins
RIG-I-like receptor signaling pathway
Apoptosis Modulation and Signaling
Negative regulators of RIG-I/MDA5 signaling
Extrinsic Pathway for Apoptosis*
Apoptosis Modulation by HSP70
immune response
Steroid hormone biosynthesis
Chemical carcinogenesis
Metabolism of xenobiotics by cytochrome P450
Irinotecan Pathway
serotonin degradation
Fatty acid degradation
Mitochondrial LC-Fatty Acid Beta-Oxidation
fatty acid beta-oxidation
Fatty acid metabolism
Biological oxidations
Drug metabolism - cytochrome P450
stearate biosynthesis I (animals)
gamma-linolenate biosynthesis II (animals)
Xenobiotics
metapathway biotransformation
Acylglycerol degradation
Synthesis of bile acids and bile salts via 7alpha-hydroxycholesterol
Synthesis of bile acids and bile salts via 24-hydroxycholesterol
bile acid biosynthesis, neutral pathway
Protein digestion and absorption
Lipid digestion, mobilization, and transport
Lipoprotein metabolism
Pancreatic secretion
Bile secretion
Synthesis of bile acids and bile salts
Bile acid and bile salt metabolism
Diseases associated with visual transduction*
Retinoid metabolism and transport
metabolism
Non-alcoholic fatty liver disease (NAFLD)
Alzheimer's disease
The citric acid (TCA) cycle and respiratory electron transport
Oxidative phosphorylation
Parkinson's disease
Electron Transport Chain
Cardiac muscle contraction
Oxidative phosphorylation
Respiratory electron transport
Respiratory electron transport, ATP synthesis by chemiosmotic coupling, and heat production by uncoupling proteins.
Huntington's disease
electron transport
Generic Transcription Pathway
Termination of O-glycan biosynthesis
3
$-\log_{10}(p)$
6

Figure S6

Heat maps showing branch specific ΔlnL4 scores of genes in pathways that score significant ($q < 0.2$) in the gene set enrichment test after pruning. The ΔlnL4 score is computed as the fourth root of the log-likelihood ratio in the branch site test for positive selection. Branches where a pathway scores significant are marked with a '*'. The genes are grouped by hierarchical clustering to visualize blocks with similar signals among branches. Genes for which ΔlnL4 scores were not available (NA) in a certain branch are depicted in grey. Genes are merged (horizontally) with their paralog(s) into an 'ancestral gene' in the branches preceding a duplication and their scores were included only once in the calculation of the SUMSTAT score for these branches. Genes with (vertically) merged branches represent cases where the sequence of one or more species is missing or excluded, resulting in a single 'average' ΔlnL4 score over multiple branches. We used this score when testing each branch separately. Full-size versions of these heat maps are available at https://dx.doi.org/10.6084/m9.figshare.3119026.


A
GPCR downstream signaling
NA
0
1
2
ΔLRT4 score
Homininae *
Hominidae
Hominoidae *
Catarrhini
B
Olfactory Signaling Pathway
NA
0
1
2
ΔLRT4 score
Homininae
Hominidae *
Hominoidae
Catarrhini
C
Metabolism of xenobiotics by cytochrome P450
NA
0
1
2
ΔLRT4 score
Homininae
Hominidae *
Hominoidae *
Catarrhini
AKR7A3
GSTM5
GSTA5
GSTA1
GSTA2
CYP3A7
CYP3A4
GSTO1
CYP2A13
ALDH3A1
AKR7A2
ADH7
ADH1B
ADH5
CYP2D6
EPHX1
HSD11B1
UGT1A1
UGT1A3
UGT1A5
UGT1A7
UGT2B11
MGST2
GSTA3
ADH4
UGT2B4
CYP1A1
DHDH
CYP1A2
CYP2S1
UGT2B15
ALDH3B2
GSTM3
CBR3
GSTP1
UGT2A3
ALDH1A3
UGT2B7
MGST3
GSTK1
MGST1
AKR1C4
GSTM1
CBR1
ADH6
AKR1C1
AKR1C2
CYP2C9
UGT1A10
SULT2A1
UGT2B17
UGT1A9
CYP2A6
CYP1B1
GSTA4
UGT2B10
GSTT2
CYP2E1
GSTO2
CYP3A5
ADH1A
D
Intestinal immune network for IgA production
NA
0
1
2
ΔLRT4 score
Homininae
Hominidae *
Hominoidae
Catarrhini
CCL25
HLA-DRB1
HLA-DMB
LTBR
CD80
MADCAM1
IL10
HLA-DPA1
AICDA
HLA-DRB5
HLA-DQB1
CCL28
TNFRSF17
CXCR4
TNFSF13B
TNFRSF13C
ITGB7
ITGA4
HLA-DMA
HLA-DPB1
CD28
CD40LG
TNFRSF13B
TGFB1
HLA-DOA
IL15RA
HLA-DRA
PIGR
IL15
CCR10
CCR9
ICOSLG
IL5
CXCL12
ICOS
CD86
IL2
IL6
HLA-DQA1
IL4


E
Synthesis of bile acids and bile salts via 7alpha-hydroxycholesterol
NA
0
1
2
ΔLRT4 score
Homininae
Hominidae *
Hominoidae
Catarrhini
AKR1C1
AKR1C2
ACOT8
AMACR
SLC27A5
PTGIS
ACOX2
C1QTNF3
ABCB11
CYP7A1
SCP2
AKR1C4
HSD17B4
AKR1D1
CYP27A1
SLC27A2
CYP7B1
AKR1C3
BAAT
CYP8B1
F
Oxidative phosphorylation
NA
0
1
2
ΔLRT4 score
Homininae
Hominidae *
Hominoidae
Catarrhini
NDUFC2-KCTD14
NDUFC2
NDUFS5
NDUFB4
NDUFA5
ATP5J2
ATP5B
ATP5G1
ATP5L
ATP6AP1
NDUFB2
NDUFB7
NDUFS6
NDUFS7
NDUFS8
NDUFC1
NDUFA2
ATP5H
ATP6AP2
NDUFS2
NDUFA4
ATP5G2
ATP5F1
NDUFV3
ATP5G3
NDUFB1
NDUFS3
ATP5I
NDUFA11
NDUFV1
NDUFAB1
NDUFS4
NDUFV2
NDUFS1
ATP5O
NDUFA10
GZMB
NDUFB9
NDUFB10
NDUFA8
NDUFB5
ATP5E
NDUFA7
ATP5J
NDUFA3
NDUFA6
NDUFB6
NDUFA9
NDUFB8
G
Fatty acid metabolism
NA
0
1
2
ΔLRT4 score
Homininae
Hominidae *
Hominoidae
Catarrhini *
ACAA2
CPT2
HSD17B12
PTPLB
ELOVL6
ELOVL5
EHHADH
CPT1B
ACSL6
ACSBG1
ACOX3
ACOX1
ACADS
ACADM
ACACB
ACAA1
ACACA
FADS1
FADS2
FASN
HADH
PPT2
PTPLA
SCD5
TECR
OXSM
ACADL
CPT1C
ACSL5
ACSL4
ACADSB
SCD
ECHS1
ACSL3
PPT1
ACAT2
ELOVL2
ACAT1
HADHB
OLAH
PECR
ACSBG2
MECR
HADHA
CPT1A
ACADVL
ACSL1
MCAT
H
Staphylococcus aureus infection
NA
0
1
2
ΔLRT4 score
Homininae
Hominidae
Hominoidae *
Catarrhini *
FPR3
CFH
ICAM1
HLA-DRB1
FCGR2A
FCGR2B
HLA-DMB
IL10
FCGR3A
HLA-DPA1
HLA-DMA
C4A
CFD
HLA-DPB1
ITGAL
ITGAM
ITGB2
MASP1
MBL2
HLA-DRA
MASP2
PTAFR
HLA-DOA
CFI
KRT10
CFB
PLG
DSG1
HLA-DQB1
HLA-DRB5
SELP
C5
FCGR1A
C5AR1
C1R
C3AR1
FGG
HLA-DQA1
C1S
SELPLG
C3
FCAR
FPR1
C1QC
I
Electron Transport Chain
NA
0
1
2
ΔLRT4 score
Homininae
Hominidae
Hominoidae *
Catarrhini
NDUFC2-KCTD14
NDUFC2
NDUFS5
NDUFB4
COX7C
ATP5J2
ATP5B
ATP5G1
ATP5L
COX15
COX6A1
COX6B1
COX7A1
COX7A2L
NDUFA12
NDUFA5
NDUFB2
NDUFB7
NDUFS6
NDUFS7
NDUFS8
SDHA
SDHD
SLC25A14
SLC25A27
SLC25A6
UCP1
UQCRB
UQCRH
COX6A2
NDUFC1
NDUFA2
NDUFS2
NDUFA4
ATP5H
COX5A
SLC25A5
SURF1
COX11
NDUFA10
NDUFV2
UCP3
NDUFS1
ATP5O
NDUFB9
NDUFB10
NDUFA8
COX7B
ATP5I
NDUFS3
UQCRQ
ATP5G2
UQCRC1
SLC25A4
ATP5F1
NDUFV3
NDUFV1
SDHB
NDUFAB1
NDUFS4
ATP5G3
NDUFA1
UQCRC2
NDUFB1
UCP2
SCO1
COX8A
ATP5C1
ATPIF1
NDUFB5
NDUFB3
COX5B
ATP5J
ATP5E
NDUFA7
COX6C
COX7A2
NDUFA3
COX4I1
NDUFA6
UQCRFS1
NDUFB6
UQCR11
NDUFA9
NDUFB8


J
Complement cascade
NA 0 1 2
ΔLRT4 score
Homininae
Hominidae
Hominoidae *
Catarrhini
C4BPB
CFH
CD59
C9
C4BPA
C1R
VTN
MBL2
MASP1
CFD
C8G
C8B
C4A
C7
CFI
C8A
MASP2
FCN2
CFB
C5
CD46
C6
CRP
FCN3
C3
C1S
C1QC
PROS1
CD55
K
Hematopoietic cell lineage
NA 0 1 2
ΔLRT4 score
Homininae
Hominidae
Hominoidae
Catarrhini *
HLA-DRB1
CD1A
CD38
CSF2
CD1C
ITGA3
CD34
HLA-DRB5
CD19
CD3G
ANPEP
CD33
CD36
IL2RA
IL5RA
DNTT
IL1R1
IL11
CD22
IL6
THPO
KITLG
ITGB3
ITGAM
ITGA4
ITGA2B
ITGA2
ITGA1
IL9R
IL6R
IL1A
GP9
EPOR
CSF1R
CD9
CD7
CD5
CD1B
CD44
ITGA5
IL1R2
IL7R
HLA-DRA
MME
CD4
CD8A
FLT3LG
KIT
GP5
CD14
TNF
FLT3
ITGA6
CD1D
CD1E
CD3E
IL4
CD37
CD55
EPO
CD8B
CSF1
IL3
GP1BA
CD3D
CD59
IL11RA
MS4A1
IL1B
TPO
IL4R
CR2
CD2
TFRC
IL7
IL5
FCGR1A
CSF3
CSF3R
L
Cytokine-cytokine receptor interaction
NA 0 1 2
ΔLRT4 score
Homininae
Hominidae
Hominoidae
Catarrhini *
M
Non-alcoholic fatty liver disease
(NAFLD)
NA 0 1 2
ΔLRT4 score
Homininae
Hominidae
Hominoidae
Catarrhini *

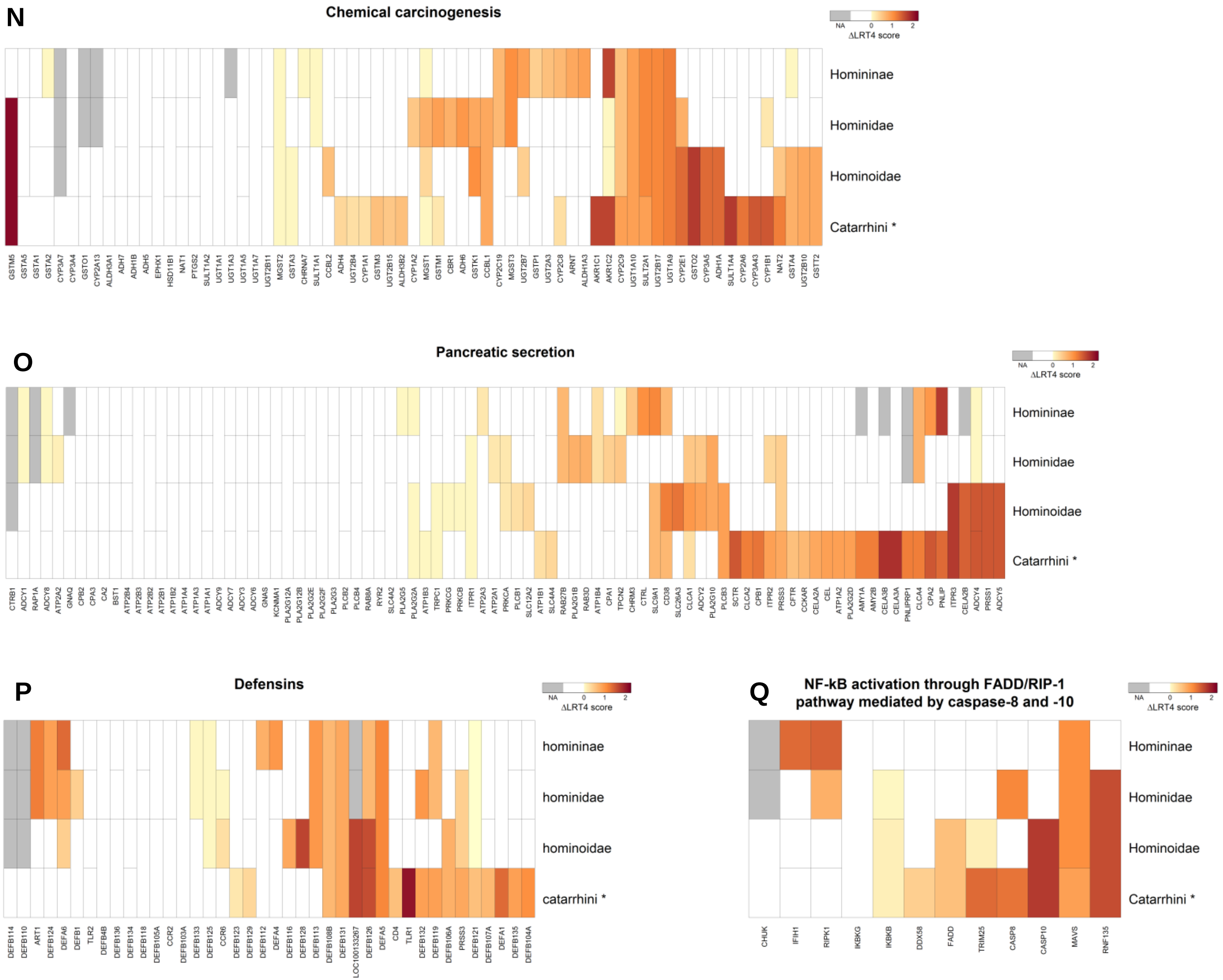

N
Chemical carcinogenesis
NA
0
1
2
ΔLRT4 score
Homininae
Hominidae
Hominoidae
Catarrhini *
GSTM5
GSTA5
GSTA1
GSTA2
CYP3A7
CYP3A4
GSTO1
CYP2A13
ALDH3A1
ADH7
ADH1B
ADH5
EPHX1
HSD11B1
NAT1
PTGS2
SULT1A2
UGT1A1
UGT1A3
UGT1A5
UGT1A7
UGT2B11
MGST2
GSTA3
CHRNA7
SULT1A1
CCBL2
ADH4
UGT2B4
CYP1A1
GSTM3
UGT2B15
ALDH3B2
CYP1A2
MGST1
GSTM1
CBR1
ADH6
GSTK1
CCBL1
CYP2C19
MGST3
UGT2B7
GSTP1
UGT2A3
CYP2C8
ARNT
ALDH1A3
AKR1C1
AKR1C2
CYP2C9
UGT1A10
SULT2A1
UGT2B17
UGT1A9
CYP2E1
GSTO2
CYP3A5
ADH1A
SULT1A4
CYP2A6
CYP3A43
CYP1B1
NAT2
GSTA4
UGT2B10
GSTT2
O
Pancreatic secretion
NA
0
1
2
ΔLRT4 score
Homininae
Hominidae
Hominoidae
Catarrhini *
CTRB1
ADCY1
RAP1A
ADCY8
ATP2A2
GNAQ
CPB2
CPA3
CA2
BST1
ATP2B4
ATP2B3
ATP2B2
ATP2B1
ATP1B2
ATP1A4
ATP1A3
ATP1A1
ADCY9
ADCY7
ADCY3
ADCY6
GNAS
KCNMA1
PLA2G12A
PLA2G12B
PLA2G2E
PLA2G2F
PLA2G3
PLCB2
PLCB4
RAB8A
RYR2
SLC4A2
PLA2G5
PLA2G2A
ATP1B3
TRPC1
PRKCG
PRKCB
ITPR1
ATP2A3
ATP2A1
PRKCA
PLCB1
SLC12A2
ATP1B1
SLC4A4
RAB27B
PLA2G1B
RAB3D
ATP1B4
CPA1
TPCN2
CHRM3
CTRL
SLC9A1
CD38
SLC26A3
CLCA1
ADCY2
PLA2G10
PLCB3
SCTR
CLCA2
CPB1
ITPR2
PRSS3
CFTR
CCKAR
CELA2A
CEL
ATP1A2
PLA2G2D
AMY1A
AMY2B
CELA3B
CELA3A
PNLIPRP1
CLCA4
CPA2
PNLIP
ITPR3
CELA2B
ADCY4
PRSS1
ADCY5
P
Defensins
NA
0
1
2
ΔLRT4 score
homininae
hominidae
hominoidae
catarrhini *
DEFB114
DEFB110
ART1
DEFB124
DEFA6
DEFB1
TLR2
DEFB4B
DEFB136
DEFB134
DEFB118
DEFB105A
CCR2
DEFB103A
DEFB133
DEFB125
CCR6
DEFB123
DEFB129
DEFB112
DEFA4
DEFB116
DEFB128
DEFB113
DEFB108B
DEFB131
LOC100133267
DEFB126
DEFA5
CD4
TLR1
DEFB132
DEFB119
DEFB106A
PRSS3
DEFB121
DEFB107A
DEFA1
DEFB135
DEFB104A
Q
NF-kB activation through FADD/RIP-1
pathway mediated by caspase-8 and -10
NA
0
1
2
ΔLRT4 score
Homininae
Hominidae
Hominoidae
Catarrhini *
CHUK
IFIH1
RIPK1
IKBKG
IKBKB
DDX58
FADD
TRIM25
CASP8
CASP10
MAVS
RNF135

Figure S7

Enrichment map of gene sets enriched with genes that were excluded in Selectome. The nodes represent the 62 pathways with q-values < 0.2 in a Fisher's exact test. See Figure S2 for a more detailed description of the enrichment map. Nodes marked with * represent unions of pathways that share more than 95% of their genes. A full-size version of this figure is available at https://dx.doi.org/10.6084/m9.figshare.3119122.

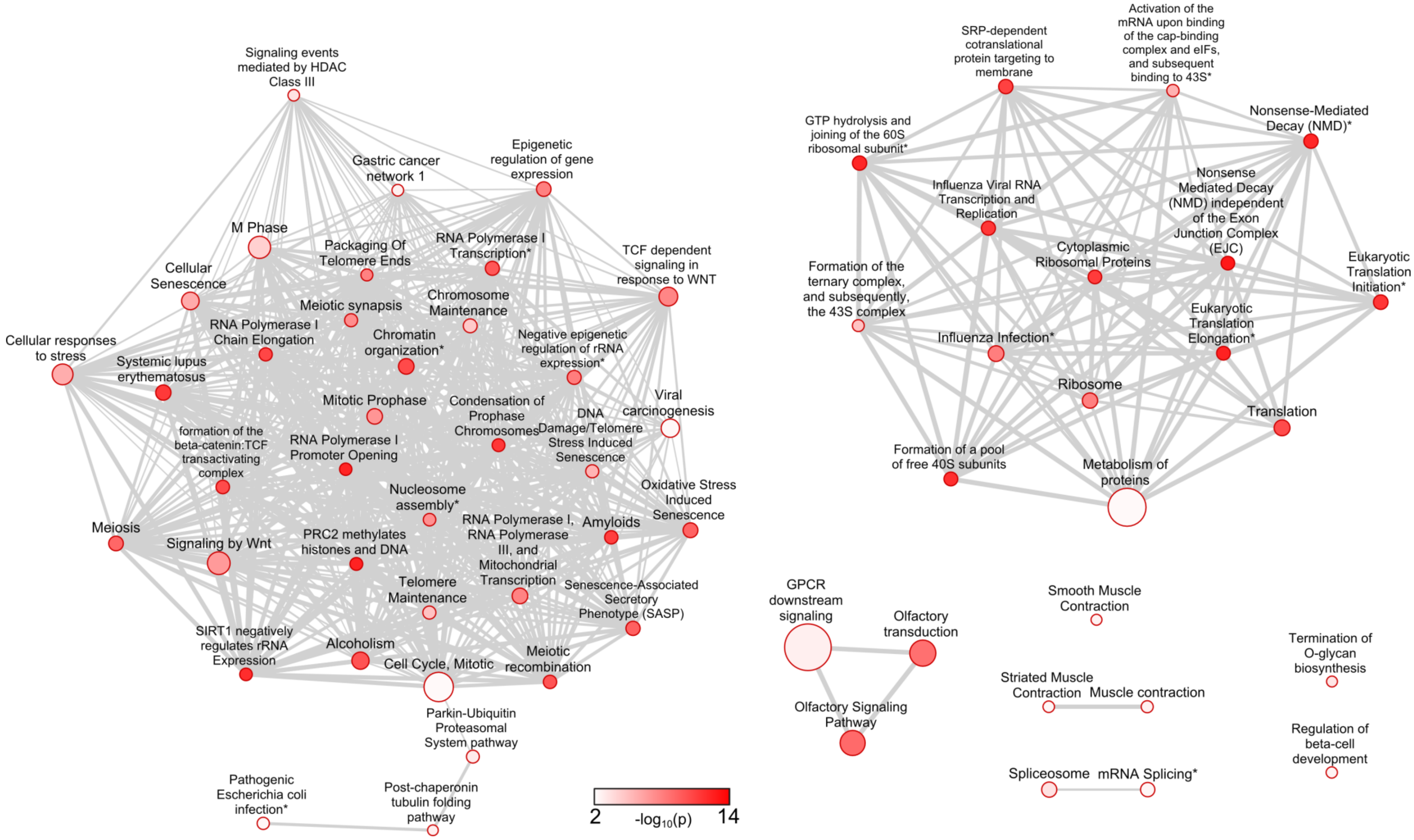
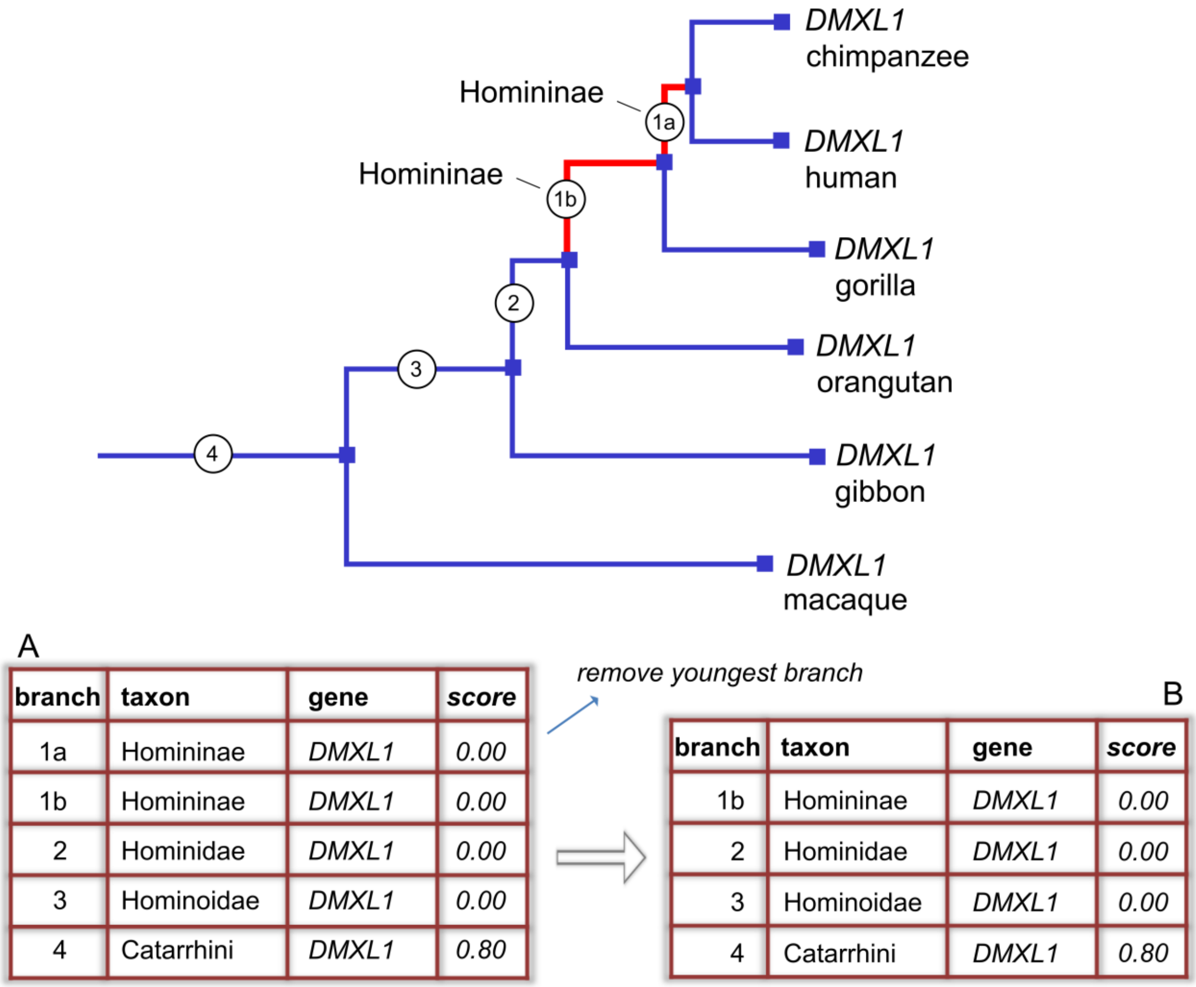

Signaling events mediated by HDAC Class III
Gastric cancer network 1
Epigenetic regulation of gene expression
M Phase
Packaging Of Telomere Ends
RNA Polymerase I Transcription*
TCF dependent signaling in response to WNT
Cellular Senescence
Meiotic synapsis
Chromosome Maintenance
RNA Polymerase I Chain Elongation
Chromatin organization*
Negative epigenetic regulation of rRNA expression*
Cellular responses to stress
Systemic lupus erythematosus
Mitotic Prophase
Condensation of Prophase Chromosomes
DNA Damage/Telomere Stress Induced Senescence
Viral carcinogenesis
formation of the beta-catenin:TCF transactivating complex
RNA Polymerase I Promoter Opening
Nucleosome assembly*
Oxidative Stress Induced Senescence
Meiosis
Signaling by Wnt
PRC2 methylates histones and DNA
RNA Polymerase I, RNA Polymerase III, and Mitochondrial Transcription
Amyloids
Telomere Maintenance
Senescence-Associated Secretory Phenotype (SASP)
SIRT1 negatively regulates rRNA Expression
Alcoholism
Cell Cycle, Mitotic
Meiotic recombination
Parkin-Ubiquitin Proteasomal System pathway
Pathogenic Escherichia coli infection*
Post-chaperonin tubulin folding pathway
2
$-\log_{10}(p)$
14
SRP-dependent cotranslational protein targeting to membrane
Activation of the mRNA upon binding of the cap-binding complex and eIFs, and subsequent binding to 43S*
GTP hydrolysis and joining of the 60S ribosomal subunit*
Nonsense-Mediated Decay (NMD)*
Influenza Viral RNA Transcription and Replication
Nonsense Mediated Decay (NMD) independent of the Exon Junction Complex (EJC)
Cytoplasmic Ribosomal Proteins
Formation of the ternary complex, and subsequently, the 43S complex
Eukaryotic Translation Initiation*
Eukaryotic Translation Elongation*
Influenza Infection*
Ribosome
Translation
Formation of a pool of free 40S subunits
Metabolism of proteins
GPCR downstream signaling
Olfactory transduction
Olfactory Signaling Pathway
Smooth Muscle Contraction
Striated Muscle Contraction
Muscle contraction
Termination of O-glycan biosynthesis
Regulation of beta-cell development
Spliceosome
mRNA Splicing*

Figure S8

Example of multiple Homininae branches (branch *1a* and *1b*, marked red) in the *DMXL1* gene tree. Only the oldest branch (*1b*) will be used in the enrichment test. Table A illustrates that initially the Hominae branches occur in double in the table, while after removing the youngest Homininae branch, only one row per branch (taxon) and human gene remains (Table B). It should be emphasized that this is not a biological phenomenon, but simply a technical result of the labeling choices made in the Ensembl database, which are carried into the Selectome database and thus into the dataset used in this study.



A

branch	taxon	gene	score
1a	Homininae	*DMXL1*	*0.00*
1b	Homininae	*DMXL1*	*0.00*
2	Hominidae	*DMXL1*	*0.00*
3	Hominoidae	*DMXL1*	*0.00*
4	Catarrhini	*DMXL1*	*0.80*

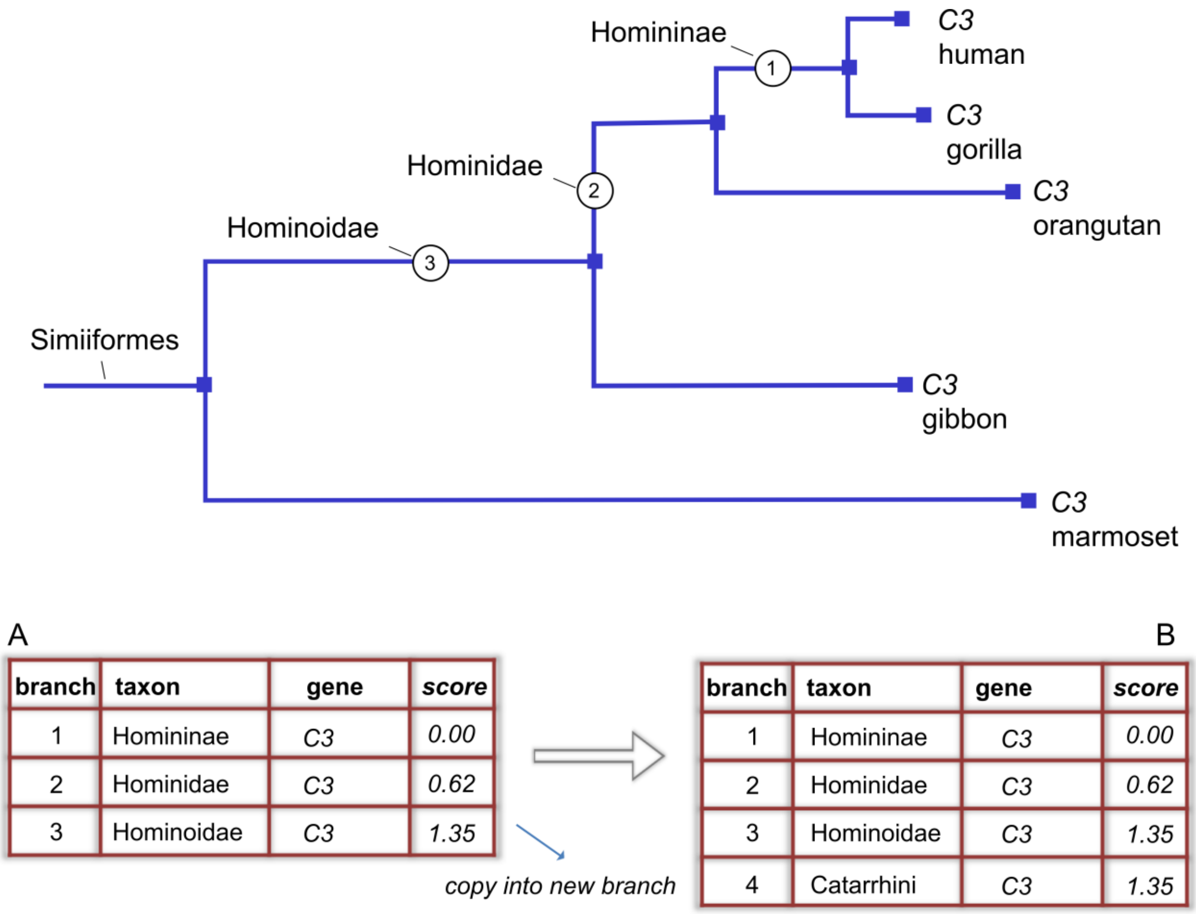
remove youngest branch

B

branch	taxon	gene	score
1b	Homininae	*DMXL1*	*0.00*
2	Hominidae	*DMXL1*	*0.00*
3	Hominoidae	*DMXL1*	*0.00*
4	Catarrhini	*DMXL1*	*0.80*

Figure S9

Example of a missing branch (Catarrhini) in the CE3 gene tree, due to an absent or low quality macaque sequence for this gene. The score assigned to the Hominoidae branch actually represents an 'average' score for both the Hominoidae and Catarrhini branch, and therefore we use this score for both branches in the enrichment test. Table A illustrates that initially one branch is missing, while after copying the Hominoidae score into a new Catarrhini row, one row per branch (taxon) and human gene is now available (Table B).



A

branch	taxon	gene	score
1	Homininae	*C3*	*0.00*
2	Hominidae	*C3*	*0.62*
3	Hominoidae	*C3*	*1.35*

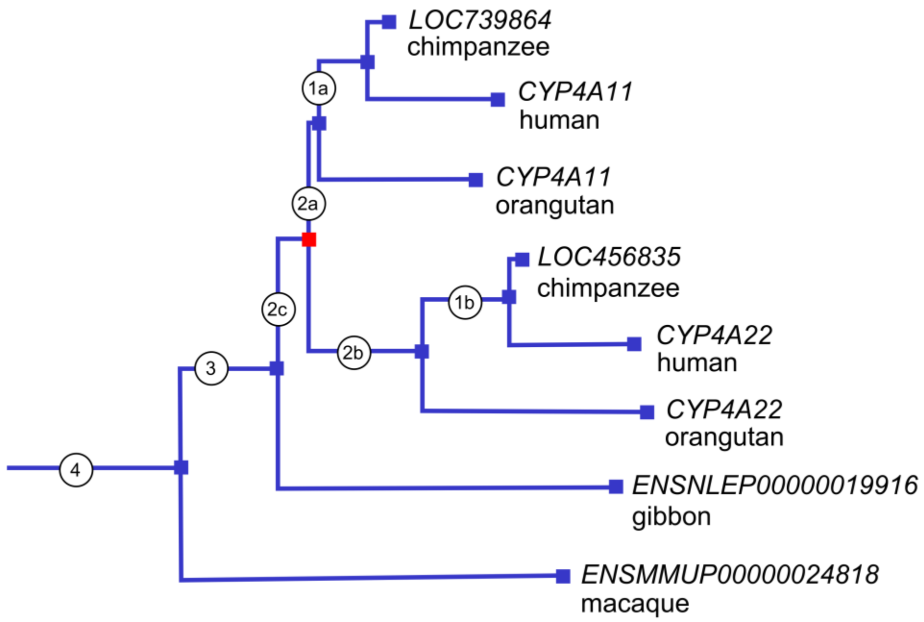
copy into new branch

B

branch	taxon	gene	score
1	Homininae	*C3*	*0.00*
2	Hominidae	*C3*	*0.62*
3	Hominoidae	*C3*	*1.35*
4	Catarrhini	*C3*	*1.35*

Figure S10

Example of a duplication event (red node) in the Hominidae branch of the *CYP4A11*/*CYP4A22* gene tree. In the enrichment tests (i) the branch leading to the duplication (*2c*) is removed and (ii) in the Hominoidae (*3*) and Catarrhini branch (*4*), the 'ancestral' gene *CYP4A11*, *CYP4A22* replaces the two human paralogs *CYP4A11*and *CYP4A22*. Table A illustrates that initially some branches (and their scores) occur in double in the table. After merging and removing branches, only one row per branch (taxon) and human gene remains (Table B).



A

branch	taxon	gene	score
1a	Homininae	CYP4A11	0.00
1b	Homininae	CYP4A22	0.00
2a	Hominidae	CYP4A11	0.00
2b	Hominidae	CYP4A22	0.57
2c	Hominidae	CYP4A11	0.00
2c	Hominidae	CYP4A22	0.00
3	Hominoidae	CYP4A11	0.00
3	Hominoidae	CYP4A22	0.00
4	Catarrhini	CYP4A11	0.00
4	Catarrhini	CYP4A22	0.00

remove branch to duplication

merge with CYP4A11 into ancestral gene

B

branch	taxon	gene	score
1a	Homininae	CYP4A11	0.00
1b	Homininae	CYP4A22	0.00
2a	Hominidae	CYP4A11	0.00
2b	Hominidae	CYP4A22	0.57
3	Hominoidae	CYP4A11, CYP4A22	0.00
4	Catarrhini	CYP4A11, CYP4A22	0.00

References

Ota R, Waddell PJ, Hasegawa M, Shimodaira H, Kishino H. 2000. Appropriate likelihood ratio tests and marginal distributions for evolutionary tree models with constraints on parameters. Mol Biol Evol 17:798-803.

Self SG, Liang KY. 1987. Asymptotic Properties of Maximum-Likelihood Estimators and Likelihood Ratio Tests under Nonstandard Conditions. Journal of the American Statistical Association 82:605-610.

Zhang J, Nielsen R, Yang Z. 2005. Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. Mol Biol Evol 22:2472-2479.