Supplementary Materials for

Ancient hybridization and strong adaptation to viruses across African vervet monkey populations

Hannes Svardal; Anna J Jasinska; Cristian Apetrei; Giovanni Coppola; Yu Huang; Christopher A Schmitt; Beatrice Jacquelin; Vasily Ramensky; Michaela Muller-Trutwin; Martin Antonio; George Weinstock; J Paul Grobler; Ken Dewar; Richard K Wilson; Trudy R Turner; Wesley C Warren; Nelson B Freimer; Magnus Nordborg*

*Correspondence to: magnus.nordborg@gmi.oeaw.ac.at

Contents

List of Figures

List of Tables

Supplementary Note 1: Alignment against rhesus macaque genome

To check whether our inference of species relationships and D-statistic are affected by reference or coverage bias, we selected one individual each from 10 different countries with 4-5 X cover-

age (Supplementary Table 4), realigned the reads of these individuals against the autosomes of the rhesus macaque reference genome Mmul_8.0.1 (https://www.ncbi.nlm.nih.gov/assembly/GCF_000772875.2/), and called variants. Alignment and variant calling methodology was similar to the original data set, except that base quality score recalibration was omitted and hard filters were slightly adapted ($\mathrm{MQ} < 40$, $(\mathrm{MQ0}/(1.0 * \mathrm{DP})) > 0.1$).

Confirming the first split in the vervet phylogeny

The neighbor-joining and UPGMA trees presented in Fig. 1b and c, respectively, are consistent with the phylogeny in Warren et al. (2015)[11] but differ from the suggested species relationships in Pfeifer (2017)[59]. In particular, Pfeifer[59] used putatively neutral fixed differences to infer divergence times, and suggested the first split to be between *aethiops* and the other taxa, while we infer *sabaeus* to split off first.

To further test whether our inferred phylogenetic relationships are correct, we first confirmed that a neighbor-joining tree of the pairwise difference matrix calculated on the macaque aligned callset yields the taxonomic relationships shown in Fig. 1b and c (Supplementary Figure 6). Aligning against the macaque genome eliminates the possible bias of using a *sabaeus* derived reference genome.

Classic phylogenetic methods such as maximum likelihood are not appropriate to resolve relationships between closely related taxa such as vervets with substantial incomplete lineage sorting and gene flow. Therefore, as a powerful and specific test which vervet taxon constitutes the outgroup to other vervet taxa, we computed two complementary statistics of the site frequency spectrum related to the average time to the most recent common ancestor (TMRCA) on different subsets of vervet samples. The rationale behind this approach is that the TMRCA should change most strongly when all samples of an outgroup branch are removed, while re-

moval of ingroup samples should have a relatively smaller effect on the average TMRCA of the whole sample. This is, because, for the many coalescent trees across the genome, the former is expected most often to be the last sample(s) coalescing with the ancestor of the rest. To eliminate any reference bias, we again used the macaque aligned callset with one representative for each of the clearly diverged taxa: VGA00010 (*sabaeus* from The Gambia), VEA1001 (*aethiops*), AGM126 (*tantalus*), VSAE3002 (*pygerythrus* from South Africa), and only considered variants that were polymporphic across vervets (that is, we excluded fixed differences between the macaque reference and our vervet samples). As we will see below, given the many (mostly) independent coalescent histories across the genome, a single sample provides more than sufficient power to discern between alternative species topologies.

For given SNP i and individual j, let $x_{i,j} = 0, 0.5,$ or 1 depending on whether sample j has 0, 1, or 2 derived alleles at this locus. For each sample k we define

$$\theta_k = \sum_i \frac{1}{n-1} \sum_{j \neq k} x_{i,j} - \prod_{j \neq k} x_{i,j} \tag{1}$$

and

$$\tau_k = \sum_i \prod_{j \neq k} x_{i,j}(1 - x_{i,k}), \tag{2}$$

where n is the total number of samples. These statistics both capture aspects of the average coalescent history of the sample (Supplementary Fig. 7a and b). θ_k is equivalent to θ_L in Zeng et al. (2006)[80] for the subsample excluding k. Simply speaking, for a given coalescent history, θ_k is a measure of the the average number of mutations on a branch from present to the most recent common ancestor (excluding sample k) and thus proportional to the average TMRCA of the samples excluding k. On the other hand, τ_k (named "anti-singleton entry" in Spence et al. 2016[81]) is the number of variants that are fixed for the non-macaque state in all samples but k, and is proportional to the average difference in TMRCA between alls amples and all samples excluding k. Hence, we expect θ_k to be lowest and τ_k to be highest for sample k corresponding

to the outgroup in the species tree. Note that the expectation of $\tau_k \neq 0$ even for k not being the outgroup, because of incomplete lineage sorting. Further note that recurrent mutation and errors in genotyping and outgroup inference are expected to increase noise in these statistics (and reduce any relative difference for the outgroup), but are not expected to give a directional bias.

For each statistic, we calculated a block-jackknife estimate of the standard deviation across windows of 50000 variants in the original VCF file to assess variability in our estimates. Statistics were normalised by the total number of SNPs segregating across vervets in this dataset. The analysis strongly confirmed the status of *sabaeus* as an outgroup relative to the other vervet taxa (Supplementary Figs. 7c-e). In particular, θ_k is clearly lowest and τ_k clearly highest for k being sabaeus with the difference to other taxa exceeding block-jacknifing standard deviations by two orders of magnitude. When calculating the statistics across windows of 50000 SNPs, θ_k is lowest for *sabaeus* in 99.4% of the windows (0.2% *aethiops*) and τ_k is highest for *sabaeus* in 93.7% of the windows (4.3% *aethiops*) (Supplementary Fig. 7d).

We suggest that the tree topology inferred by Pfeifer[59], which places *aethiops* as outgroup to other vervet taxa, is confounded by the fact that they apparently only used fixed differences to estimate divergence times without taking segregating variants into account. As opposed to pairwise genetic differences used to construct the neigbor-joining tree presented here, the number of fixed differences depends on effective population sizes within the taxa. The overall lower effective population size in *aethiops* (see section *Genetic relationships among vervet groups and SIV strains*) leads to an increased number of fixed differences between *aethiops* and other taxa (see relatively high $\mathrm{F_{ST}}$ values of *aethiops* with other taxa; Fig.1c, below diagonal) without increasing absolute divergence (Fig. 1c, above diagonal). Furthermore, the inferred gene flow between *sabaeus* and *tantalus* is expected to substantially reduce fixed differences between the two even if average pairwise sequence divergence is only mildly affected.

D-statistic on the macaque-aligned callset

We used admixtools to recompute the D-statistic for all triplets of African vervet samples consistent with the taxonomic tree. The macaque reference sequence was used as outgroup. The results are highly consistent with D-statstic calculated across whole populations on the original callset aligned against the vervet reference genome (Spearman's r $= 0.98$, $p < 10^{-38}$, Supplementary Figure 36, Supplementary Data Table S2), suggesting that the results presented in Fig. 2, d and e are robust with respect to choice of reference genome, population subsampling, and individual coverage. In particular, for 51 out of 56 comparisons the D-statistic has the same sign in both analyses. The five comparisons for which this is not the case have small values of the D-statistic (D < 0.015) and are not highly significant in either the original or the macaque-aligned analysis (at least one Z-score < 5; Supplementary Fig. 36). D scores are generally smaller in the macaque-aligned data set which might be the result of increased noise from imperfect alignment and erroneous variant calls. Comparisons that have *sabaeus* from Ghana and Gambia as h1 and h2 show the largest difference in D between the analyses ($\approx 60\%$ smaller in the macaque aligned callset). To check whether this discrepancy is caused by population structure (i.e. choice of individuals), we re-ran Admixtools on the original vervet aligned callset, but only included the individuals used in the macaque aligned data set. We found that this third datset is very close to the original vervet aligned callset that used all individuals (Supplementary Data S2), so the difference between the two alignments is more likely explained by some kind of reference bias which does not change the overall conclusions, however.

Supplementary Note 2: Additional results on population histories

MSMC results support the prevalence of ancient gene flow across vervet taxa (Fig. 2, b and c; Supplementary Figs. 15 and 16). In particular, for the south- and east-African group of *hilgerti/cynosuros/pygerythrus*, MSMC suggests that *cynosuros* and *pygerythrus* first started to become reproductively isolated from *hilgerti* approximately 500 thousand years ago (kya), possibly representing the southward migration of their common ancestral population (Supplementary Fig. 16). However, $\sim$220 kya South African *pygerythrus* start to become more isolated with their level of gene flow with *cynosuros/hilgerti* dropping below the level of gene flow between those two populations. Then, $\sim$70 kya the picture changes a second time, with the level of genetic exchange between *cynosuros* and South African *pygerythrus* plateauing at a level of $\sim$10% until total separation $\sim$20 kya. In contrast, *hilgerti* became totally isolated 40 to 50 kya.

Supplementary Note 3: Complementary analyses of selection scans

Pairwise XP-CLR comparisons

To reduce the amount of data and to enrich for genomic regions repeatedly under diversifying selection, we computed and analyzed a root mean square summary of the 30 XP-CLR comparisons (Supplementary Figs. 17 to 22). However, summarizing over many comparisons also harbors a certain risk of amplifying confounding signals and biases.

To test whether the strong enrichment of high selection scores in the GO category *viral process* (GO:0016032) critically depends on the way selection scans were averaged, we performed independent GO enrichment analyses for each of the pairwise 30 XP-CLR comparisons across taxa, following the same methodology as for the root mean square average. For each of the

enrichment studies, we extracted the p-value of the category *viral process* and reported its relative rank among GO terms. We found *viral process* to be significantly enriched in 28 out of 30 XP-CLR scans, with a median p-value of 1.06×10^{-6} and a median rank of 3 (among ~ 5700 GO terms; mean rank 25). We conclude that the significant enrichment of this category is not specific to the root mean square average but a general feature of multiple XP-CLR comparisons.

Background selection checks

Negative selection on linked deleterious alleles (background selection) can produce genomic signatures similar to local adaptation. To check whether our results are dominated by signals of negative selection we compared our selection scores (and the resulting GO enrichments) to two proxies for background selection: Neutrality index and percentage of conserved elements.

Neutrality index of genes and gene categories

If our selection scores were mainly determined by the strength of background selection rather than diversifying selection, then we would expect sites with higher impact to show less evolutionary change, and hence a strong positive correlation between selection score and the neutrality index, $NI = (Pn/Ps)/(Dn/Ds)$. ($NI > 1$ suggests negative selection, while $NI < 1$ is a sign of positive selection).

Using the nucmer vervet-macaque whole genome alignment (Online Methods), macaque alleles were added to the beagle-imputed and shapeit-phased vervet SNP calls for sites where vervet and macaque alleles differed. Variant effects were annotated for the combined set of vervet specific and vervet-macaque differentiating variants using SNPeff and the ensembl vervet annotation 1.78. The ensembl annotation was used rather than the NCBI annotation, because the latter led to numerous errors in the effect annotation. However, the vast majority of genes overlap between the two annotations.

For each gene, we calculated Dn and Ds as the number of SNPs for which the macaque allele was different to the major allele in the vervet callset, for non-synonymous and synonymous mutations, respectively. Similarly, we calculated Pn and Ps as the respective numbers of variants polymorphic in the vervet callset. We calculated NI the 5214 genes for which we both obtained an XPCLR score and all of Pn, Ps, Dn, Ds were greater than zero ($\log NI$ in the interval $(-\inf, \inf)$).

We found that selection scores and NI are negatively correlated (Supplementary Figure 27, Pearson's $r = -0.137$, $p < 10^{-22}$, using $\log NI$), suggesting that overall there is no signal of stronger background selection in genes with higher selection score. Results were similar when using maximum selection scores across exons.

To further test whether gene category enrichments found in our selection scan were primarily driven by purifying or background selection, we next calculated the neutrality index for each GO category and compared the results to their enrichment p-values. For each GO category, we summed the respective numbers of Dn, Ds, Pn, and Ps across genes and calculated NI from the sums. This approach has the advantage that it does not require Pn, Ps, Dn, Ds to be greater than zero for every single gene, and it reduces the stochasticity coming from counts in very short genes.

We obtain an average NI across GO categories of 1.07. We note that due to Simpsons paradox, when calculating NI across multiple genes, the expected NI under neutrality is not necessarily 1 if the level of polymorphism correlates with the level of constraint[82]. Comparing relative values of NI between GO categories, we do not see particularly large NI in GO categories that are highly enriched for selection signals (Supplementary Fig. 28), suggesting no strong excess of purifying selection in these categories. In particular, the category viral process shows an NI within the 11% lowest percentile compared to random sampling of the same number of genes (Supplementary Fig. 29; $\log NI = -0.022$), suggesting that genes in this category

where overall under weaker purifying selection than average.

Percentage of conserved elements

As a second control whether our results could be dominated by background selection, we downloaded evolutionary conservation scores for the 10-primate-subset of a 46 vertebrate species alignment (phastCons46way, http://hgdownload.cse.ucsc.edu/goldenPath/hg19/phastCons46way/primates/)[32], and calculated for each gene the percentage of conserved coding sequence, weighting all splice variants equally. Conserved sequence was determined as sites for which the phastCon score was greater than 0.5, corresponding to a greater than 50% change of being conserved. We find that the percentage of conserved elements across genes shows weak but significant correlation with our selection scores (Pearson's r$= 0.11$, $p < 10^{-50}$). We used percentage of conserved sequence of each gene to run the same GO enrichment analysis as for our selection scores using TopGO. Comparing the results, we note that GO enrichments are generally much less significant for conserved elements (38 vs. 155 scores significant at $p < 0.01$, smallest p-value $2.3 * 10^{-4}$ vs. $2.6 * 10^{-17}$, Supplementary Fig. 30). There is some overlap in significantly enriched GO terms, most notably, *proximal/distal pattern formation* and *G1/S transition of mitotic cell cycle* (Supplementary Fig. 30), but most of the significant selection score enrichments are not significant using the percentage of conserved elements. In particular, the highly significant and virus-related categories *positive/negative regulation of transcription of RNA polymerase II promoter*, and *viral process*, are not significantly enriched for conserved elements.

Next we computed the sumstat test for enrichment of conserved elements in human-HIV-1 interaction gene categories and compared it to the enrichments for vervet selection scores (Supplementary Fig. 32). We found that many human-HIV-1 interaction categories are highly enriched for conserved elements, suggesting that these categories are not very specific. How-

ever, there is no particularly strong correlation between the two analysis. While, some of the very large gene categories such as *Tat_interacts_with*, *Vif_interacts_with* and *Rev_interacts_with* are highly significant in both analysis, others are only highly significant for selection scores, e.g., *Gag-Pol_complexes_with*, *Envelope_surface_glycoprotein_gp120_complexes_with*, and *Pr55(Gag)_complexes_with*.

Finally, we tested whether the genes in any of the 35 WGCNA modules show significant enrichment for conserved elements. While we have shown that five out of 33 modules with vervet-specific expression changes show significant enrichment for high selection scores (FWER < 0.05), there is only one (different) module marginally enriched for conserved elements (FWER $= 0.049$). This category ("grey60", row 5, column 2 in Supplementary Fig. 34) loads on a biologically irrelevant expression change in rhesus macaque 40 days before exposure to the virus.

Taken together, the results of this note lead us to conclude that while our selection scan is likely to pick up some signals of negative selection, in particular in developmental categories, there is no indication that the most significant and virus-related enrichments are driven by background selection.

Supplementary Data Tables

Data 1: Table of sample IDs, taxonomic group attribution (c.f., Fig. 1a), taxonomic classification from the Integrated Taxonomic Information System (http://www.itis.gov/ last accessed June 2016), collection site, country, coordinates, actual fold coverage, percentage of mapped reads (including all scaffolds), SRA Sample ID, and BioProject Accession number.

Data 2: Results for all D-statistic (ABBA-BABA test) comparisons that are consistent with the UPGMA clustering tree of pairwise differences. Z-scores were obtained through block-jackknifing. Samples were grouped by country. Fig. 2e and Supplementary Figs. 14 and

36 show a subset of the data. See Supplementary Table 4 for IDs of the samples used in the single-sample analysis.

Data 3: Average and maximum of XP-CLR root mean square average selection scores for each gene. Details on on how these scores were obtained are given in the Online Methods.

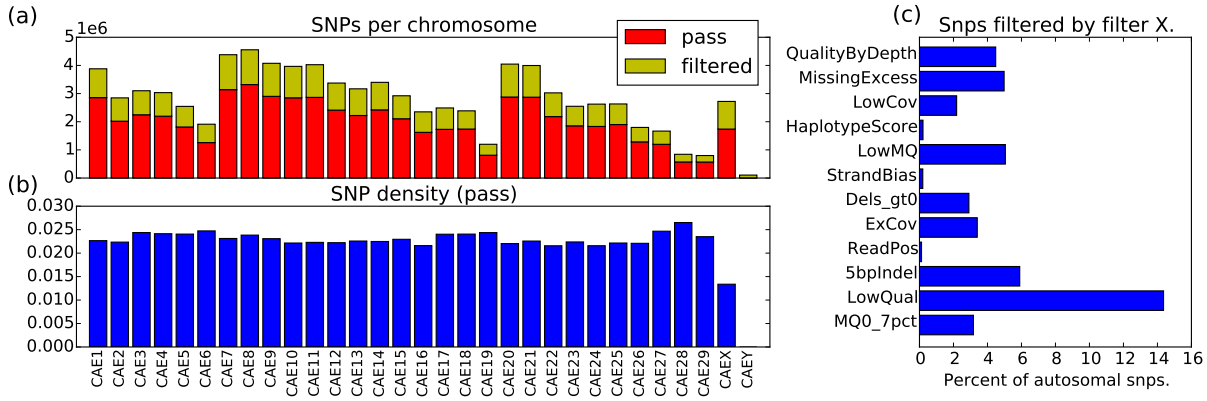
Data 4: Significance p-values for enrichment of selection scores in gene ontology categories using the R-package TopGO with a Kolmogorov-Smirnov test and the weight01 algorithm for all categories with $p < 0.1$.

Data 5: Significance p-values for sumstat enrichment of selection scores in NCBI HIV-1 human interaction gene categories (See Online Methods).

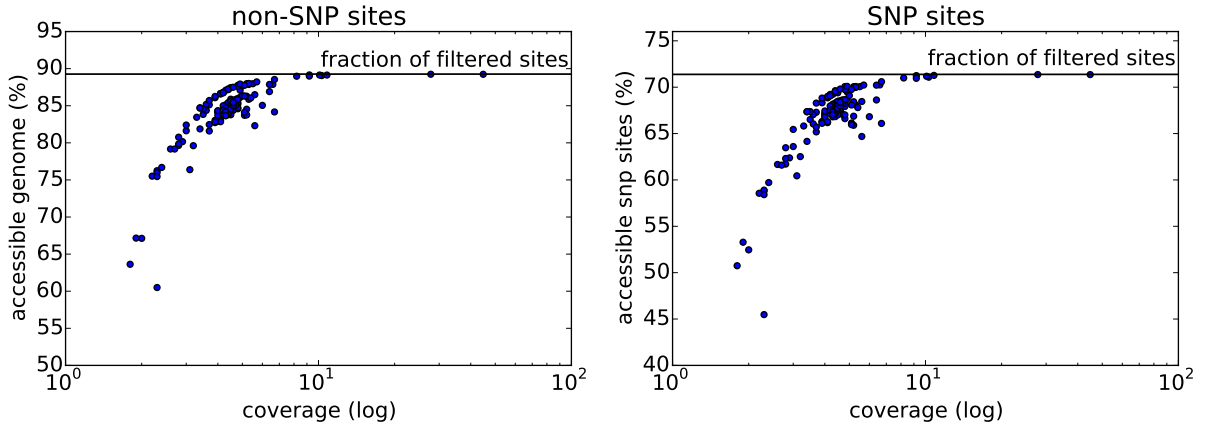
Data 6: Significance p-values for enrichment of selection scores in gene expression categories (See Online Methods).

Data 7: Significant GO enrichments for WGCNA modules significantly enriched in high selection scores that only show a short term response in vervet (mainly day 6 post infection), i.e., for genes from the green, blue, and magenta modules with asterisk in Supplementary Fig. 35. R-package TopGO with Fisher's exact test and weight01 algorithm was used.

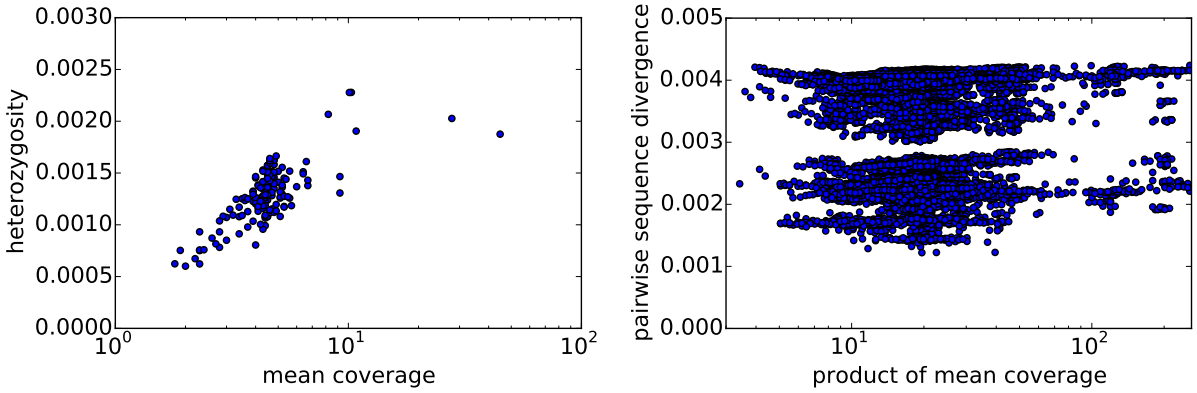
Data 8: Significant GO enrichments for GO-enrichment for WGCNA modules significantly enriched in high selection scores that show a long-term response in vervet (day 115 post infection), i.e., for genes from the yellow and tan modules with asterisk in Supplementary Fig. 35. R-package TopGO with Fisher's exact test and weight01 algorithm was used.



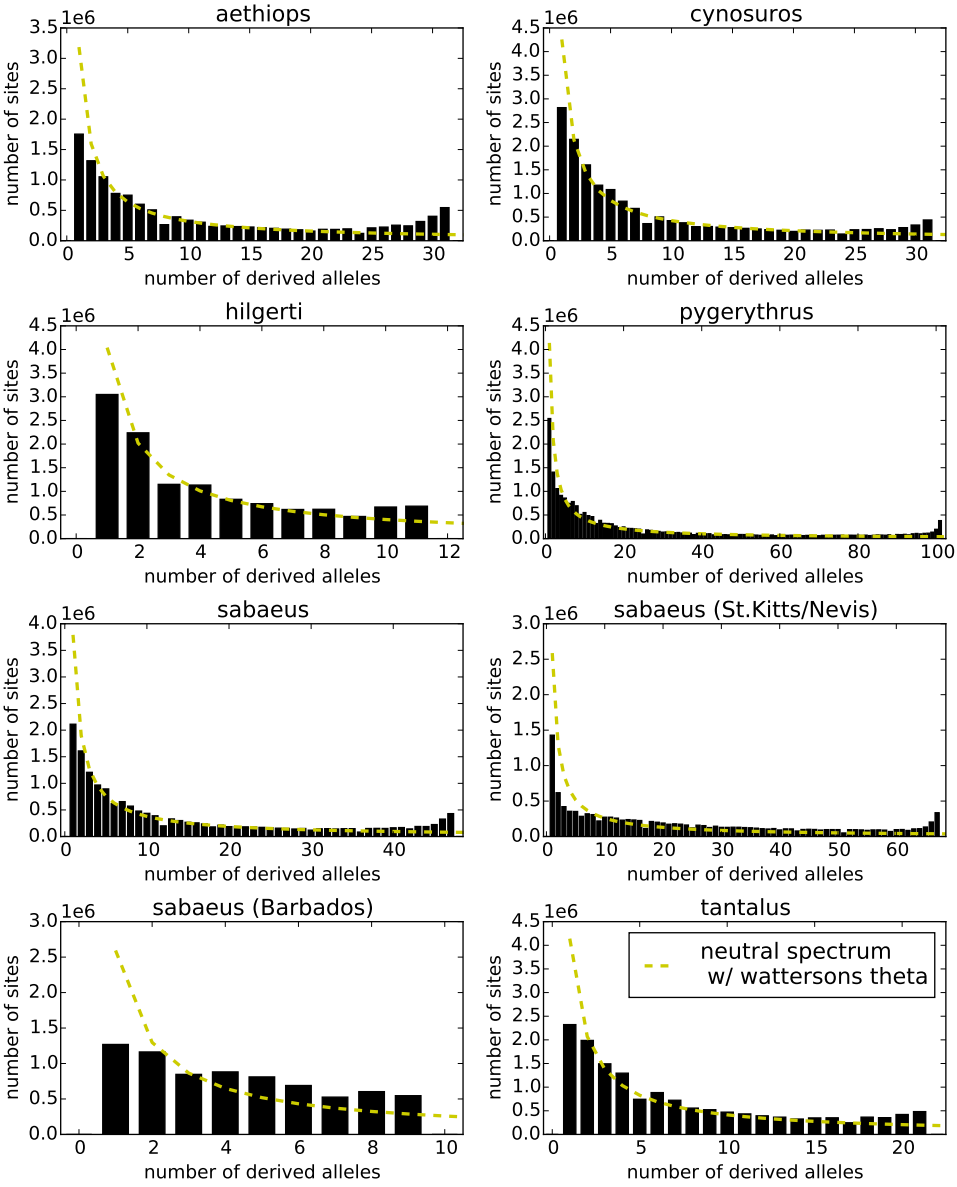
Supplementary Figure 1: SNP filtering statistics. (a) SNPs per chromosome. (b) SNP density of pass SNPs per chromosome. (c) Percentage of autosomal SNPs that do not pass a filter. Filters are described in Supplementary Table 1. Note that a SNP can fail multiple filters. Hence, the percentages sum up to more than the total fraction of filtered SNPs.



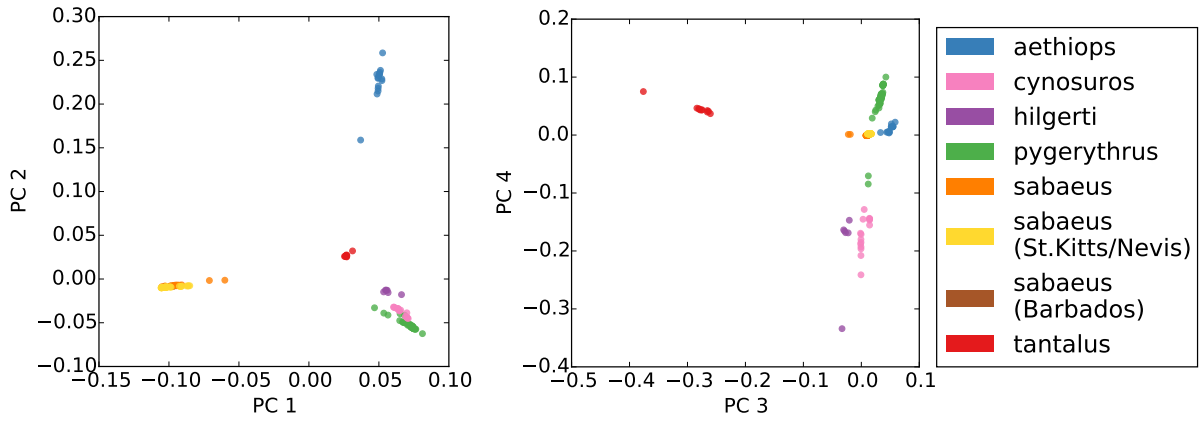
Supplementary Figure 2: Proportion of the genome that is accessible for analysis for non-ariant sites (left) and SNP-sites (right) as a function of coverage. A site is inaccessible either because it did not pass our filtering criteria (c.f., Supplementary Table 1 and Supplementary Fig. 1) – this contribution is the same for all individuals and represented by the horizontal line – or because because the site has a missing genotype call in a given individual.



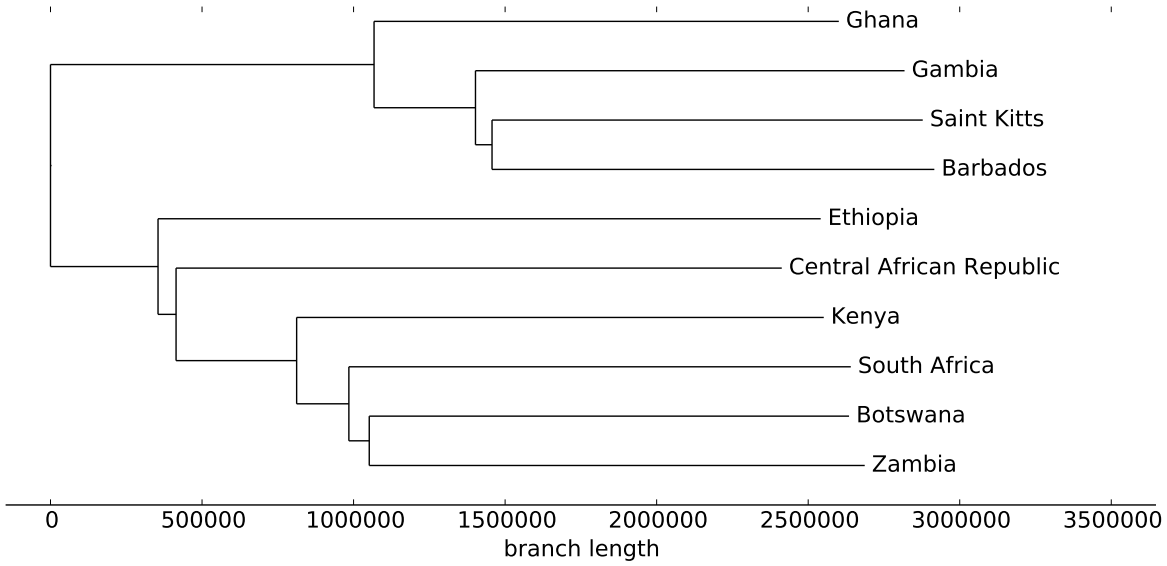
Supplementary Figure 3: Comparison of individual heterozygosity (left panel) and pairwise sequence divergence (right panel) to individual mean coverage and product of individual coverages (x-axis), respectively.



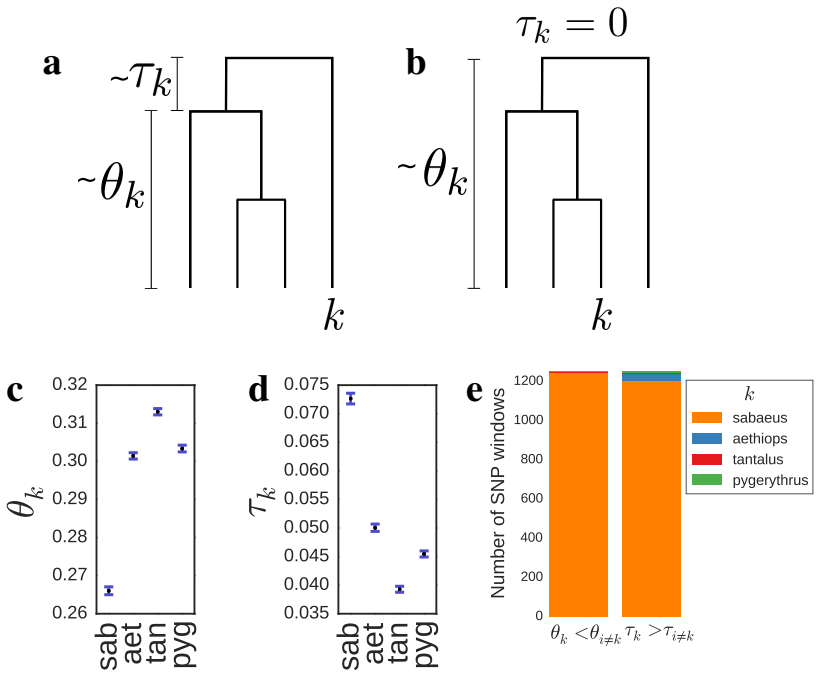
Supplementary Figure 4: Site frequency spectra within taxonomic groups. Note that singleton numbers are biased by sequencing coverage because their calling-rate is much lower in low coverage individuals.



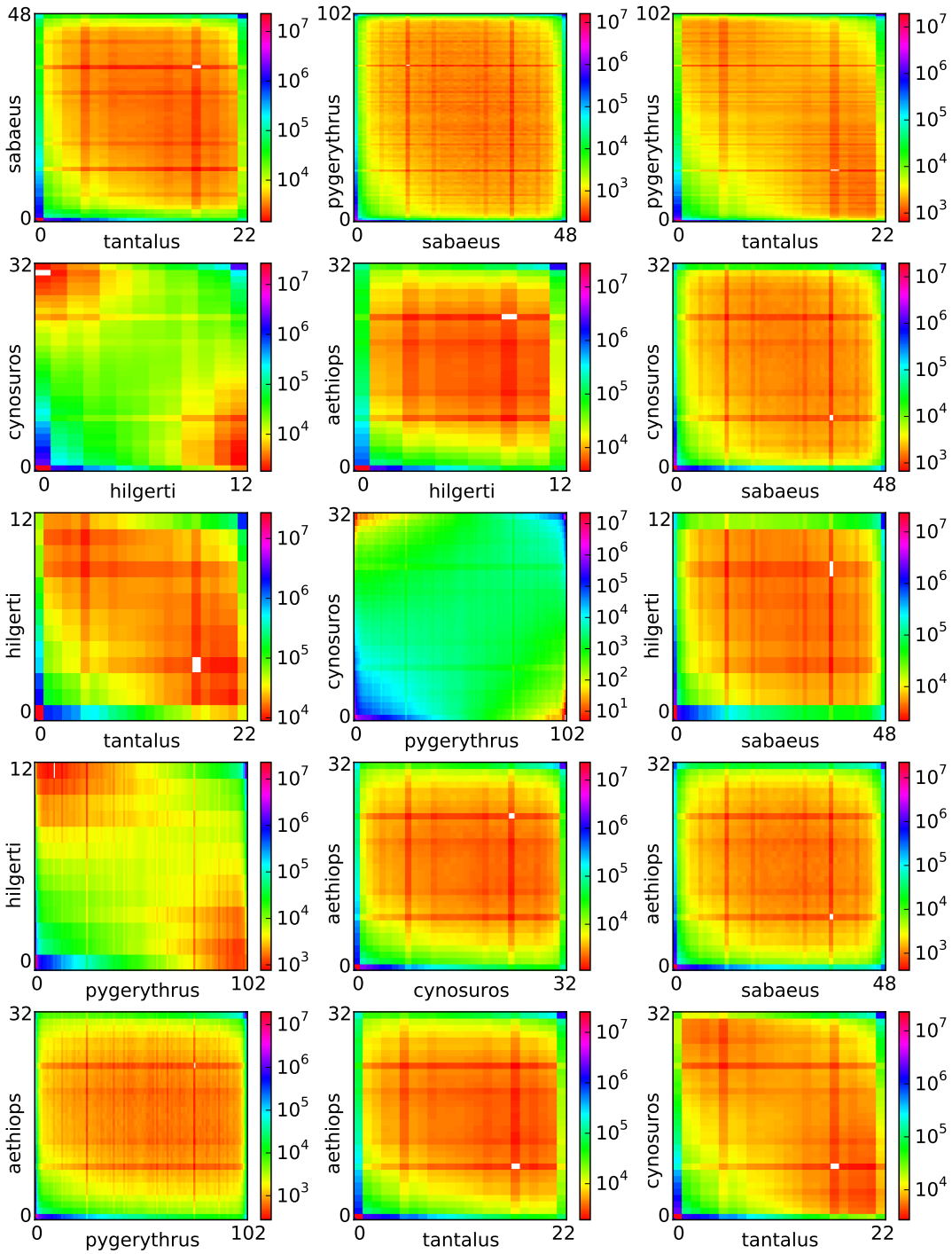
Supplementary Figure 5: Principal component analysis of biallelic SNPs computed using PCAdapt[70] in mode *fast*.



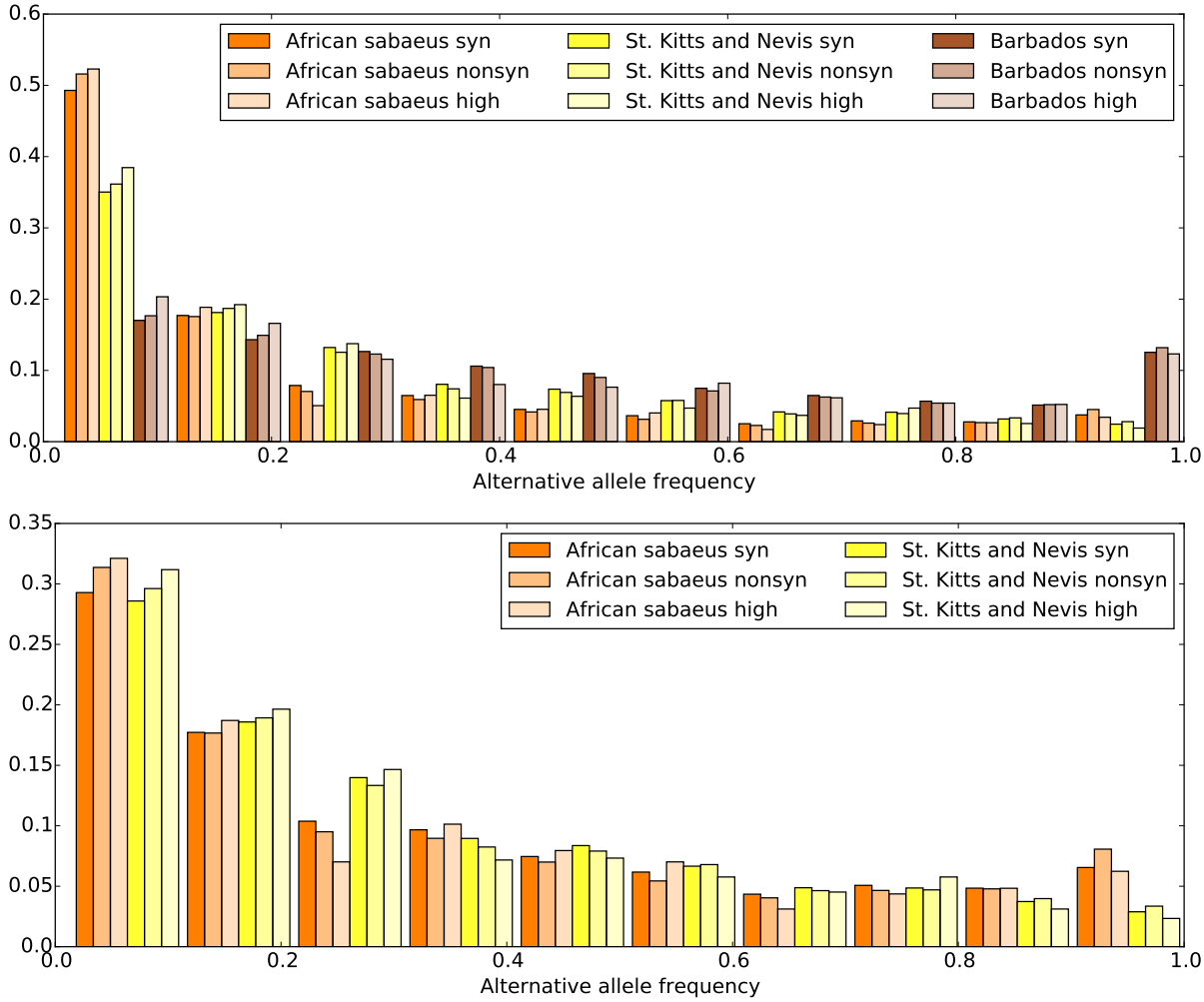
Supplementary Figure 6: Neighbor-joining tree of pairwise differences using single individuals and alignment against rhesus macaque (Mmul 8.1.). Sample ids are given in Supplementary Table 4.



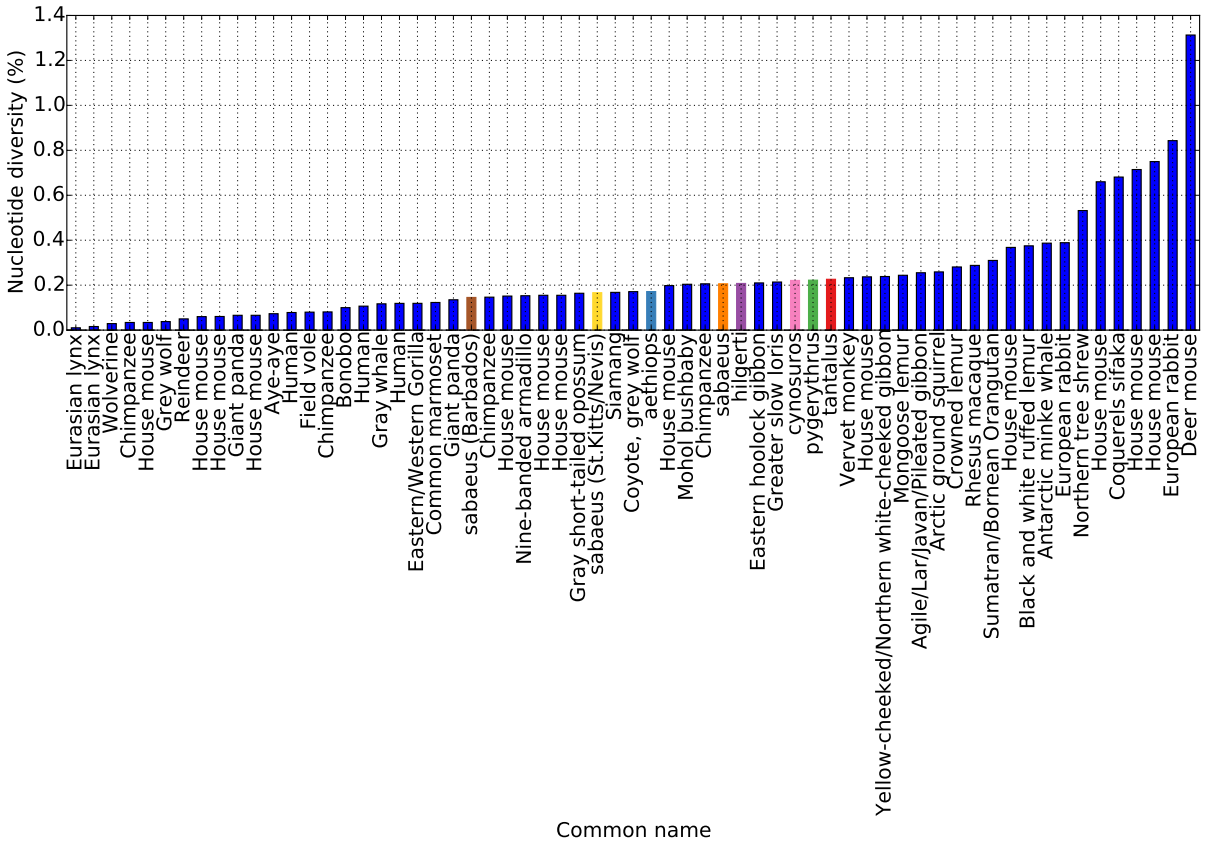
Supplementary Figure 7: Coalescent statistics for outgroup inference. (a) and (b) Symbolic representation of the information captured by statistics θ_k and τ_k on a given coalescent tree if (a) k coalesces with the other samples in the most recent common ancestor (MRCA) or (b) if k coalesces with an other sample before the MRCA. Note that in practice the computed values are averaged over multiple coalescent trees and over the two alleles in each diploid individual. (c) Whole genome estimate of θ_k for different k. (d) Whole genome estimate of τ_k for different k. (c) and (d) Values are normalised by the total number of SNPs. Error bars correspond to ± 3 block-jackknifing standard deviations. (e) Number of windows of 50000 SNPs each, for which sample k shows lower θ_k and higher τ_k than all other taxa (left and right bar, respectively) consistent with falling onto the outgroup branch of the species tree.



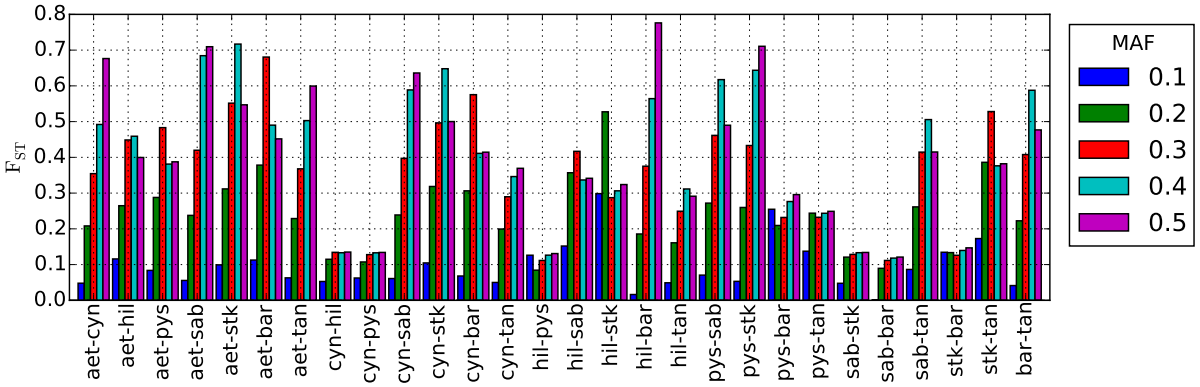
Supplementary Figure 8: Two taxon joint site frequency spectra. The axis units are the number of derived alleles (allele count) and the colour gives the number of SNPs for each combination of allele counts in the two population.



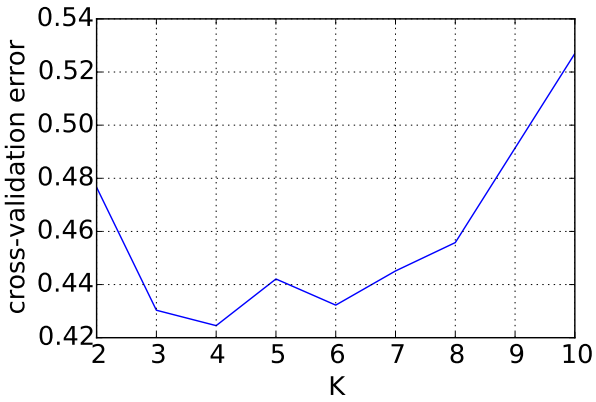
Supplementary Figure 9: Site frequency spectra for synonymous, non-synonymous and high impact mutations for African and Caribbean *sabaeus* vervets. The top panel shows all variants, and the bottom panel only variants which are shared between African *sabaeus* and *sabaeus* from Saint Kitts and Nevis.



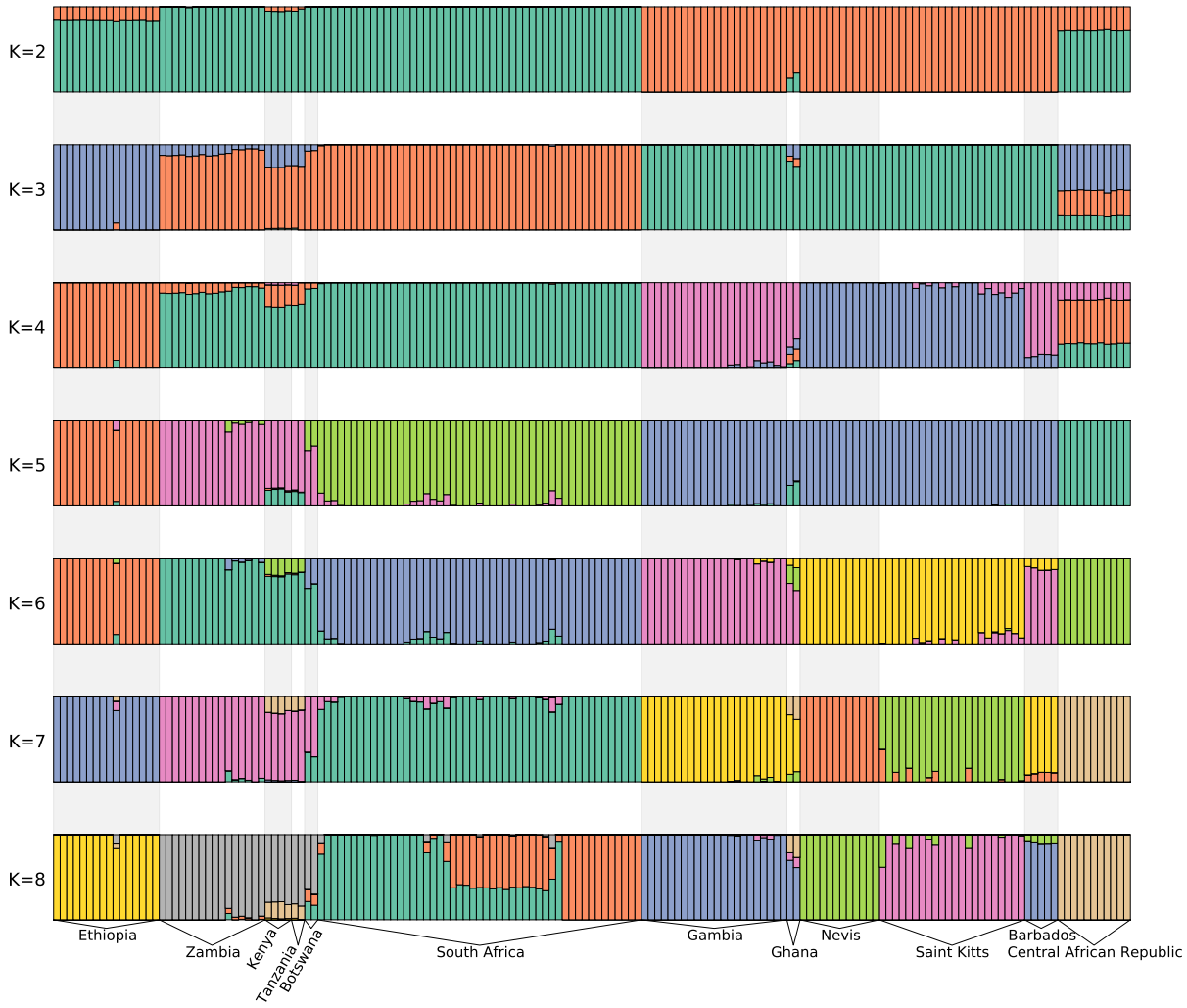
Supplementary Figure 10: Nucleotide diversity (π) compared between vervet monkey groups and other mammals. Data for other mammals (blue bars) were taken from Leffler et al.[19]. References and more details on how these values were obtained can be found there.



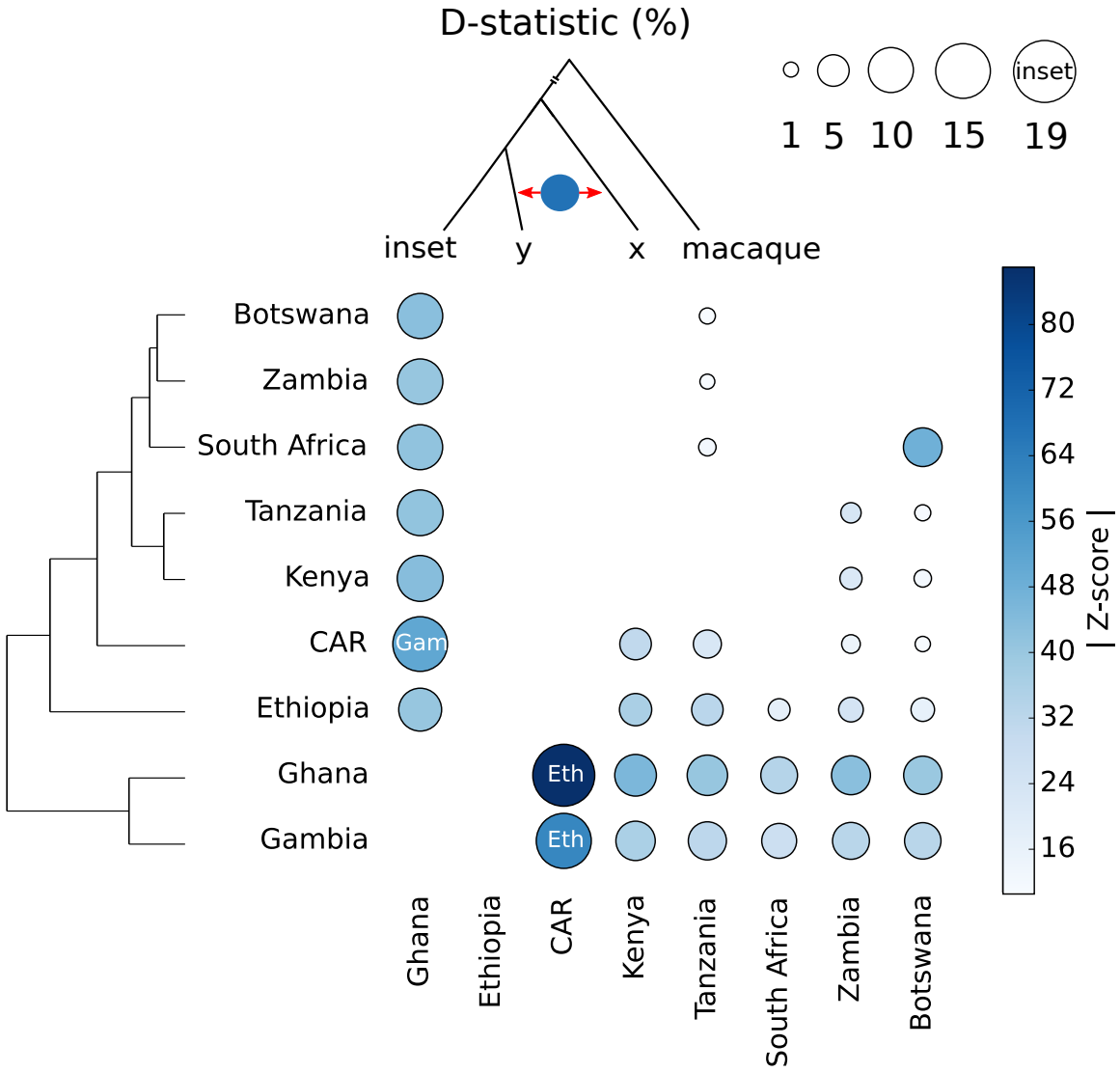
Supplementary Figure 11: Average F_{ST} values binned by minor allele frequency for each pairwise comparison. Label abbreviations correspond to the first three letter of the taxon name in Fig. 1a for African taxa, and the first three letters of the country name for Caribbean groups.



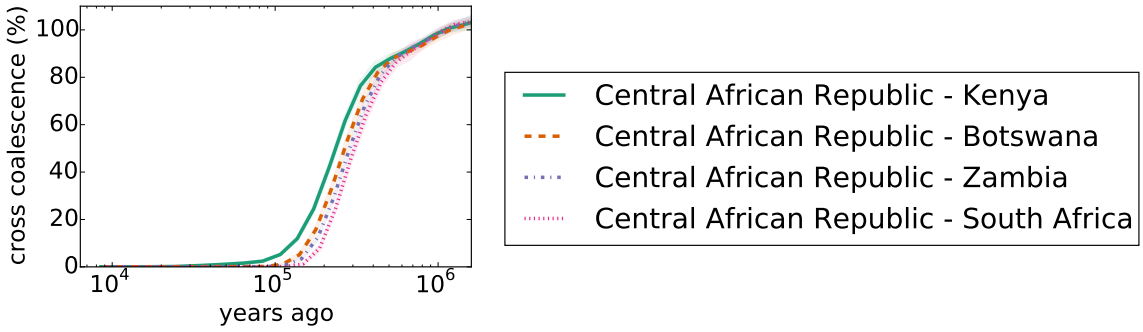
Supplementary Figure 12: Admixture cross-validation error for $K = 2 - 10$. Note that the run for $K = 9$ is missing due to convergence problems. While there are local minima of cross-validation error at $K = 4$ and $K = 6$, we chose to present $K = 5$ in Fig. 2a, because at the global minimum of cross-validation error at $K = 4$ all individuals of the group *tantalus* are consistently shown as three-way mixtures between three other groups. We do not think that this pattern corresponds to any biologically relevant admixture event, but think that it is an artefact of not using a large enough K value. We think the reason that at $K = 4$ the Caribbean populations of St. Kitts and Nevis get their own cluster rather than *tantalus* is simply because we have a larger sample of St. Kitts and Nevis and they are genetically more homogeneous than other groups.



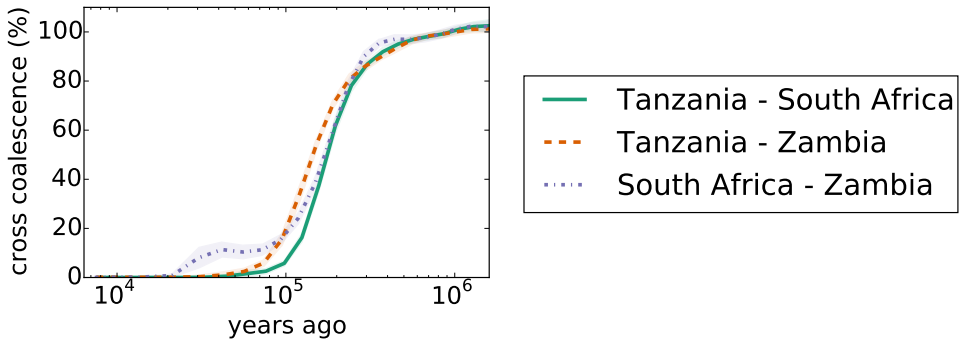
Supplementary Figure 13: Admixture results for $K = 2 - 8$.



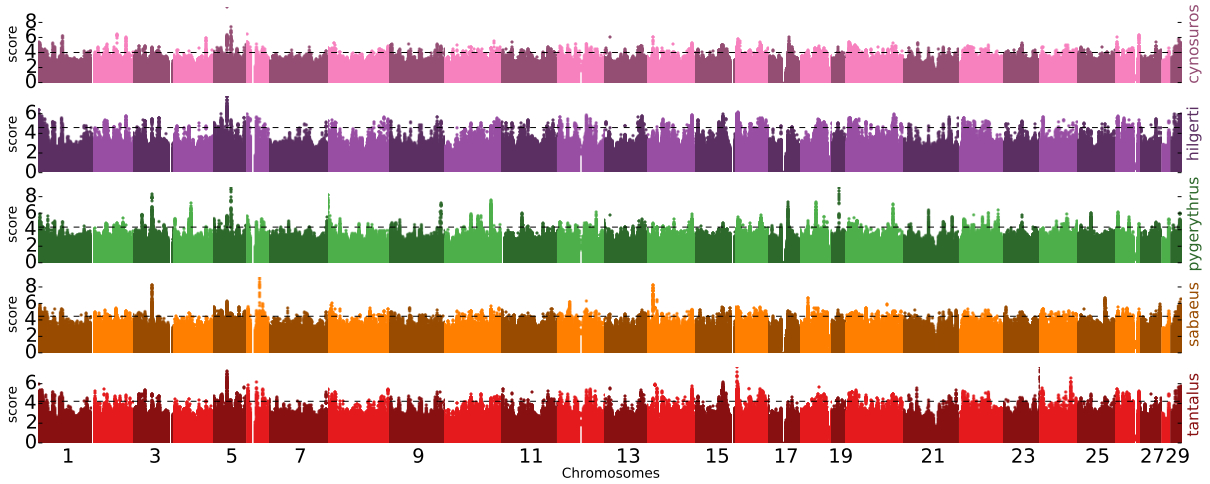
Supplementary Figure 14: D-statistics across countrie. Maximum value of D supported for each pair of individuals across all possible ingroup controls. For the largest values, the ingroup control is given as inset (for all comparisons, see Supplementary data 2). Missing values represent no support for gene-flow between these groups. Note that for a given row or column, all but the largest circles can be correlated responses due to shared drift of these taxa with the taxon corresponding to the largest circle.



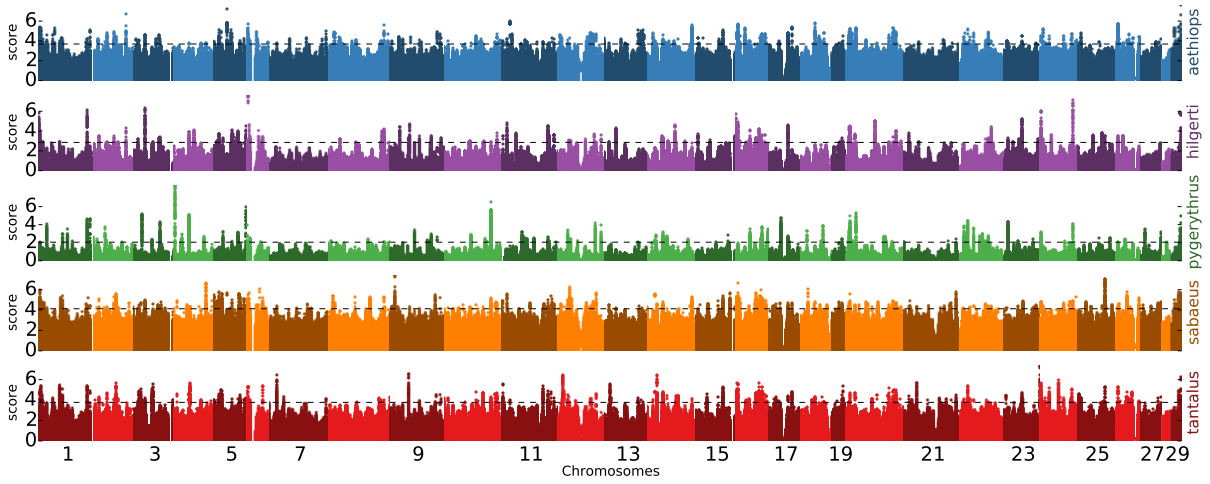
Supplementary Figure 15: MSMC cross-coalescence rate for *tantalus* with *hilgerti*, *cynosuros*, and *pygerythrus*. Shaded areas correspond to ± three block-jackknifing standard deviations.



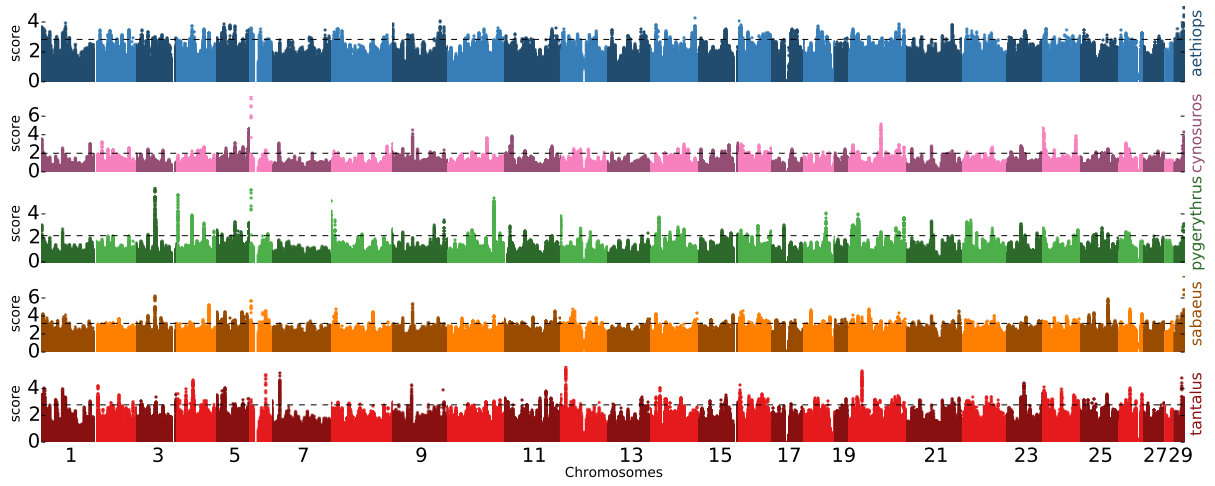
Supplementary Figure 16: MSMC cross-coalescence rate for *hilgerti*, *cynosuros*, and *pygerythrus*. Shaded areas correspond to ± three block-jackknifing standard deviations.



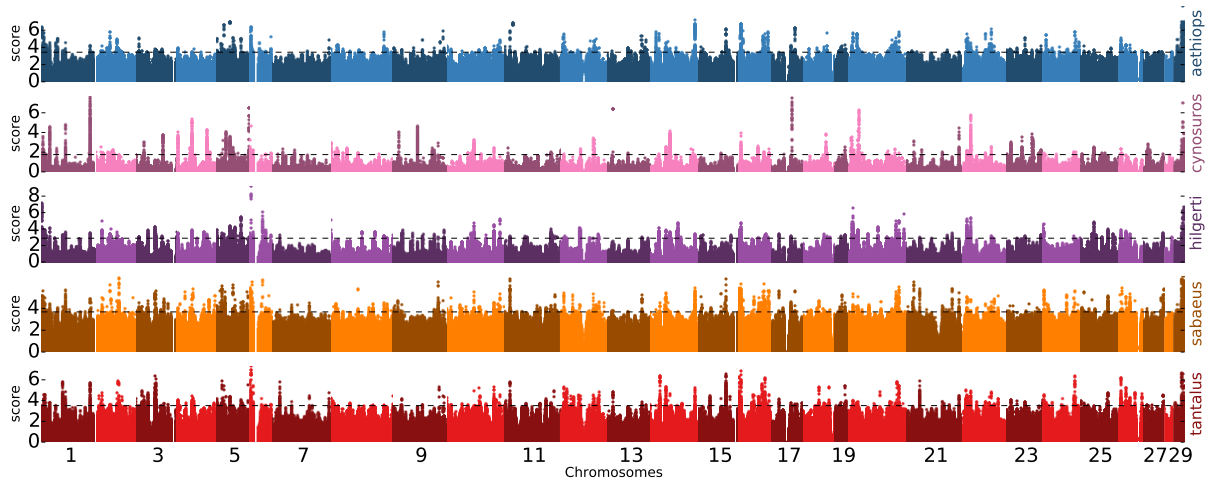
Supplementary Figure 17: Manhattan plot of XP-CLR scores for comparisons of *aethiops* against all other taxa. The dashed line indicates the top 0.1% quantile.



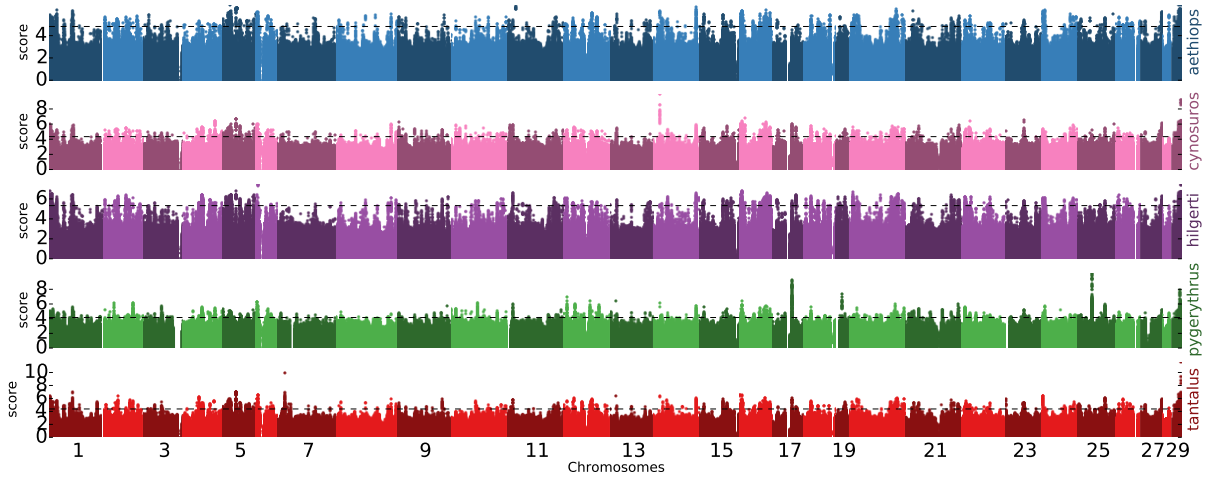
Supplementary Figure 18: Manhattan plot of XP-CLR scores for comparisons of *cynosuros* against all other taxa. The dashed line indicates the top 0.1% quantile.



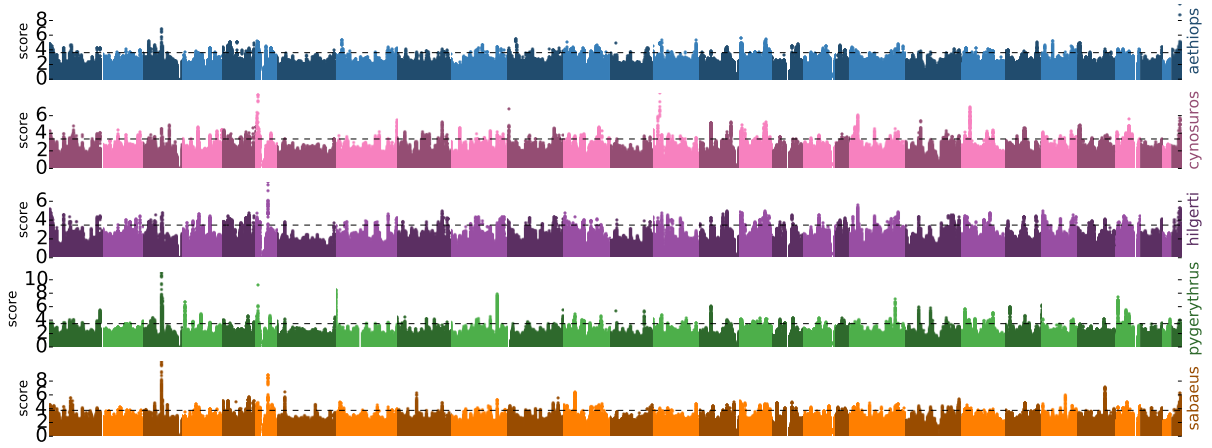
Supplementary Figure 19: Manhattan plot of XP-CLR scores for comparisons of *hilgerti* against all other taxa. The dashed line indicates the top 0.1% quantile.



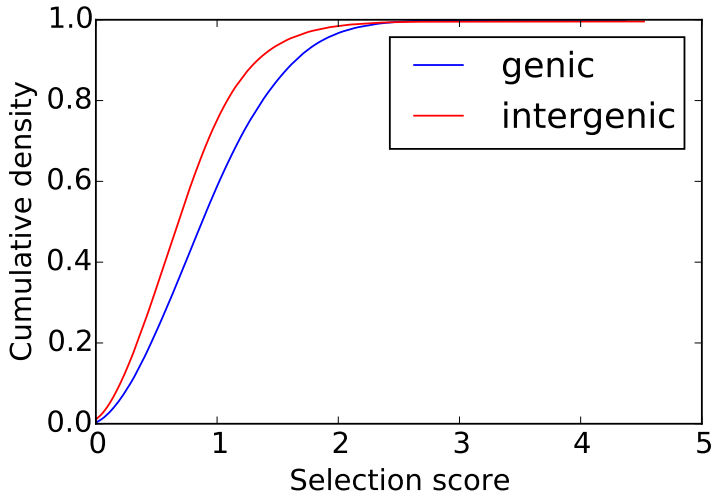
Supplementary Figure 20: Manhattan plot of XP-CLR scores for comparisons of *pygerythrus* against all other taxa. The dashed line indicates the top 0.1% quantile.



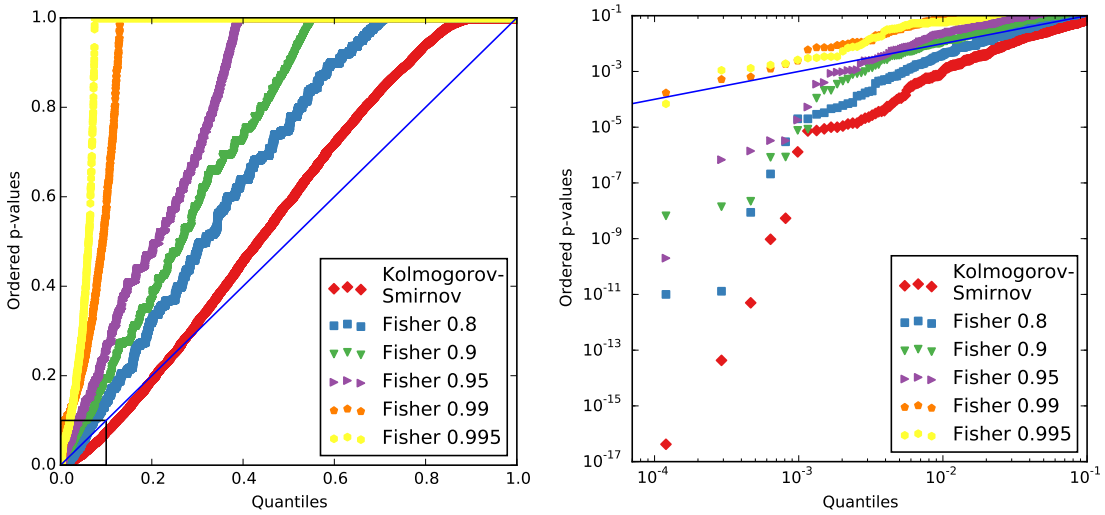
Supplementary Figure 21: Manhattan plot of XP-CLR scores for comparisons of *sabaeus* against all other taxa. The dashed line indicates the top 0.1% quantile.



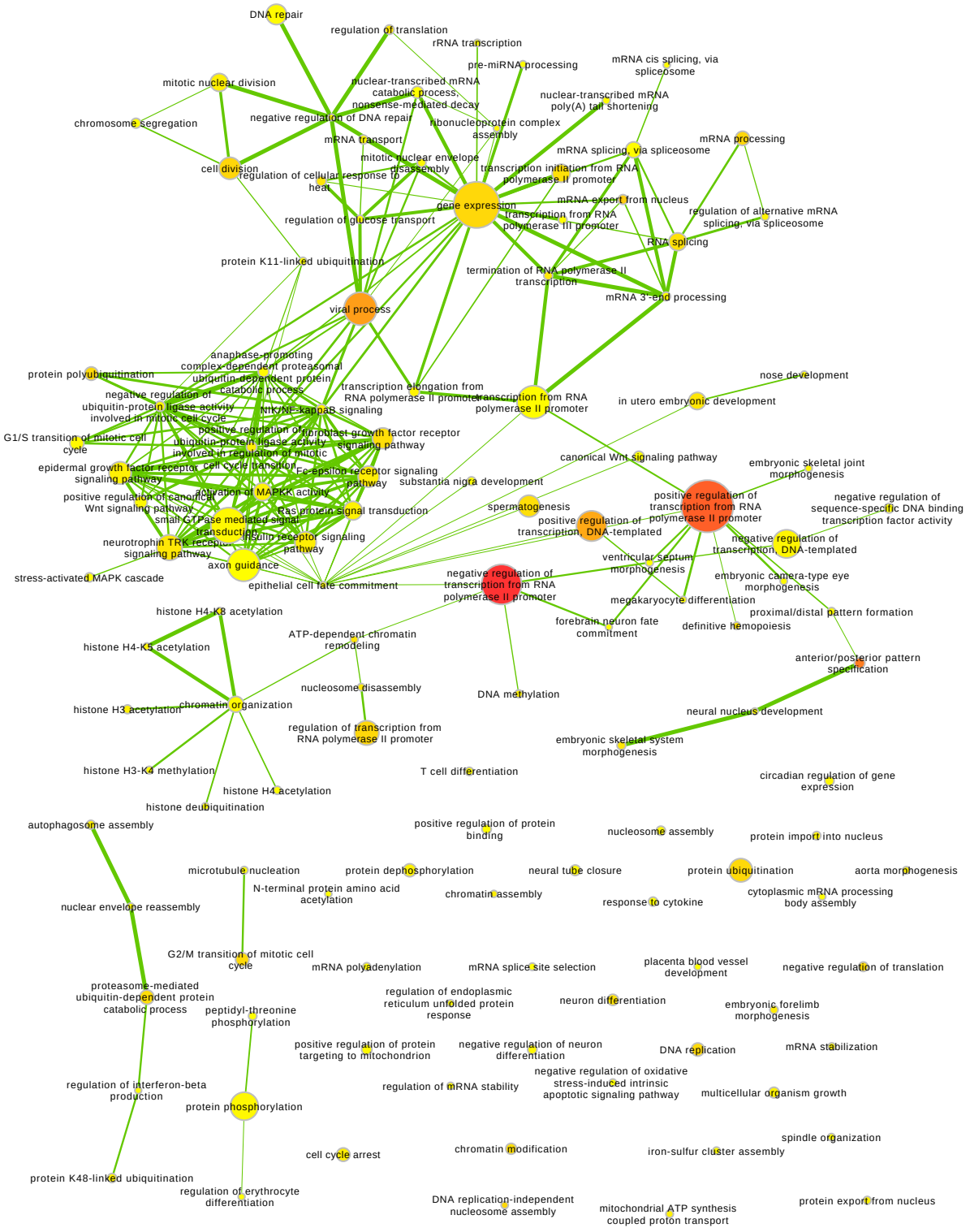
Supplementary Figure 22: Manhattan plot of XP-CLR scores for comparisons of *tantalus* against all other taxa. The dashed line indicates the top 0.1% quantile.



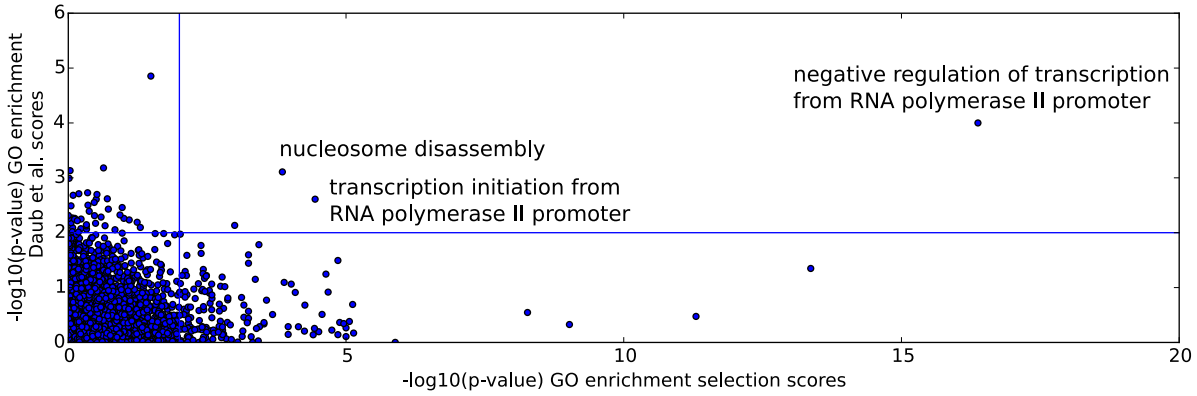
Supplementary Figure 23: Empirical cumulative density function of genic and intergenic selection scores. Genic scores are enriched for higher values.



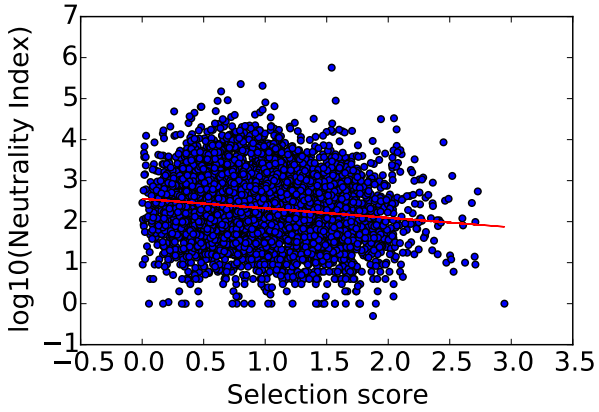
Supplementary Figure 24: TopGo enrichment p-values plotted against the quantiles of a uniform distribution. For Fisher's exact test the method compares a top quantile of genes against the background. The Kolmogorov-Smirnov test takes the scores of all genes into account. The right panel corresponds to a zoom into the rectangle in the left panel.



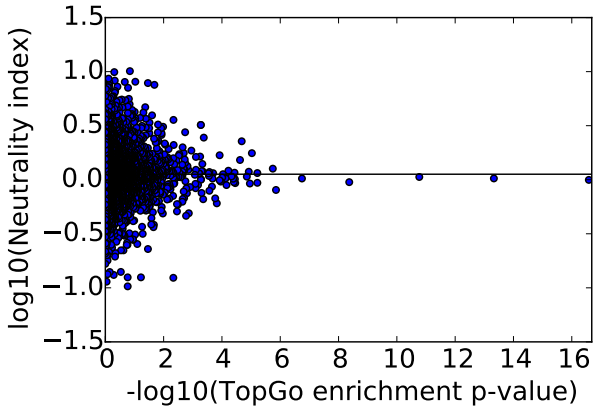
Supplementary Figure 25: Cytoscape enrichment network of enriched GO terms ($p < 0.005$) with TopGo Kolmogorov-Smirnov test with weight01 algorithm. Yellow to red represents linear mapping of -log(p-value) from 2.4 to 16.4. Size represents gene number, edge-width overlap in genes.



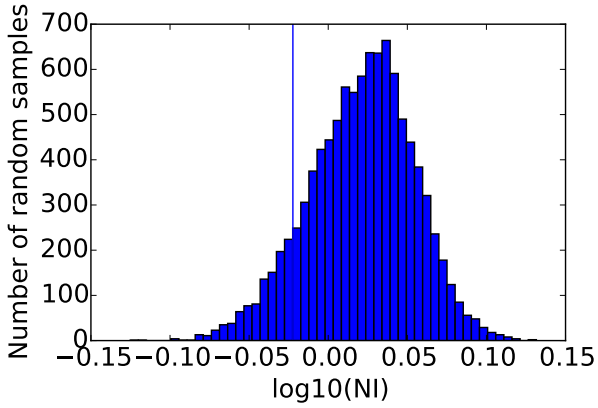
Supplementary Figure 26: Comparisons of TopGO enrichment p-values (on log scale) for vervet selection scores (x-axis) and human selection scores from Daub et al.[13] (y-axis). The blue lines indicate a significance level of $p = 0.01$.



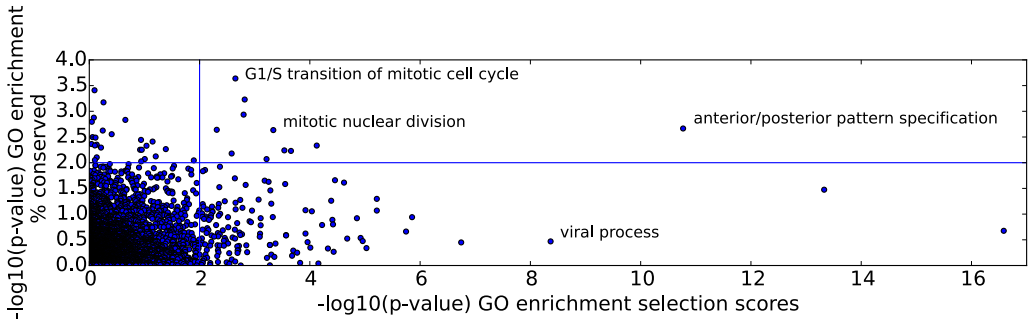
Supplementary Figure 27: Gene mean selection scores plotted against neutrality index (on log-scale). The red line shows the linear regression with slope -0.23 and $R^2 = 0.019$.



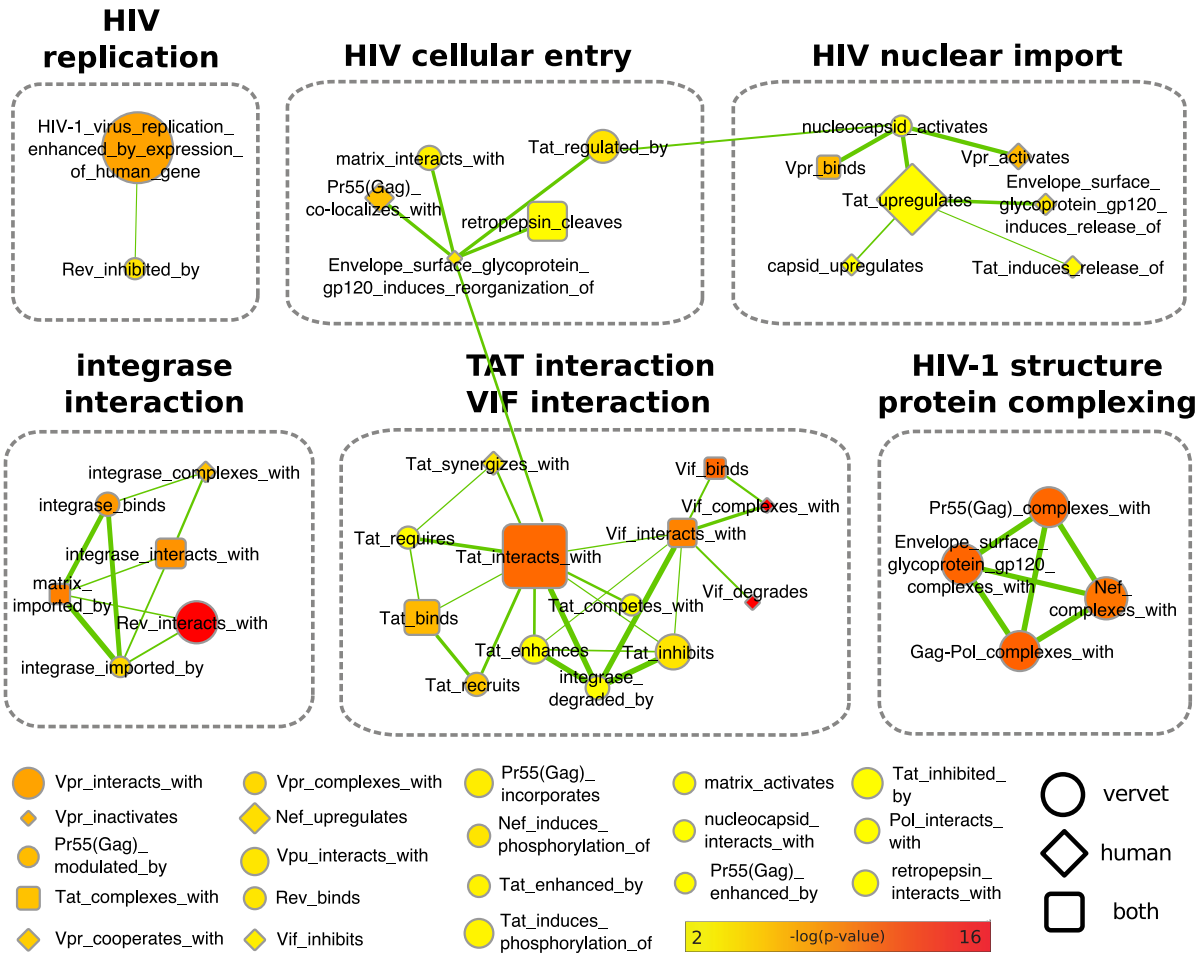
Supplementary Figure 28: TopGO Kolmogorv-Smirnov selection enrichment p-values (log-scale) compared to neutrality index (log-scale). The black horizontal line shows the mean neutrality index across GO categories ($\log_{10}(NI) = 0.05$). Note that both enrichment p-values and neutrality index are effectively calculated as a summary across all genes, in which case the expected NI under neutrality is not necessarily 1[82].



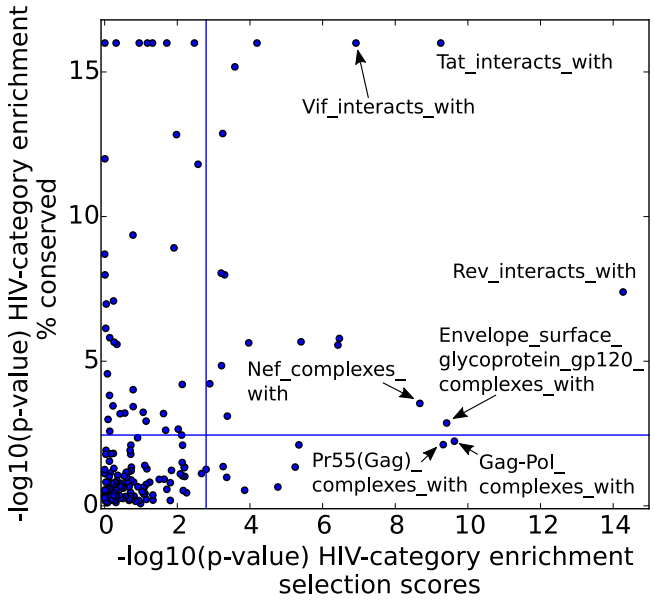
Supplementary Figure 29: Histogram of joint neutrality indices of 10000 random samples of 544 genes compared to the real value of GO:0016032 (viral process; blue line), which contains 544 genes.



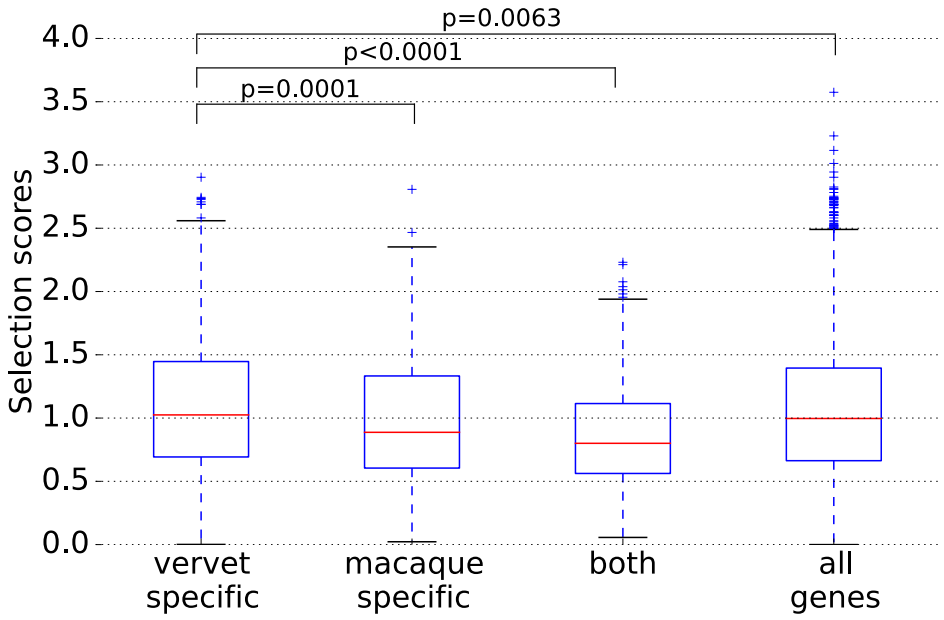
Supplementary Figure 30: Comparisons of TopGO enrichment p-values (on log scale) for selection scores (x-axis) and fraction of the gene conserved, as measured by PhastCons elements (y-axis). The blue lines indicate a significance level of $p = 0.01$.



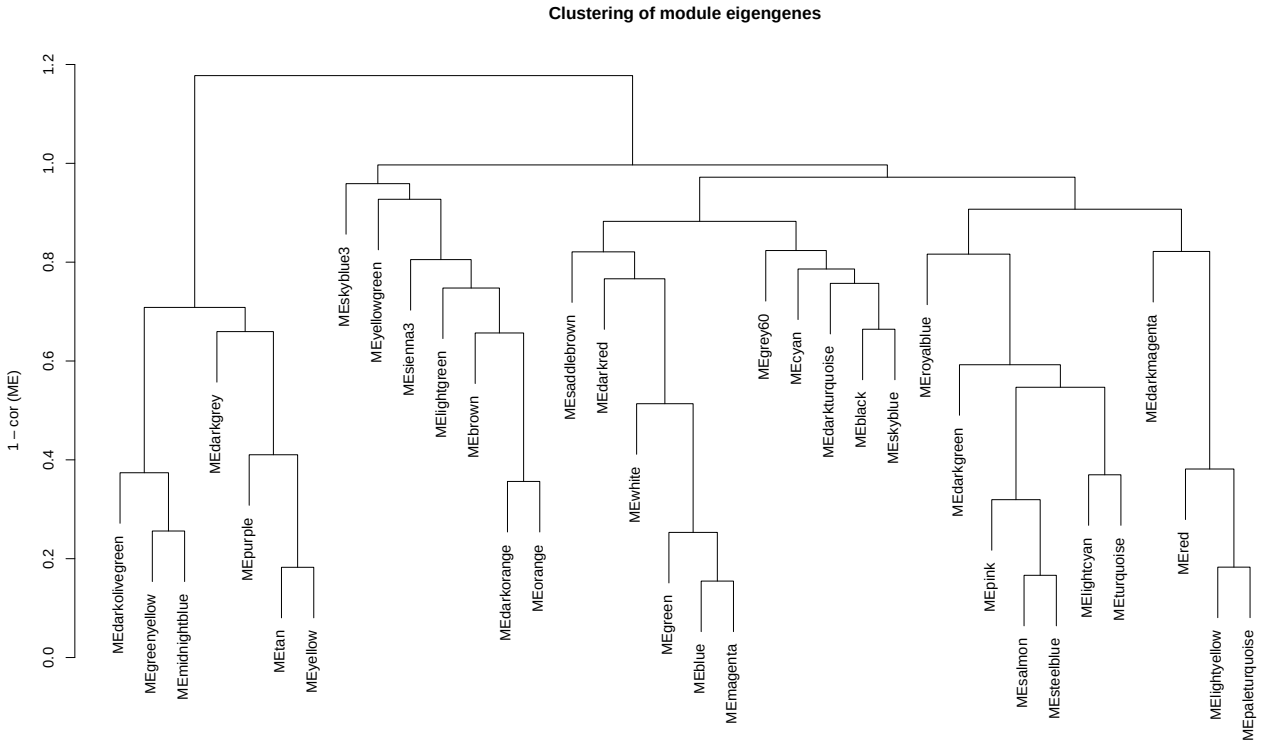
Supplementary Figure 31: Enrichment map network of human-HIV-1 interaction categories enriched for high average gene selection scores in vervet monkeys and humans. Human selection data was taken from Daub et al. 2013[13]. Colors represent p-values on a log scale (red most highly significant, sumstat enrichment test $p < 0.01$ and Bonferroni-Holm FWER < 0.05). Shapes correspond to categories significantly enriched in only vervets (circles), only human (diamonds), or both (squares). Edges represent overlap in genes. Node size represents number of genes in a category (min genes$= 10$; max$= 474$). Terms are grouped using Cytoscape clustermaker[34].



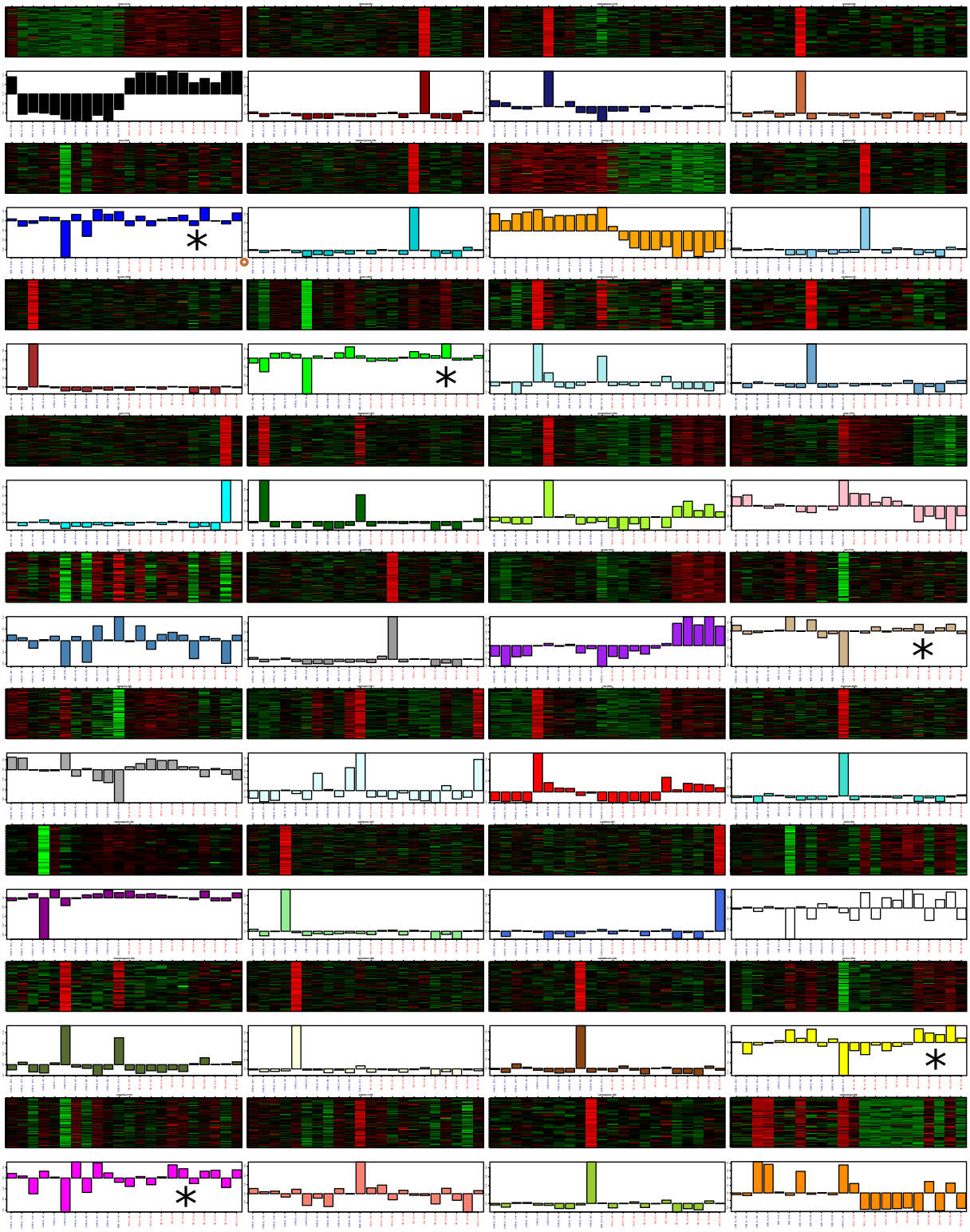
Supplementary Figure 32: Comparisons of sumstat human-HIV-1 category enrichment p-values (on $\log$ scale) for selection scores (x-axis) and fraction of the gene conserved, as measured by ConScores (y-axis). The blue lines indicate a significance above Holm-Bonferroni FWER of 1%.



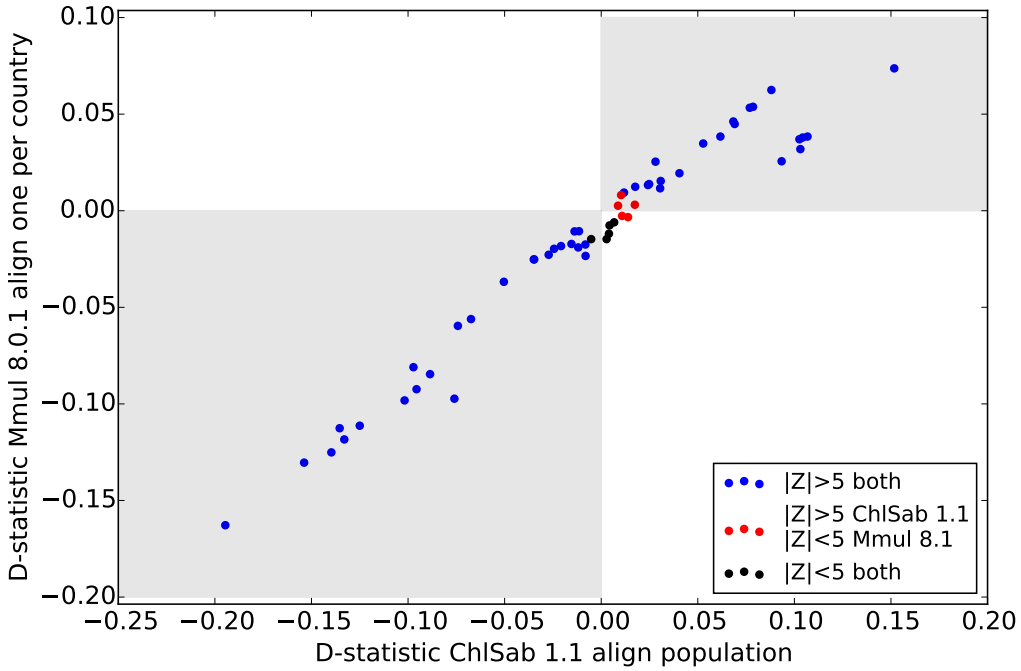
Supplementary Figure 33: Boxplot of selection scores in genes from expression study in Jacquelin et al.[40]. The bars correspond to genes with significant expression changes pre- and post-SIV-infection in (from left to right) only vervet, only macaque, both species, and all genes on the expression array. Significance was calculated using the one-sided Mann-Whitney-U-test. The categories contain 1593, 405, and 254 genes, respectively.



Supplementary Figure 34: Cluster dendrogram representing relative distance of the 36 modules identified by WGCNA. 1-correlation among module eigengenes (1-cor(ME)) was used as a measure of similarity. Modules with 1-cor(ME)<0.1 were merged.



Supplementary Figure 35: Co-expression modules of genes with significant expression difference pre- and post-SIV-infection. For each module, the top panel represents a heatmap of normalized relative expression levels of all genes (rows) across samples (columns). The bottom panel shows the first principal component of gene expression as a bar-plot. The column labels are as given in Fig. 3c. Number of transcripts in each module is in the title in parentheses next to module name. Modules with significantly enrichment for high selection scores are marked with an asterisk (FWER$<$0.05) and also represented in Fig. 4a.



Supplementary Figure 36: Comparison of D-statistic results between the original callset aligned against the vervet genome (Chl.Sab.1.1; x-axis) and realignment of a subset of the samples (one individual per sample, c.f., Supplementary Table 4) aligned against the macaque genome (Mmul 8.0.1). Country labels for each test are given in Supplementary Data S2.

Supplementary Table 1: Hard filters applied to SNP calls.

name	JEXL expression	description
QualityByDepth	$\mathrm{QD} < 2.0$	Low quality by depth.
MissingExcess	$\mathrm{NCC} > 81$	More than 25% of the alleles in genotypes missing.
LowCov	Custom mask	Raw coverage over all individuals $< 1/2$ median.
HaplotypeScore	$\mathrm{HaplotypeScore} > 13.0$	Low consistency of the site with two segregating haplotypes.
LowMQ	$\mathrm{MQ} < 50.0$	Low mapping quality.
StrandBias	$\mathrm{FS} > 60.0$	Highly significant strand bias.
Dels_gt0	$\mathrm{Dels} > 0.0$	Locus spanning deletion as reported by GATK.
ExCov	Custom mask	Raw coverage over all individuals > 1.5 median
ReadPos	$\mathrm{ReadPosRankSum} < -8.0$	Bias in read position.
5bpIndel	Custom mask	Reported indels $+ -$ 5bp.
Low Qual	$\mathrm{QUAL} < 50.0$	Low SNP quality.
MQ0_7pct	$(\mathrm{MQ0}/(1.0 * \mathrm{DP})) > 0.07$	Large amount of mapping quality 0 reads.

Supplementary Table 2: P-values for anova (likelihood-ratio test) of adding geography or taxon to a linear model explaining the first six principal components.

PC	adding geography	adding taxon
1	$7.31 * 10^{-21}$	$3.65 * 10^{-103}$
2	$1.13 * 10^{-6}$	10^{-300}
3	$1.39 * 10^{-3}$	10^{-300}
4	$7.32 * 10^{-7}$	2.1210^{-284}
5	$1.90 * 10^{-4}$	4.5310^{-123}
6	$1.09 * 10^{-13}$	$6.51 * 10^{-53}$

Supplementary Table 3: Average number of IBD segments in the autosome that are 10kb or longer. Values normalised by number of pairwise comparisons.

	aethiops	*cynosuros*	*hilgerti*	*pygerythrus*	sabaeus Africa	*St. Kitts Nevis*	*Barbados*	*tantalus*
aethiops	9185	37	115	22	6	5	2	48
cynosuros		3153	424	1117	3	3	1	44
hilgerti			2051	290	16	13	13	125
pygerythrus				6752	0	0	0	29
sabaeus Africa					15742	17471	13457	107
St. Kitts Nevis						66819	26241	132
Barbados							43631	91
tantalus								2631

Supplementary Table 4: Samples used for alignment against rhesus macaque Mmul 8.0.1.

ucla_id	country	coverage
VGHA1001	Ghana	4.3
VEA1001	Ethiopia	4.5
VWP00201	Saint Kitts	4.4
VKA3	Kenya	4.6
VGA00010	Gambia	4.1
VZA1001	Zambia	4.6
VSAE3002	South Africa	4.5
VBOA1005	Botswana	4.9
C2166	Barbados	4.5
AGM126	Central African Republic	4.1

Supplementary references

[80] Zeng, K., Fu, Y.-X., Shi, S. & Wu, C.-I. Statistical tests for detecting positive selection by utilizing high-frequency variants. *Genetics* **174**, 1431–9 (2006).

[81] Spence, J. P., Kamm, J. A. & Song, Y. S. The site frequency spectrum for general coalescents. *Genetics* **202**, 1549–1561 (2016).

[82] Shapiro, J. A. *et al.* Adaptive genic evolution in the drosophila genomes. *Proc Natl Acad Sci U S A* **104**, 2271–6 (2007).