

**Supplemental Information for:
Lindblad-Toh et al: “A high-resolution map of evolutionary constraint in the
human genome based on 29 eutherian mammals.”**

Contents

Supplementary Notes.....	2
S1 Samples and sequencing.....	2
S2 Assembly.....	2
S3 Error estimation and correction.....	3
S4 Alignments details.....	4
S5 Estimating constraint.....	5
S6 Detection of constrained elements.....	6
S7 Constraint vs. polymorphism.....	7
S8 Protein coding genes.....	8
S9 Readthrough and SCEs.....	9
S10 RNA structures.....	9
S11 Patterns of promoter constraint.....	14
S12 Regulatory motif discovery.....	15
S13 Chromatin state information.....	16
S14 Overall accounting of constraint elements.....	17
S15 Disease associated variants.....	17
S16 Codon specific positive selection.....	17
S17 Exaptation of ancestral repeat elements.....	21
S18 Human and primate accelerated regions.....	22
Supplementary Data Sets.....	26
Supplementary Reference.....	33
Supplementary Tables 1-15.....	37
Supplementary Figures 1-24.....	90

Supplementary Notes

This supplementary information follows the structure and layout of the main paper. Here we describe the methodology in detail and include additional analysis not covered in depth in the main text.

Supplementary Section S1: Samples and sequencing

Supplementary Methods 1.1 - Sequencing strategy:

Our goal was to identify constrained elements in the human genome based on evolutionary constraint across eutherian mammals. To obtain the maximum power at a reasonable cost we developed a strategy of sequencing genomes to 2x coverage instead of the standard 8x¹. At 2x, theoretically ~85% of the sequence should be recovered, although in smaller pieces.

Supplementary Methods 1.2 - Species selection and preparation:

Species were selected to generate the maximum novel branch length, while spanning all the four major clades. When appropriate the relevance of the species as a model organism was taken into account. For non-endangered species, 8 females were selected and heterozygosity tested across 200 loci. The individual with the lowest level of polymorphism was selected for sequencing. All individuals sequenced are female. Most species had heterozygosity rates of 1/500-1/1000 bp, while the elephant, guinea pig and rabbit had extremely low heterozygosity. DNA was prepared from blood or available tissues using standard protocols for high molecular weight DNA. To determine the amount of sequence required for 2x coverage of each genome, ~50,000 Fosmid pairs were generated and an estimation of the euchromatic genome size was performed by placing these Fosmid pairs on the canine genome and estimating the relative genome size.

Supplementary Methods 1.3 - Sequencing coverage:

Eight of the nine high coverage draft or finished genomes were previously described (Table S1). Guinea pig was sequenced with a standard mix of paired end reads (4kb and 10kb plasmids, 40-kb Fosmid and BAC ends²) totaling 6.8x coverage of the genome and assembled using ARACHNE³.

For each of the organisms sequenced to low coverage, ~2x coverage of sequence typically was generated from 4kb paired end reads (~1.8x) and 40kb Fosmids (~0.2x) using ABI3730 sequencers. Species and sample source information is described in Table S1.

Supplementary Section S2: Assembly

Supplementary Methods 2.1 - Genome assembly strategy:

All 2x genomes were assembled using an assisted assembly method⁴ where the *de novo* assembly is improved by placing all reads on related genomes (in this case on dog, CanFam2.0 and human, hg18) and leveraging the placement information to support and extend information in the *de novo* assembly.

Importantly, no novel information was introduced in the assemblies based on the alignment to the related genomes. Instead, read-read alignments confirmed by read placement on related genomes

were used to extend existing contigs; and confirmed single links were allowed to join different scaffolds, hence improving contiguity. Moreover, read pairs aligning onto the reference uniquely and consistently were used to detect and eventually fix misassemblies introduced in an earlier stage due to the presence of weak joins. All assembly statistics are described in Table S1 and assemblies are available at <http://www.broadinstitute.org/ftp/pub/assemblies/mammals/>.

Supplementary Text 2.2 – 2x sequencing versus next generation sequencing:

The 2x genomes utilized here were produced over a time span of several years, but to achieve consistency in the data set we utilized the same sequencing and assembly strategy throughout. The current availability of next generation sequencing technologies will reduce the cost of genome sequencing by an order of magnitude and allow sequencing to high coverage of many mammals. For comparative genome analysis understanding the quality of the assemblies, alignments and error modes in the data are important and so while future mammalian analysis will be less expensive and time consuming, they should still be performed in a consistent manner.

Supplementary Section S3: Error estimation and error correction

A detailed analysis of sequencing errors in the 2x genomes, including both miscalled bases and erroneous indels, has been published as a satellite paper⁵, with key results summarized briefly below.

Supplementary Text 3.1 - Quality score distribution for assembly:

Even with only 2x average genomic coverage, these assemblies predominantly consist of high-quality bases. Roughly 82% of bases, across all of the assemblies, have quality scores >45, and only ~4% have scores <20.

Supplementary Text 3.2 - Comparison with ENCODE assemblies to estimate error rates:

The mismatch rates for aligned 2x and ENCODE bases ranged from 2.6 (elephant) to 25.1 (hedgehog) mismatches per kb, while the indel rates showed somewhat less variation. These rates generally decreased with increasing quality scores, as expected, but the limiting rates—at the highest quality scores—differed considerably among species.

At the highest quality scores, mismatch rates were highest for hedgehog, for which separate species were sequenced in the ENCODE and 2X project, and microbat, known to show elevated levels of intraspecific genetic variation⁵. The smallest mismatch rates occurred with the African savannah elephant, which has been reported to have low average genetic diversity⁵.

We thus used observed difference rates at the highest quality bases (quality score >45) as rough estimates of polymorphism rates, and subtracted these estimates from the overall observed rates to obtain approximate polymorphism-corrected error rates. After this adjustment, the estimated base-call error rates were much more concordant across species, at 0.72–3.43 per kb.

The 2x species with slightly higher coverage, such as megabat (2.6x) and rock hyrax (2.2x), have the lowest residual error rates, as expected.

Supplementary Text 3.3 - Generation of error-corrected assemblies and alignments:

Given the localized nature of sequencing errors, with a large majority of errors coming from a small minority of bases, we developed automatic methods to generate a version of the alignments that mitigate the effects of errors for downstream analyses. These are available from <http://compgen.bscb.cornell.edu/projects/32way-masked/>

Supplementary Section S4: Alignment details**Supplementary Text S4.1 - Alignment choice in the detection of constraint:**

Although we have used MultiZ alignments for the main bulk of this work, we also evaluated the use of Enredo and Pecan⁶ to build an alternative set of alignments, for the specific purpose of evaluating segmental duplications, since MultiZ was not designed to handle them. In short, Enredo is a graph-based method that defines sets of co-linear genomic sequences and Pecan, a consistency-based multiple aligner, aligns them. On the one hand MultiZ provides a higher coverage of the genome. On the other hand Pecan works in a global fashion and, together with Enredo, can provide alignments on segmental duplications.

We compared the sets of phastCons constrained elements we obtain from both sets of alignments. We restricted the comparison to the regions covered by the Pecan alignments. Globally, we found approximately the same amount of genome under constraint: 4.34% on the MultiZ alignments vs. 4.01% on the Pecan alignments. When comparing the constrained regions at the base pair level, we found that 83.48% of the bases called by phastCons on the Pecan alignments are also detected as constrained when looking at the MultiZ alignments. The majority of the conflicting elements are shorter than 20 nucleotides. For instance, 93% of the MultiZ-phastCons elements of at least 30bp overlap with a Pecan-phastCons element (Fig S2).

We next focused on the human segmental duplications as defined by Enredo where Pecan and MultiZ alignments are likely to be more different. In these regions, Pecan assigned all the human copies of the genomic segmental duplication to the same alignment. We observed a drop in the amount of nucleotides under constraint in the MultiZ alignments (from 4.34% to 2.71%) while we see an increase in constraint in the Pecan alignments (from 4.01% to 4.94%). Again, both sets of elements tend to disagree more on smaller elements, but this time a third of the elements longer than 50 nucleotides are found in the Pecan-phastCons set only (Fig S2). PhastCons can over-predict constrained regions on Pecan alignments where long insertions have happened. We remove these from the final set by ensuring that all constrained elements have no more than 10% gaps in the human sequence and at least 80% informative nucleotides (non gaps or Ns) in at least one other sequence.

In segmental duplications, we note a reduction in the amount of detectable constrained elements. In some examples, we see how Pecan-phastCons elements can detect coding exons missed in the MultiZ-phastCons set, although we note that phastCons is not designed to study segmental duplications. Importantly, the constrained elements detected on both sets of alignments are comparable for the non-duplicated portion of the genome. For the remaining analysis in this paper we use the MultiZ alignments.

Supplementary Section S5: Estimating constraint

Supplementary Methods S5.1 - Neutral tree estimation:

For each chromosome we extracted 4-fold third codon positions from all Ensembl genes annotated as protein that had a single homolog in both mouse and dog. We used PAML (version 4, June 2007)⁷ with default parameters to estimate branch lengths fixing the tree topology. The estimated nucleotide substitution rate matrix and tree branch lengths were used as the neutral model for constraint estimation.

Supplementary methods S5.2 - Supplementary methods - Estimates of selection - Siphy- ω , Siphy- π :

We used Siphy⁸ for constraint estimation, at a resolution of 12bp and 50bp as previously described. Constraint was estimated in the full 29way-alignment, an alignment containing only a subset of species Human-Mouse-Rat-Dog (HMRD), and an alignment containing high coverage genomes for the ENCODE regions. Constraint was estimated using both a rate based method (Siphy- ω) and a biased substitution based method (Siphy- π).

Supplementary Text S5.3 - Challenges of varying branch length:

As a measure of constraint for the Siphy- π method, the log-odds scores were used. For a given strength of constraint, elements with log-odds scores are proportional to alignment depth, and as a result, elements with higher branch length will receive higher score. Since Ancestral Repeats tend to have lower branch length, this may lead to an over-estimation of the excess under constraint. To conservatively estimate constraint purely based on differences in substitution patterns, we have employed a branch length correction. Briefly, genomic windows were divided to bins based on their average branch length. In each bin, the mean (μ) and standard deviation (s) of the log-odds scores were computed for the Ancestral Repeats background sets. All scores in a bin were standardized by subtracting the mean and then dividing by the standard deviation to produce a Z-score. All Z-scores from all bins were then collected to produce one genomic distribution and one Ancestral Repeat distribution. Constraint was called based on these distributions as described in the previous section. The correction we have employed eliminates the bias towards windows with higher branch length.

Supplementary Methods S5.4 - Comparison with Encode:

In its most recent data freeze (January 2009) the ENCODE dataset includes a multiple alignment of a superset of the species used in this paper. This alignment consist of assemblies sequenced at “comparative analysis grade” offer us the possibility to benchmark the impact of low coverage data on our predictions.

The usage of 2x mammals presents potential challenges and biases in comparing the background set to the entire genome, due to problematic alignability issues and varying branch lengths. To assess the impact of the incomplete data resulting from low coverage

assemblies we compared predictions in the ENCODE regions using both alignments: The ENCODE alignment composed of high coverage assemblies and the genome wide alignment composed of low coverage assemblies restricted to the ENCODE regions. The ENCODE alignments are missing 6 species from our sets, and these species were excluded for the purpose of this comparison, resulting in the comparison of 23 species.

Supplementary Text 5.5 - Notes for Table S2 - Constraint Estimation and Detection statistics:

We used three different methods for detection and estimation of constrained elements (blue, red, green). The first (SiPhy- ω) is looking for a reduction in the neutral divergence rate estimate ω , and setting a cutoff on the value of ω . The second (Siphy- ω lods) is looking at ω , but setting a cutoff on the log odds probability that ω is less than 1. The third (Siphy- π) is looking at the log odds probability that the stationary distribution of the mutation matrix is different from uniform. We applied all three methods to 12-mers across the 29-way alignment using ancestral repeats as the background (top row), to 12-mers across the 29-way alignments using ancestral repeats that are conserved over species covering the same branch length as the elements, thus correcting for the varying branch lengths across the genome, but possibly being too stringent because some of the conservation of ancestral repeats is due to selection (second row), to 12-mers across the four mammals human, mouse, rat, dog (third row), and to 50-mers across the four mammals (last row).

Supplementary Section S6: Detection of constrained elements

Supplementary Text 6.1 - Detecting constraint elements:

We used a window-based approach to call constrained elements.

First, we divided the genome into (overlapping) kmers. Each kmer was scored using SiPhy to get $S(i)$ the score of k-mer i . Since neutral evolutionary rate differ between different chromosomes, subsequent analysis was done separately for each chromosome. For each chromosome we constructed two histograms, a genomic histogram comprising all scores of k-mers for this chromosome, and a background histogram, comprising scores of all k-mers overlapping Ancestral Repeats for this chromosome. The two histograms were then used to compute an empirical FDR cutoff ω_c at level 10%, defined as the maximum value such that the ratio of areas to the right of this value is no more than 10%. Windows with $S(i) > \omega_c$ were declared as significant k-mers. Finally, we clustered overlapping significant k-mers to yield larger elements. The above procedure was repeated for two different window sizes (12bp and 50bp), different conservation scores (SiPhy- ω , SiPhy- π) and different alignments (29way, HMRD and ENCODE comparative grade).

The different alignments gave vastly different significance cutoffs. For example, when calling elements at a 10% FDR level and using the full 29way alignments at a 12bp resolution, the average constraint cutoff across all chromosomes was $\omega \sim 0.29$ - hence a roughly 3.4-fold reduction in average substitution rate was sufficient to declare an individual 12-mer as constrained. Using the HMRD alignment, a much stricter cutoff of essentially $\omega=0$ (no observed substitutions in the alignment) was required to declare a 12-mer as constrained at the same 10% FDR level.

Supplementary Text S6.2 - Comparison of SiPhy- ω , SiPhy- π and PhastCons elements:

We set out to examine the robustness of the set of elements called based on the method used. We have used both SiPhy versions, as well as elements called using phastCons. Overall, there was a strong agreement between the different methods. PhastCons elements overlap 93% of SiPhy- ω (5% FDR) bases, whereas phastCons overlap only 75% of SiPhy- ω (10% FDR) bases. This suggests that the core data set is similar but that 10% FDR of SiPhy- ω elements allows the inclusion of additional bases not found in the phastCons elements. This is supported by the fact that when these three data sets are compared, the SiPhy- ω 5% FDR contains 1,674 elements unique to this data set, whereas phastCons has 285,039 unique elements and SiPhy- ω 5% FDR has 1,449,816. While some of these elements may be false positive due to the higher FDR, this dataset should contain many novel elements.

Supplementary Text S6.3: Correlation between 29 mammals and previously defined constrained elements:

For each Megabase in the genome we computed HMRD 50mer based element and 29-way eutherian 12mer based (SiPhy- ω 10% FDR) element density, and found a very strong correlation between the two, suggesting that newly discovered elements have a similar genome-wide distribution as previously discovered elements (Figure S6).

In addition, we compared the elements detected for 29 mammals with both SiPhy ω and π at 5% and 10% FDR and PhastCons elements and compared the overlap for both elements and bases with the HMRD 50 bp set, the Siepel five vertebrate data set and the union of these elements. While the union of the Siepel and HMRD elements cover a larger fraction on the genome, the overlap with the 29 mammals elements is only marginally bigger (see Supplemental Table S3 and Figure S4).

Supplementary Text S6.4: Identification of newly-detected elements: Within intronic and intergenic elements, the majority of constraint sequence was not previously detected as constraint with HMRD. Taking the lack of granularity of the 50mer HMRD analysis into account, we identify 55.3% of elements as potentially part of elements previously detected by HMRD (overlapping or falling within 50 bp of a previously detected constraint element). An additional 19.4% of elements fall 50 to 500 bp from previously identified elements suggesting they could be part of the same regulatory unit. The remaining 25.3% of elements (~916,000 elements), fall >500 bp from previously identified constraint elements, suggesting that they may be novel entities.

To examine the clustering of elements on a larger scale we examined the element numbers in 10kb windows of the genome. Under Poisson distribution the 95 percentile for the number of elements contained in a 10K window is 18. Since there are 300K such windows in the genome, we expect about 15K windows having 18 or more elements. We observe ~5x this number which indicates that elements do indeed tend to cluster

Supplementary Section S7: Constraint vs. polymorphism**Supplementary Methods S7.1 - SNP analysis methods:**

For all analysis we have used an unbiased SNP set from Keinan et al.⁹. To allow the maximal divergence between human populations we have used the Yoruban (YRI) SNP set. Each position in the 29way-alignment was collapsed to the IUPAC code symbol closest to the Siphy- π vector for this position. A total of 92,906 SNPs for which Siphy- π indicated a two-fold degenerate site were used for comparison with the human SNPs.

Supplementary Methods S7.2 - SNP density vs. constraint:

We computed the average SNP density in both masked genomic regions, and in constrained elements and performed a correlation.

Supplementary Methods S7.3 - Shifted allelic spectrum for two-fold mutating sites:

We counted the joint frequency of all di-nucleotide patterns in the mammalian phylogeny for the ancestral and derived allele in the human YRI populations.

The value $C(ij,kl)$ represent the counts in which the two mammalian alleles are i and j , the human ancestral allele is k and the derived allele is l .

We computed an enrichment matrix $E(ij,kl)$ given by:

$$E(ij,kl) = C(ij,kl) / C(ij)*C(kl)$$

where $C(ij)$ is the marginal frequency of the di-nucleotide pattern (ij) in the mammalian phylogeny, and $C(kl)$ is the marginal frequency in the human ancestral and derived alleles. This enrichment matrix measures the tendency of the non-human allele and the derived allele to coincide, with $E(ij,kl) = 1$ representing no enrichment beyond background.

Supplementary Section S8: Protein-coding genes

Supplementary Methods S8.1 - Exon prediction:

To predict novel conserved exons, we applied an enhanced version of CONGO, an algorithm we previously developed for the same purpose in *Drosophila* genomes¹⁰. Briefly, CONGO incorporates discriminative metrics of protein-coding evolutionary signatures -- including reading frame conservation and codon substitution frequencies¹¹ -- within the framework of a semi-Markov conditional random field (SMCRF), a type of probabilistic graphical model that can combine such metrics in order to produce a segmentation of the genome into predicted exons and non-coding regions. The enhancements to the previous version are mainly adaptations to the mammalian exon prediction task, including a semi-Markov feature to model the short length distribution of mammalian exons, a synteny feature helpful for recognizing duplicated and pseudogenic regions, and an alternative training objective function similar to Conrad's¹², which improves the accuracy of the algorithm given the unbalanced prediction task (only ~1.5% of the human genome being protein-coding).

We applied CONGO to the MULTIZ alignments of 29 eutherian genomes, which we also postprocessed to mask likely sequencing errors⁵. We trained CONGO's SMCRF using the GENCODE

annotations of the ENCODE ‘random’ regions¹³, approximately 0.5% of the human genome. We then applied the trained model to decode each human chromosome assembly. From the ~175,000 resulting exon predictions, we subtracted those overlapping any annotated coding exon from the major human gene catalogs (including RefSeq, Ensembl, GENCODE and UCSC Genes downloaded on March 27, 2010 – see supplementary data), leaving the 3,788 predictions reported in the main text.

Supplementary Methods S8.2 – Transcript models

We next used Scripture¹⁴ to reconstruct transcript models based on high-throughput transcriptome sequencing in 16 human tissues by Illumina, Inc (BodyMap2). We used the resulting transcript models to assess expression evidence and tissue specificity for the predicted exons, and also to ‘link’ them with known gene structures annotated by GENCODE, as described in the main text. We also collected several other lines of evidence not used by CONGO to provide initial support for the exon predictions, including existing EST/cDNA expression evidence, similarity to Pfam protein domains identified using HMMER¹⁵, and coding gene-associated chromatin states¹⁶⁻¹⁷.

Supplementary Section S9: Readthrough and synonymous constraint elements

We searched for potential examples of stop codon readthrough in mammals by searching for continued protein-coding evolutionary signatures in the regions immediately downstream of the stop codon in known human coding transcripts. This follows our previous approach in 12 *Drosophila* genomes¹⁰. In addition to known selenoprotein-encoding genes, we found four novel candidates (*OPRK1*, *OPRL1*, *BRI3BP*, and *SACM1L*) with no apparent alternative explanations other than translational readthrough (such as splicing or RNA editing). Further analysis of these mammalian examples, and many others in animal species, are reported in a separate manuscript (Jungreis et al., “Evidence of widespread stop codon readthrough in *Drosophila* and other metazoa”, in revision).

To identify protein-coding regions under selection for additional, overlapping functions, we developed a method based on phylogenetic codon models to measure synonymous substitution rates in short windows and report the statistical significance of their reduction. We applied this method to all human CCDS ORFs and identified more than 10,000 “Synonymous Constraint Elements” (SCEs) with resolution for nine-codon windows, covering about 2% of all synonymous sites (FDR < 0.01). A typical example shows a 77% reduced synonymous rate compared to genome-wide averages across placental mammals. The SCEs show strong positional enrichments for exon and ORF boundaries, suggesting widespread constraints on splicing and translational regulatory elements embedded within mammalian ORFs. Many SCEs can also be associated with other overlapping functions such as miRNA targeting, dual-coding regions, A-to-I editing, and RNA secondary structures, but further study will be needed in order to associate most with biological functions. Full details of this work are reported in a companion papers¹⁸⁻¹⁹.

Supplementary Section S10: RNA structures

Supplementary Text 10.1 - EvoFold screen

The EvoFold RNA structure screen was based on a 41 species subset of the genome-wide 44-way multiZ alignment, which includes additional vertebrate genomes and is available from the UCSC

Genome Browser. We used a 31-way subset for the structure prediction and profile-model training, consisting of 28 of the 29 eutherian mammals, together with opossum, chicken, and tetraodon as outgroups. An additional 10 species that were not used for structure inference, consisting of primarily non-mammalian vertebrates and a single eutherian mammal, were used as an independent test set for structure validation.

The screen was restricted to a set of conserved alignment segments based on the PhastCons predicted elements²⁰, which span 5.56% of the genome. Since EvoFold is sensitive to misaligned sequences, we applied a conservative sequence filter to the extracted alignment segments, which discards sequences with a surprising number of mismatches given the branch-lengths of the relating phylogenetic tree (see Alignment Filtering below). EvoFold (v.2.0)²¹ was then applied to these filtered alignments in both their forward and reverse directions. Low-confidence predictions that are short (< 6 base-pairs); harbor excessive amount of bulges; based on shallow or low quality alignments; or overlap repeats or pseudogenes were eliminated from the prediction set. See supplement of²² for details on the screen.

The UCSC Genes set (as of May 25, 2009) was used to define genomic regions. Each prediction was assigned to the genomic region it had the greatest overlap with. Protein-coding regions were excluded from the study to focus it on non-coding regions.

Supplementary Text 10.2 - EvoFam clustering pipeline

We clustered the EvoFold predictions into candidate families using a novel approach (EvoFam), as described below. A probabilistic model based on a profile stochastic context free grammar (pSCFG), a.k.a. a covariance model, was built from each EvoFold prediction, using the Infernal RNA tools v 1.0 (cmbuild utility)²³. An all-against-all similarity (homology) comparison between structural RNA predictions was performed, based on a probabilistic similarity measure between pairs of pSCFG models. This new similarity measure is based on a form of Kullback-Leibler divergence, modified to correct for varying false positive rates due to varying model size and complexity (see²²).

A similarity graph was defined with vertices corresponding to the pSCFG models of RNA structures and with edges connecting pairs of models with a dissimilarity below a threshold selected to control false discovery rate (FDR). Families were defined as highly connected subgraphs, where a highly connected subgraph (HCS) is defined as a subgraph S of n vertices with edge connectivity $k(S) > n/2$, with edge connectivity $k(S)$ defined as the minimum number of edges whose removal disconnects S . These families were computed using the iterated HCS algorithm of²⁴.

Additional paralogous matches to the UTR EvoFold predictions were detected by searching the conserved UTR regions of the human genome with the corresponding pSCFG (using cmsearch with global search option). The paralogous hits were filtered by requiring E-value < 0.1 (relative to a 1 Megabase database) and strong double substitution evidence (p-value < 0.2; Monte Carlo test applied to all species excluding human). Repeat regions and known pseudogene matches were removed. For the analysis limited to UTR regions, this set of putative paralogs was then combined with the original EvoFold set and analyzed by the subsequent family identification stages.

Supplementary Text 10.3 - Filtered candidate families

After initial definition of the candidate families through cluster analysis, we further evaluated the statistical significance and biological evidence for the candidate sets. The disjunction of a series of enrichment tests was used to produce the final high confidence filtered sets:

(i) We evaluated the statistical significance of the compensatory substitutions supporting each member of a family using a Monte Carlo test on the 31-way alignment and, importantly, on the independent set of ten held-out species not used for structure inference. The test, called EvoP, measures how surprising it is to achieve the observed number of double substitutions, given the total number of substitutions and a random substitution process on the branches of the tree relating the aligned species (see S10.5 – EvoP double substitution significance test). Considering each member as an independent test for the overall significance of a family, the p-values of all family members were combined multiplicatively using the Fisher method and used as an overall measure of evidence as well as for ranking.

(ii) For predictions within known protein-coding genes (i.e. UTR and intronic genomic regions), gene ontology (GO) enrichment statistics were computed for each cluster with three or more members, using the topGO library²⁵. We additionally required that an enriched GO term had evidential support in two or more family members to prevent a single unusual gene flagging the entire family. The GO analysis was conducted against a background set of the original EvoFold structure predictions, and so estimated the additional enrichment of families beyond the possible enrichments or biases of the original EvoFold set. Families were filtered based on the most significant p-value in each ontology.

(iii) The degree of enrichment of family members for a particular genomic region (5'UTR, 3'UTR, intron, intergenic) was computed by chi-squared statistic relative to the background proportions of the entire EvoFold prediction set.

(iv) We calculated the mean structure length in terms of pairing bases for each family: longer structures have a lower prior probability and thus higher confidence.

(v) Enrichment relative to an immunity-related gene set consisting of the human homologues of mouse macrophage-related genes as defined in²⁶ was estimated by Fisher's exact test.

We defined a final set of high-confidence predictions as the disjunction of the families deemed biologically significant via any of these significance estimates: those for which any of the following measures had p-value smaller than a defined threshold (0.05 for double substitution p-values; < 0.005 for region enrichment; < 0.01 for maximal GO enrichment p-values); or mean base-pair length > 11. Combining these statistical measures of confidence, the original full set of candidate families was filtered to a smaller set of high-confidence families.

Supplementary Text 10.4 - Alignment filtering

Since both the EvoFold RNA structure screen and the EvoFam family identification pipeline are sensitive to the assumption that the alignments are correct, we filter sequences that are likely to be misaligned from the alignments before using them. We filter an alignment by first identifying every outer branch on which there is significantly more substitutions than expected given its length and

the substitution rate in the rest of the tree. If such a branch exists, with at least two substitutions and more than 5-fold rate change relative to expected, we then mark the entire sequence from the species to which they lead as unobserved in the given alignment.

Both the number of substitutions on the outer branches and the substitution rate in the remaining tree are estimated using PAML⁷.

Supplementary Text 10.5 - EvoP double substitution significance test

We have developed a new, easily interpretable p-value for evaluating and ranking the predicted structures based purely on substitution evidence. See supplement of ^{f22} for details.

For a given predicted structure and an alignment that corresponds to the stem bases of this predicted structure, the p-value measures how probable it would be to see at least as many double substitutions in the aligned sequences if they do not encode a structural RNA (assuming the total number of substitutions in the alignment is fixed). We applied a previously-defined methodology²⁷, but whereas most current methods for counting substitutions are based on pairwise sequence comparisons, the counts for this p-value are performed on the underlying phylogenetic tree to take into account the treelike nature of evolution. Similarly, we only count a pair of substitutions as a double substitution if the two substitutions occur not only in the same base-pair, but also on the same branch in the tree, since the two substitutions are only likely to be correlated if they happen in close proximity time-wise. All of the substitution counts are based on the most probable ancestral tree which we infer using PAML⁷.

When calculating the p-values we assume that if the aligned sequences do not encode a structural RNA each of the substitutions will have happened with equal probability along the sequence, with probabilities proportional to the branch lengths along the phylogenetic tree and with equal probabilities to all different types of substitutions. These are estimated using a Monte Carlo approach.

When a structure prediction is based on only a subset *S* of the species in the alignment, we use a very similar approach to assess to what extent the additional species support the original prediction. The only difference to the p-value defined above is that we include only the branches that do not connect the species in *S* in our analyses.

Supplementary Text 10.6 - Summary and enrichment statistics

To evaluate the statistical evidence for the entire predictions sets, we calculated the following summary and enrichment statistics (Figure 9a), which in all cases are based on the subset of novel predictions as defined in ²²: (1) EvoFold score: mean EvoFold log-odds score of novel structures in the sets. (2) RNAz overlap enrichment: Enrichment in overlap of novel EvoFold predictions with RNAz predictions compared to a random null set. (3) DNase hypersensitivity overlap: % overlap of novel intergenic predictions with DNase I hypersensitivity sites compared to random conserved regions (see ²² for details). (4) Avg. correlation of tissue-specific expression within families: The average Pearson correlation coefficient of expression of novel structures within families, across tissues based on a multi-tissue RNA-seq dataset (Illumina Body Map 2 data) (P-values based on

shuffled null) (see ²² for details). (5) Intergenic expression enrichment: The expression of novel conserved intergenic elements relative to randomly selected intergenic regions.

Using the Illumina BodyMap 2 ribo-depleted, non-polyA selected, total RNA dataset [Illumina], we compared the expression evidence overlying the novel intergenic structures compared with a shuffled set of random structure positions chosen from the conserved intergenic regions of the genome. For the expression analysis, regions overlapping RNA repeat elements and known pseudogenes were removed from both unshuffled and random background sets, as reads are often wrongly mapped in these regions giving false expression signals. Similarly, we removed regions showing mitochondrial chromosome homology as they have increased probability of representing either ribosomal RNA or tRNA pseudogenes. 1000 shuffles were used to estimate p-values by permutation test (see ²² for details).

Compared to the random background, the fold increase in expression (mean reads per base) of intergenic elements that show any expression for the EvoFold, GW unfiltered, and GW filtered input sets was 1.20 X ($P < 1E-3$), 1.46 X ($P < 1E-3$), and 2.33 X ($P < 1E-3$), respectively. Similarly, the fold increase of the average expression levels for the entire predictions sets (including elements with no expression evidence) were 1.02 X ($P < 0.21$), 1.20 X ($P < 6E-3$), and 1.70 X ($P < 1E-3$), respectively, for the EvoFold, GW unfiltered, and GW filtered input sets.

Supplementary Text 10.7 - Thermodynamic analysis of EvoFold predictions using RNAz

We used RNAz 2.0²⁸⁻²⁹ to analyze the initial EvoFold predictions as well as the structures in the clustered families before and after filtering. RNAz was run on the same alignments as EvoFold. However, since the classification algorithm of RNAz is not trained for short structures, we added flanking regions to all EvoFold predictions shorter than 120 nt to obtain a minimum length of 120.

As positive control, we used a set of 356 known structural RNAs. As negative controls we (i) chose random locations within the PhastCons conserved regions that were used as input for the EvoFold analysis, (ii) we shuffled the alignment³⁰ (iii) we simulated random alignments preserving the dinucleotide content in the alignment³¹.

We focused on two metrics calculated by RNAz: the z-score and the classification score. The z-score is a normalized value measuring the thermodynamic stability of an RNA structure. It is the number of standard deviations a given RNA structure is more/less stable than structures for random sequences of the same length and dinucleotide content. RNAz calculates the average z-score of all sequences in an alignment. By convention, negative z-scores denote more stable structures. RNAz combines the stability z-score with an evolutionary score measuring the structural conservation and uses a support vector machine to classify an alignment as "Functional RNA" or "other".

The results of the z-score analysis show that EvoFold predictions are significantly more stable than the random controls and that the clustering and the filtering of clusters enrich for even more stable RNA secondary structures (Supplementary Fig. S10a). The same trend can be observed for the RNAz classification (Supplementary Fig. S10b). While only 1-2% of the random controls are predicted as functional RNA, 11% of the EvoFold predictions are supported by RNAz. For the

clustered families this fraction increases to 15% and 24%, for the unfiltered and filtered sets, respectively.²²

Supplementary Text 10.8 - Data availability

The complete set of structure predictions from the EvoFold screen as well as the candidate family predictions sets can be downloaded in bulk or browsed through a UCSC Genome Mirror from the following web-site: <http://moma.ki.au.dk/prj/mammals/>. In addition, individual families are listed and annotated in²² and its supplement.

Supplementary Section S11: Patterns of promoter constraint

Supplementary Text 11.1 - Identification of peaks in constraint in the core promoters of transcripts

A peak in constraint was defined by an increase in Siphy- π value over the surrounding region. Raw π -scores were smoothed with a windowed-mean algorithm in 8 bp windows. Local maxima were identified in the smoothed signal, and a peak recorded if the conservation score at the maximum was $\geq 1.5X$ the average score at the base of the peak. To reduce spurious signals produced by low conservation, peaks with smoothed π -score less than 4 (roughly the genome-wide average) were discarded.

Supplementary Text 11.2 - GO enrichment

Enrichments for Gene Ontology (GO) terms were run using the topGO package, version 1.16.2 from BioConductor²⁵. Gene Ontology annotations for human transcripts were collected from version 58 of the Ensembl database using the Ensembl Perl API: each term within the Biological Process ontology annotated by Ensembl as being associated with a given transcript was collected, along with all of its ancestor terms, with the associated transcript and gene name. In order to avoid a lack of power due to small numbers of annotated genes, all GO annotations were used regardless of evidence code. Each independent list of genes was tested for GO term enrichment using the entire set of human transcripts as the universe, using the "classic" algorithm and Fisher's exact test statistic as implemented in topGO. P-values for enrichment were corrected for multiple testing using the Bonferroni method.

Supplementary Text 11.3 - Comparison to CpG and TATA-box promoter classifications

To evaluate the relationship between the observed promoter constraint patterns with CpG and non-CpG promoters, we defined CpG promoters as those with an annotated CpG island within 200 bp of the TSS (CpG island coordinates were downloaded from UCSC) and compared these to our groups of promoter classification. CpG promoters were abundant in all classes; 66% of genes with both 'high' and 'intermittent' constraint were CpG promoters (5,083 and 9,342 genes respectively), while 41% (1,188 genes) of 'low' constraint genes were CpG promoters. In total, 63% (15,613) of the genes in our set are associated with a

CpG island. For reference, genes with TATA boxes of the “right” distance and orientation in the promoter are represented at 2-3% in each of the 3 categories.

Supplementary Section S12: Regulatory motif discovery

Supplemental Methods S12.1 – Regulatory motif instance identification

We use a method similar to Kheradpour³² with extensions for using position frequency matrices (PFMs). We build a catalog of 688 motifs (for 345 factors) from TRANSFAC (version 11.3)³³, Jaspar (version 2008)³⁴, and large scale systematic motifs generated by Protein Binding Microarrays³⁵⁻³⁷.

For each motif, 100 shuffled motifs are generated by randomly shuffling the columns of each PFM. Because the way the information content is ordered may affect the background level of conservation (e.g. a group of specified bases surrounded by unspecified bases may be more likely to be conserved by chance than the converse due to conservation typically being “blocky”), we create three bins of information content and shuffle only within each bin. Each of the 100 shuffled motifs is then matched to the human genome and only those that have +/- 20% the number of matches are considered. The remaining motifs are then clustered at a 0.8 correlation and up to 10 control motifs are chosen in random order, allowing only one motif per cluster. Together, these procedures attempt to choose control motifs that have the same base-composition and similar higher-order compositions (through frequency matching), while being diverse.

A branch length score (BLS) is computed for each motif match in human by computing the branch length of the smallest subtree that contains human and the informant species that contain an aligned motif match. We then produce a mapping between BLS and confidence ($1 - \text{false discovery rate}$) for each branch length score using $1 - r_c / r$ where r and r_c are the fraction of motif instances and control instances, respectively, that reach at least the specified BLS. Wilson score interval with $z=1$ is applied to both r (correcting downward) and r_c (correcting upward) in order to produce a conservative estimate of confidence in situations with few instances.

We also permit motif movement by repeating the procedure for each of the 32 windows $w=0, 5, 10, 20, \dots, 100, 120, \dots, 500$ allowing both the motif and the control motifs to move w bases in the informant genomes relative to the position aligned to human. Consequently, for each confidence cutoff from 0.1, 0.2, ..., 0.9 the BLS and w combination that results in the highest sensitivity is chosen.

The computation of the confidence mapping benefits from having homogeneity in the regions scanned. It is important to exclude regions that may have other sources of evolutionary constraint (e.g. coding sequence) and regions that are difficult to align (e.g. repeats). Consequently, we exclude all simple repeats and repeat masked regions (downloaded from the UCSC genome browser on April 13, 2006) and use the Ensembl gene annotations (downloaded from UCSC on September 18, 2007) to exclude coding regions, 3' untranslated regions and exons from non-coding genes.

Confidence prediction is done on only autosomes, and then instances are identified on the chromosome X using the mapping produced on the autosomes but with a tree produced on

chromosome X. This is important to correct for the higher background level of conservation of chromosome X. Chromosome Y is ignored. Scaling analysis ignores instances on chromosome X.

When matching the PFMs we use a uniform background and a pseudo count of 0.001. The analysis in this paper is done at a match p-value of 4^{-8} as determined by TFM-Pvalue³⁸. Of the 688 motifs that we started with, 630 (representing 335 factors) were able to be matched at this stringency and have at least one shuffle motif.

Supplementary Text S12.2 - Statistics on motif instances (Table S6). Matching motifs against the genome generally matches a very large proportion of the 1,558,114,353 bases we scanned. Only considering conserved motif instances, even at a low stringency, dramatically reduces the number of bases matching a motif. Statistics also reported for best motif for each factor (i.e. the one with the highest number of instances).

Supplementary Text S12.3 - Datasets and motifs used in motif analysis. Datasets were identified from the literature and peaks identified in the study were used after mapping to the appropriate assembly (if necessary). For factors that also had a dataset available in mouse, we also show the number of peaks found in human that were conserved in mouse. When multiple motifs were available for a factor, we chose the one with the highest enrichment in the human dataset (ignoring conservation).

Supplementary Section S13: Chromatin state information

Supplementary Methods S13.1 – Overlap with 51 chromatin states in CD4 cells. For the chromatin state analysis we first masked the portion of the genome corresponding to coding exons, RNAs, regions +/-2kb of a TSS, and pseudogenes (see Figure S18). We then evaluated the overlap of unmasked Siphy- ω 10% FDR elements with a set of 51 chromatin states previously defined in CD4T cells based on the maximum posterior state assignments¹⁶. For each state we computed the ratio of the number of bases of unmasked conserved elements to the total number of bases in the state, which fell in the region that was unmasked. The reported fold enrichment was this ratio relative to the fraction of bases in unmasked regions, which overlapped unmasked bases of Siphy- ω elements.

Supplementary Methods S13.2 – Overlap with 15 chromatin states across nine cell types. We also evaluated overlap of Siphy- ω elements with a set of 15 chromatin states defined across 9 cell types (Figure S19). As the same location may be assigned to different chromatin states in different cell types we ordered the chromatin states in a greedy manner, that is selecting the chromatin state with the maximum enrichment for conserved bases after excluding masked regions and chromatin states previously selected. A location was associated with a state if it was assigned to that state in at least one cell type and was not assigned to any previously selected state in any cell type. Using that same ordering of states we also evaluated the cumulative enrichment for unmasked Siphy- ω bases for locations assigned to that state or a higher ranking state in at least N cell types for N=1,...,9 (Figure S20).

Supplementary Section S14: Overall accounting of constrained elements

Supplementary Text S14.1 Supplementary Methods

Bases in constrained elements were assigned in a hierarchical manner based on the Gencode Version 2 (Levels 1-3) annotations from hg18, such that if a constrained base overlapped multiple categories, it was assigned only to the first annotation on the list it overlapped. The annotations used and their order were: coding exons, 5' UTR, 3' UTR, core promoter (200 bp of transcription start), extended promoter (2 kb of transcription start), RNA genes, pseudogenes, introns and intergenic sequence. RNA genes consisted of both linc RNAs¹⁴ and structural RNAs as defined in <http://moma.ki.au.dk/~jsp/data/rna/>. Among the constrained elements not accounted from this set, chromatin states were considered to account for unconstrained elements if they overlapped one of the 20 chromatin states with the greatest conservation in the CD4T model or one of the eight candidate promoter, enhancer, and insulator states in at least one cell type of the nine cell type model. Motif instances at the 40% confidence were used to account for constrained elements after the chromatin states.

Supplementary Section S15: Disease-associated variants

Supplementary Methods 15.1

To investigate the overlap between genomic variants associated with clinical phenotypes and mammalian conservation, we used results from the genome-wide association study (GWAS) database curated by NHGRI and accessed on May 30, 2011³⁹. The overlap between GWAS hits and SiPhy- ω 10% FDR elements was calculated for the autosomes and X chromosome, excluding regions that were masked during prediction of the SiPhy- ω elements, exons, and regions within 2kb of the starts of transcripts as defined by GENCODE version 3c⁴⁰. We also calculated the enrichment relative to conservation of HapMap phase 2 and 3 SNPs that were assayed in the Utah residents with ancestry from northern and western Europe (the CEPH population, abbreviated CEU), from release 28⁴¹. Significance of the fold enrichment was expressed as a binomial p-value.

To find examples of conservation and regulatory motif instances overlapping GWAS-linked SNPs, we used LD from HapMap release 27⁴¹. We first used r^2 measures from the CEU population to list blocks of SNPs in perfect linkage ($r^2 = 1.0$) with each of the GWAS results, and scanned them against SiPhy- ω elements and the library of motif instances described in Supplementary Section S12.

Supplementary Section S16: Codon-specific positive selection

Supplementary Methods S16.1 - Phylogenetic analysis with the sitewise likelihood ratio

Sitewise dN/dS values were estimated by running the Sitewise Likelihood Ratio (SLR) software⁴² on the trees and alignments from each of the parallel datasets produced by the pipeline described below. Characterizing the sitewise behavior of protein evolution in mammals was of primary interest, as the increased phylogenetic depth afforded by the added low-coverage genomes was

expected to give enough evolutionary branch length to make a sitewise analysis sensitive and specific enough to discover new regions subject to pervasive positive selection pressures within the mammalian clade. The SLR method was employed for this study, as it has been shown in simulation experiments to have desirable performance when compared to the sitewise Bayes Empirical Bayes decoding implemented in PAML, especially with regards to consistent performance between different genes⁴².

A brief overview of SLR's sitewise inference is as follows: first, all gene-wide parameters are estimated for a probabilistic codon model of evolution (including branch lengths, equilibrium codon frequencies, the transition/transversion ratio, and dN/dS_{gene} corresponding to the overall dN/dS across the gene) using information from all sites in the gene. The dN/dS_i at each site is then individually and independently re-estimated while holding all other parameters fixed, and a per-site likelihood ratio test is performed, calculating the independently maximized likelihood scores for each site given the null model (neutrality, dN/dS=1) and the alternative model (dN/dS ≠ 1) and comparing the resulting test statistic (twice the log-likelihood difference between the null and alternative model) to a χ^2 distribution. The Benjamini-Hochberg procedure is employed to correct for multiple tests. In addition to the likelihood ratio test results, SLR provides the estimated dN/dS_i at each site and the lower and upper limits of an approximate 95% confidence interval on dN/dS formed by thresholding the likelihood curve. See⁴² for more information and results on the size and power of the SLR test. We refer to these sitewise patterns of positive selection as localized positive selection.

Supplementary Methods S16.2 - Ensembl Compara gene trees pipeline

Orthologous gene trees were generated from gene family trees in the Ensembl Compara v57 database. The Ensembl Compara gene trees pipeline⁴³ generates gene families and alignments with the following steps: (1) run all-against-all BLASTP sequence similarity searches using all annotated Ensembl proteins, (2) generate sequence family clusters from the resulting BLAST similarity matrix using the hcluster_sg algorithm, (3) align sequences using Mcoffee's meta-alignment method, and (4) calculate gene trees using TreeBeST, which uses a given known species tree and multiple evolutionary models (including protein, codon, and DNA models) to guide the creation of gene family trees with resolved duplication events. TreeBeST combines different evolutionary models into one final gene tree topology that is then fixed. Finally, branch lengths are optimized with maximum likelihood using the HKY model. The ability of TreeBeST to combine information from multiple evolutionary models is especially important in obtaining robust tree topologies from the often partial gene predictions resulting from the low-coverage genomes in this dataset.

The Ensembl pipeline uses a novel gene-building pipeline for its annotation of low-coverage shotgun genome assemblies, as the nature of low-coverage genomes causes problems with the standard pipeline⁴⁴. Ensembl's low-coverage gene building pipeline is based on whole-genome alignments of each low-coverage genome to an annotated reference genome (human in the case of the present mammalian genomes), built using BLASTz and axtTools. Scaffolds are arranged into "gene-scaffolds" based on the reference genome annotation, and frame-shifting indels in the low-coverage sequence (which are the result of sequence / assembly error in the vast majority of cases) are corrected by inserting 1- or 2-bp "frame-shifting" introns into the low-coverage gene model.

When the whole-genome alignment implies that the low-coverage sequence is missing an entire internal exon, a run of 'X's is inserted into the gene model in order to produce the correct translation length.

Supplementary Methods S16.3 - Isolating nearly-orthologous mammalian sub-trees

The full Compara gene family trees represent the deep evolutionary history of a gene, including duplication and speciation events going as far back as the protein-protein BLAST searches will reach. As a result, some gene trees comprise several hundred sequences with many nearly complete sets of orthologous sub-trees related by ancient duplication events. Although this property of the trees may be useful for other analysis, we wished to avoid measuring functional divergence following gene duplication in this study. Furthermore, we also wanted to restrict our analysis to genes that are well-represented in mammals to avoid including artifactual lineage-specific genes resulting from over-annotation or assembly error.

To mitigate both of the above issues for this analysis, a simple algorithm was used to filter out trees without sufficient taxonomic coverage and to split trees with ancient duplication events into roughly orthologous sub-trees. For each node starting from the root, the tree was split into two sub-trees and both sub-trees were included in the analysis if they both satisfied the following criteria: (1) contained genes from at least two out of four mammalian families (Primates, Glires, Laurasiatheria, Afrotheria) and (2) contained at least one gene from one of the outgroup species (chicken, tetraodon, opossum). If the entire tree did not satisfy the two conditions, it was dropped from the analysis. The application of this tree-splitting algorithm to the Ensembl Compara v57 database, which contains 19,298 full gene trees covering 21,317 human genes, yielded the set of 15,451 trees (or sub-trees) containing 17,709 human genes used in this analysis.

Supplementary Methods S16.4 - Sequence quality and alignment filtering

The accuracy of homology assignments is a major concern in all evolutionary studies, and the issue becomes critically important in site-wise analyses: because there is no averaging of evolutionary measures over the length of a gene, a single mis-aligned codon could directly lead to a false positive in the sitewise analysis of selective pressures, artificially raising the proportion of observed positively-selected codons. In the present study, there was also concern that the higher sequencing error rates in low-coverage genomes could lead to elevated estimates of evolutionary rates and dN/dS ratios.

In order to minimize these potential sources of error, we masked out regions of the multiple alignments based on sequence quality scores and an alignment quality metric. For sequence quality, we masked out codons (replacing the codon's nucleotides with 'N' characters) in low-coverage genomes where any of the three nucleotides in the codon had a PHRED or PHRED-equivalent score equal to or below 20. Quality scores for the source assemblies in Ensembl v57 were retrieved from Ensembl, UCSC or the source sequencing centers. For alignment quality, we used the approach employed by Pollard et al⁴⁵ of filtering out alignment columns which did not contain at least 3 Primate sequences, 2 Glires sequences, and 1 outgroup sequence. Outgroup species were defined as any species not within the Primate, Glires, or Laurasiatheria clades.

Supplementary Methods S16.5 - Clade-specific datasets

Although the SLR method does not include models comparable to PAML's branch-site models⁴⁶ for lineage-specific sitewise analysis, the mammalian evolutionary tree contains a number of similarly-sized and closely-related clades which are amenable to independent evolutionary analysis. In order to investigate independent sitewise selective pressures in mammalian sub-clades, we ran the sitewise pipeline on alignments restricted to each of the Primates, Glires, and Laurasiatheria clades. Although the Afrotheria clade is also situated nearby in the mammalian tree, it was excluded from this analysis due to the low number of available genomes (3). Each sub-clade pipeline run used alignments generated from the full mammalian gene alignments by removing any sequences from species outside the clade, and subsequently removing any columns containing only gaps. The sequence and alignment quality filtering was applied to the full mammalian alignments, and any masked sequences or alignment columns were carried over to the sub-clade alignments.

Supplementary Methods S16.5 - Gene Ontology term enrichment analysis

The topGO package for R/Bioconductor²⁵ was used to evaluate genes containing elevated dN/dS or positively-selected codons for enrichment in Gene Ontology (GO) terms. GO term annotations from v60 of the Ensembl human annotation set were used. Two sets of genes were defined: genes under weak constraint having $dN/dS[gene] > 0.4$, and genes undergoing significant positive selection with at least one confidently positively selected codon (at $FDR < 0.05$). The topGO package was used to perform a series of Fisher's exact tests for enrichment, using the set of all genes analyzed in this study as the gene 'universe'. A Bonferroni correction was applied to enrichment p-values to correct for multiple testing, and only terms enriched at $p < 0.05$ after correction were retained.

Supplementary Methods S16.6 - Pfam protein domain analysis

Alignment sites were annotated with Pfam domains using v57 of the Ensembl human annotation set, yielding 2.2 million sitewise values annotated with 3,116 Pfam IDs. Each Pfam domain was summarized according to its number of annotated sites, number of genes covered, mean estimated dN/dS value, and fraction of positively-selected sites. Because the overall proportion of positive selection in mammals is quite low, domains covering either fewer than 500 sites or fewer than 5 genes were dropped from the analysis to avoid spurious results from a small number of misalignments or stochastic effects.

For the present analysis, domains were separately sorted by the mean dN/dS value and the fraction of positive selection. In order to separate domains best characterized by weak selective constraint and intermittent positive selection from those best characterized by strong constraint and localized selection, the top 10 domains for each sorting criterion were compared; results are shown in Table 5.1.3a. Domains placed within the top 10 under both sorting criteria are shaded in grey, and potentially interesting domains under the fraction of positive selection criterion are highlighted in bold.

Supplementary Methods S16.7 - Data availability

See <http://www.ebi.ac.uk/~greg/mammals/> for data resulting from the sitewise analysis. Summary statistics and plots for each gene can be accessed online, and all data generated (including trees, alignments, sitewise results, and tables of domains & GO terms) are available either to download in bulk or as UCSC browser tracks.

Supplemental Text S16.8 (Table S9) - Summary of sitewise selection pressures in mammals, primates, glires, and laurasiatheria

Notes: A number of summary calculations were performed on each of the four sitewise datasets. Columns labeled 'dN/dS X' contain the fraction of sites where the maximum-likelihood estimate of dN/dS is above or below the given value. 'Domain instances' shows the total number of Pfam domain instances gathered from Ensembl v57, and 'Domain types' shows the number of unique Pfam IDs gathered. 'Positive domain instances' and 'Positive domain types' show the same calculations for the subset of domain instances containing significantly positively-selected sites. 'Positive genes' and 'Positive sites' similarly show the number of genes and sites containing significant evidence for positive selection (after multiple-testing correction).

Supplemental Text S16.10 (Table S12) - GO enrichments

Notes: All terms were first sorted by 'pval.fis.bonf' (the Bonferroni-corrected p-value for enrichment) and a threshold of $p < 0.05$ was applied. Terms were subsequently sorted by 'pval.elim' for display purposes. Terms mentioned in the text are in bold.

Tables containing complete results for all GO enrichments are available from the above webpage.

Supplemental Text S16.11 (Table S13) - Domain analysis

Notes: Rows were sorted either by mean dN/dS or by fraction of positive sites, and the top 10 domains for each set were retained. Those domains showing up in the top 10 on both lists were grayed out. Domains mentioned in the text are in bold.

Supplementary Section S17: Exaptation of ancestral repeat elements

Supplementary Methods S17.1 - A detailed analysis of mobile elements exapted to act as putative regulatory elements in the human genome is described in a companion paper⁴⁷.

To create a subset of putative regulatory regions we began with the set of conserved elements defined by phastCons²⁰. We then removed all conserved elements that overlap protein-coding exons, untranslated regions, or exons from non-coding RNA genes. The resulting set consists of ~2.6 million conserved non-exonic elements (CNEEs). These CNEEs are under selection, but do not appear in mature transcripts, suggesting that they are likely to be functional at the DNA level and act to regulate the expression of nearby genes⁴⁸.

To understand which of these putative regulatory elements in the human genome are the results of mobile element insertions we examined the overlap of our CNEEs with mobile element annotations generated by RepeatMasker (www.repeatmasker.org). To be conservative, we did not keep all CNEEs that overlapped a mobile element insertion, but only those, which had a majority of the

bases annotated as originating in a SINE, LINE, LTR, or DNA transposon insertion. This resulted in a set of 284,857 conserved non-exonic elements, totaling nearly 7Mb of sequence, which have been exapted from mobile element insertions.

To date exaptation events we used the genome-wide multi-species alignment. For each exaptation we began with the most divergent group of species and calculated if half, or more, of the bases in the CNEE were aligning to any species in the group. This was iterated with progressively closer species until more than half of the CNEE bases were present in the most recent common ancestor of human and the group of species being used. The exaptation event was placed on the branch of the human lineage above the ancestor that appears to have contained at least half of the CNEE bases.

Supplementary Section S18: Human and Primate accelerated regions

Supplementary Text S18.1 - Human acceleration in Primate conserved elements

A second set of HARs was identified using candidate elements that were only required to show conservation in primates (though many are also more deeply conserved). These 920,486 primate-conserved sequences are similar in their genomic distribution to the 1.3 million mammalian-conserved regions used in the primary HAR analysis. They are 92.4% non-coding, with 11.1% overlapping experimentally identified enhancers. Interestingly, a much larger proportion of these primate-conserved regions are accelerated in human (1930 or 0.2% of primate-conserved elements vs. 0.04% of mammalian-conserved elements), suggesting that lineage-specific evolution is more common in less deeply conserved functional elements. Like the primary set of HARs, these 1,930 HARs are slightly depleted in coding sequences and occur as often as expected in enhancers compared to the set of primate-conserved regions from which they were identified.

Supplementary Text S18.2 - Evidence of biased gene conversion in accelerated regions

GC-biased gene conversion (gBGC) is a non-adaptive, recombination-associated process that increases the rate of fixation of AT to GC polymorphisms (i.e., weak-to-strong changes). Since gBGC can mimic selection and accelerate the overall rate of substitutions, we checked for biases in the pattern of human substitutions in HARs in order to assess evidence that these regions could have been shaped by gBGC.

On average, HAR sequence GC-content is nearly identical in human (39.0%) and the inferred human-chimp ancestor (38.1%). Of 550 HARs with sufficient outgroup alignment data to count substitutions of each type, 284 (51.6%) have more weak-to-strong than strong-to-weak substitutions on the human lineage. These HARs, and not any others, show an increase in GC-content on the human lineage. On average, this change represents a 1% increase in GC-content. A few HARs show more extreme patterns of weak-to-strong bias and larger increases in GC-content. Human substitutions in seventy-three HARs (13.3%) are all weak-to-strong (compared to 6.4% all strong-to-weak) and result in greater than 5% increases in GC-content on the human lineage. Since HAR ancestral GC-content is 38.1%, there is greater mutational opportunity for weak-to-strong compared to strong-to-weak changes on the human lineage. Hence, these findings suggest that while gBGC may have shaped the evolution of some HARs, most of the signal of human acceleration in HARs cannot be explained by gBGC or other GC-biased mutation or fixation processes.

We did not conduct a parallel analysis of substitutions in PARs, because gBGC is not a likely explanation for the substitution patterns in these elements. In particular, our filtering and statistical testing methods ensure that most PARs have higher than expected rates of substitutions in multiple primate lineages. Since recombination rates are highly variable at the PAR length-scale (hundreds of base pairs) between even closely related primates and recombination hotspots are not expected to occur at the same place on multiple branches on the primate phylogeny, it is highly unlikely that gBGC would have shaped substitutions across the primates.

Supplementary Text S18.3 - Gene Ontology (GO) enrichment analysis of accelerated regions versus positively and negatively selected codons

To explore potential functional effects of sequence changes in fast-evolving regions of the mammalian genome, we conducted statistical enrichment analyses of Gene Ontology (GO) terms associated with HARs and PARs (which are mostly non-coding) and compared these results with those for positively and negatively selected codons. Interestingly, there are many more enriched terms in common between the HARs or PARs and negatively selected codons than between HARs or PARs and positively selected codons. One notable exception is a set of GO terms associated with extra-cellular signaling (e.g., “cell surface”, “extra cellular space”, “receptor activity”, “signal transducer activity”) that are enriched in loci that are fast-evolving at both the coding and non-coding levels. Also, terms related to immunity feature in both sets, albeit somewhat distinct aspects of the immune system. In contrast, developmental pathways and processes (e.g., “axon guidance”, “ureteric bud development”) are much more commonly enriched in the HARs and PARs, suggesting a prominent role for developmental gene expression divergence in primate and human adaptations.

Although these results could be affected by myriad factors including gene density and genome-scale mutation rate heterogeneity, they seem to suggest that regulatory changes tend to occur more frequently in genes under strong purifying constraint, and rarely in genes experiencing adaptive or diversifying selection. Thus, distinct processes and pathways appear to have been under selection at the regulatory versus protein level in mammals.

Supplementary Text S18.4 - Gene Ontology (GO) enrichment analysis of HARs versus PARs

Contrasting the HAR and PAR GO enrichment results may indicate specific functions underlying human-specific biology. In fact, HARs and PARs are associated with largely different sets of enriched GO terms, although they share a small number of terms including “homophilic cell adhesion” and “astacin activity”. Interestingly, HARs are preferentially enriched for GO terms related to transcriptional regulation, the cell cycle, and MHC receptor activity compared to PARs. Furthermore, while HARs are enriched for some neuronal terms, they do not show strong enrichment for annotations related to axonogenesis, which are highly enriched among PARs. These results are not sensitive to the use of a 1Kb versus 100Kb window for mapping GO terms onto HARs and PARs. These findings indicate that although some biological systems are universally fast-evolving, adaptation appears to occur through somewhat distinct mechanisms in different clades.

Supplementary Methods S18.5 - HAR and PAR detection

Genomic regions with accelerated substitution rates in the human and primate lineages were identified by first defining candidate elements conserved across all mammals excluding the lineage of interest (human or primates). Conserved elements were identified using the phastCons program from the PHAST package (<http://compugen.bscb.cornell.edu/phast>) with expected length = 45, target coverage = 0.3, and expected conservation level = 0.3. Conserved elements were filtered to remove potential alignment and assembly errors using annotations from the UCSC genome browser database. Our strict inclusion criteria were: level 1 or level 2 non-gap synteny between human and all of macaque, mouse, and dog (netSynteny); no pseudogenes (luNega and pseudoYale); no segmental duplications (genomicSupDups); no repeat elements (rmsk); and no human paralogs (selfChain). After filtering, there were 1,322,576 mammalian conserved elements for human lineage-specific tests and 1,255,037 for primate clade-specific tests. Note that the exclusion of either human or all primates when running phastCons means that these sets are not identical to each other or to the set of phastCons elements used elsewhere in this study.

These sets of filtered conserved elements were scored for accelerated substitution rates in the subtree of interest (human or primates) compared to the rest of the tree using the likelihood ratio test (LRT) method implemented in the phyloP program from the PHAST package^{45,49}. Acceleration p-values were adjusted for multiple comparisons using the FDR-controlling method of Benjamini and Hochberg⁵⁰.

Intersections with genome annotations and experimentally identified enhancers were performed using the featureBits program from the kent libraries and custom scripts. Genomic coordinates of experimentally identified enhancers obtained from the supplemental materials of Visel *et al.*⁵¹ and Heintzman *et al.*⁵² were mapped to the human genome assembly (hg18) using the liftOver program from the kent libraries and alignment chains from the UCSC genome browser database.

Transcription factor binding sites were predicted on both strands of the human and chimpanzee sequences of each HAR using position specific weight matrices derived from motifs of the 11 JASPAR transcription factor families with pseudo-counts³⁴. For each family, statistically significant binding potential was determined using the balanced cutoff method of Rahmann *et al.*⁵³. The total number of predicted binding sites was determined separately for human and chimp in each HAR for each JASPAR family. For each HAR and each family, these counts were transformed into a divergence score by taking the absolute difference in the number of sites between human and chimp and dividing by the length of the HAR. The minimum, maximum, and mean values of the divergence score across all HARs were computed for each family and for all families together.

Supplementary Methods S18.6 – Checking HARs for evidence of gBGC

To check if the accelerated substitution rates in HARs might be driven by GC-biased gene conversion, we calculated GC content and proportion of changes from weak-to-weak, weak-to-strong, strong-to-strong, and strong-to-weak along the human branch for each HAR. To do so, we first reconstructed the ancestral sequence, using chimp, gorilla, and organutan as outgroups. The procedure for reconstructing the ancestral sequence was majority rule parsimony, but also minimized the number of changes from weak to strong and strong to weak in cases where there was no majority. Specifically, the procedure was as follows: If any one outgroup nucleotide

matched the ingroup, the ancestral nucleotide was assigned that nucleotide. However, if none of the outgroup nucleotides matched the ingroup, the nucleotide with the maximum number of outgroup species in agreement was used (e.g., ingroup = A, outgroups = {CCG}; ancestor is assigned C). If there was no majority, the nucleotide was assigned based on whether the ingroup nucleotide was weak or strong (e.g., ingroup = A, outgroups = {C,G,T}; ancestor is assigned T).

Supplementary Methods S18.7 - Gene Ontology enrichment analyses

Functional enrichment/depletion analysis of fast-evolving regions was conducted by mapping Gene Ontology (GO) terms onto our data sets as follows. In all cases, we considered both the full set of GO terms and the GOSlim subset from the GOA project (<ftp://ftp.ebi.ac.uk/pub/databases/GO/goa/goslim/>).

We first mapped GO terms to UCSC known genes using the go.goaPart table of the UCSC genome browser. Then, we associated a term with any of the 1.32 million candidate phastCons elements located within 1Kb (or 100Kb) of a transcribed known gene annotated with that term. In this case, the “universe” for enrichment/depletion analysis is the set of candidate phastCons elements and the “foreground” set is the subset of these elements that is significantly accelerated in humans (HARs). By conducting enrichment/depletion tests at the level of phastCons elements, we automatically account for the many-to-many mappings of these elements to genes and GO terms. A parallel mapping pipeline was applied to the PARs. Both pipelines were implemented using Perl and MySQL (DBI) with local copies of the UCSC databases.

Each GO term was tested for enrichment and depletion in each of the “universe” versus “foreground” comparisons described above using two one-sided Fisher exact tests implemented in custom R (<http://r-project.org>) scripts. Unadjusted and Bonferroni-corrected p-values were used to evaluate statistical significance.

Supplemental Text S18.7 – Explanation of Table S15-predicted transcription factor binding sites per 100bp across all HARs for each of the 11 JASPAR families. Min=minimum number of hits for a given family in a single HAR. Max=maximum number of hits for a given family in a single HAR. Mean=average number of hits for a given family per HAR.

Supplementary Data Sets

Data access: A complete set of data files can be downloaded from or viewed in:

1. The Broad website (<https://www.broadinstitute.org/scientific-community/science/projects/mammals-models/29-mammals-project-supplementary-info>)
2. UCSC (<http://genomewiki.cse.ucsc.edu/index.php/29mammals>)
3. IGV (<http://www.broadinstitute.org/igv/projects/29mammals>).

Constrained Elements (SiPhy -omega, & pi)

Summary: Lists of constrained elements. For each 12-mer in the human genome a measure of constraint was scored using SiPhy (see reference below), both as a rate-based score (omega), and a measure that includes biased substitution patterns (pi). Those falling in annotated Ancestral Repeats were used as a background. An empirical cutoff score was set corresponding to 10% FDR, and all 12-mers above this score were considered significant. Overlapping significant 12-mers were clustered to yield larger elements. The HMRD SiPhy -omega 50-bp set used for comparison is also made available.

Files:

29way_omega_lods_elements_12mers.chr_specific.fdr_0.1_with_scores.txt.gz
 29way_pi_lods_elements_12mers.chr_specific.fdr_0.1_with_scores.txt.gz
 HMRD_omega_lods_elements_50mers.chr_specific.fdr_0.1_with_scores.txt.gz

Format: 'Chromosome Start End Lods-score Branch-length'

Format note: coordinates are 0-based, inclusive (meaning the End position is considered part of the element), and on hg18

Contact: Or Zuk <orzuk@broad.mit.edu>, Manuel Garber <mgrarber@broadinstitute.org>

Reference: Garber, M. *et al.* Identifying novel constrained elements by exploiting biased substitution patterns. *Bioinformatics* **25**, i54-62, doi:btp190 [pii] 10.1093/bioinformatics/btp190 (2009).

Heights (omega,pi)

Summary: Base-level measure of constraint scored using SiPhy (see reference above), both as a rate-based score (omega) and a measure that includes biased substitution patterns (pi).

Files:

omega.12mers.wig.gz
 Format: 'position log_odds_score'
 pi.ewig.gz
 Format: 'postion %A %C %G %G log_odds_score'

Format note: coordinates on hg18

Contact: Manuel Garber <mgarber@broadinstitute.org>

Protein-coding exons

Summary: A list of identified previous annotation and (Reference annotation) novel conserved exons (Congo). Exons were identified using a version of CONGO (previously developed for the *Drosophila* genomes, see reference below) enhanced to handle mammalian exon prediction. The enhancements include a semi-Markov feature to model the short length distribution of mammalian exons, a synteny feature for recognizing duplicated regions, and an alternative training function to improve accuracy when performing an unbalanced prediction task (only ~1.5% of the human genome is protein-coding).

Files: ReferenceAnnotation_hg18.20101019.new.syn.gtf
CONGO_hg18.20101019.new.syn.gtf

Format: GTF

Format note: coordinates on hg18

Contact: Mike Lin <mikelin@mit.edu>

Reference: Lin, M. F. *et al.* Revisiting the protein-coding gene catalog of *Drosophila melanogaster* using 12 fly genomes. *Genome Res* **17**, 1823-1836, doi:gr.6679507 [pii] 10.1101/gr.6679507 (2007).

Synonymous Constraint Elements

Summary: Identified coding regions with a very low synonymous substitution rate – indicating additional sequence constraints beyond the amino acid level. The Synonymous Constraint Elements (SCEs) are defined at three different resolutions (9-, 15-, and 30-codon). There is also a bedGraph track for the local estimate of the synonymous substitution rate (λ_s). Also available at: <http://compbio.mit.edu/SCE/>

File: SynonymousConstraintElements.tar.gz

Files in archive:

SCE9.hg18.bed.gz

SCE15.hg18.bed.gz SCE30.hg18.bed.gz

λ_s _ORF.hg18.bedGraph.gz

File formats: BED, bedGraph

Format note: coordinates are on hg18

Contact: Mike Lin <mikelin@mit.edu>

RNA structures

Summary: The list of candidate predictions for structural RNA families. EvoFold structural predictions were based on a 31-way subset of the genome-wide 44-way multiZ alignment

(consisting of 28 of the 29 eutherian mammals, together with opossum, chicken, and tetraodon as outgroups) and clustered into candidate families using the novel EvoFam algorithm. This data, as well as the complete set of structure predictions from the EvoFold screen can be downloaded in bulk or browsed through a UCSC Genome Mirror from the following web- site:

<http://moma.ki.au.dk/prj/mammals/>.

In addition, individual families are listed and annotated in the following reference and its supplement.

File: StructuralRNAfamilies.tar.gz

Files in archive:

Genome_wide_prediction_set
Genome_wide_with_paralogs_prediction_set
UTR_with_paralogs_prediction_set
data_format.txt

Format: described in the file: data_format.txt

Format note: coordinates are hg18

Contacts: Brian Parker bparker@binf.ku.dk, Jakob Skou Pedersen jakob.skou@ki.au.dk

Reference: Parker, B. J. *et al.* New families of human regulatory RNA structures identified by comparative analysis of vertebrate genomes. *Genome Research* (2011).

Constraint Structure in Promoters

Summary: A list of local maxima identified from the smoothed pi-scores in the core promoters of genes.

File: peaks.positions.gz

Format: 'Chromosome Start End Score'

Format note: Coordinates are 1-based and exclusive (meaning the End base is not included in the peak position). All positions are at a single base, and are on hg18.

Contact: Evan Mauceli <evan@broadinstitute.org>

Motif instances

Summary: A list of instances of identified regulatory motifs. A motif catalog was built from TRANSFAC, Jaspar, and Protein Binding Microarrays using a method similar to that described in the reference below, with extensions for position frequency matrices. Motif instances were identified genome-wide using a FDR of 60%.

File: instances-thresh8-0.4.txt.gz

Format: Motif-name Chromosome Start End Strand

Format note: coordinates are 1-based, inclusive (meaning the End position is considered part of the element), and on hg18

Contact: Pouya Kheradpour pouyak@mit.edu

Reference: Kheradpour, P., Stark, A., Roy, S. & Kellis, M. Reliable prediction of regulator targets using 12 Drosophila genomes. *Genome Res* **17**, 1919-1931, doi:gr.7090407 [pii] 10.1101/gr.7090407 (2007).

Chromatin Mark Data

Summary: ENCODE segmentation of hg18 into chromatin states for each of nine human cell types. States were learned using a Hidden Markov Model that computationally integrated CHIP-seq data into fifteen states associated with different types of functionality. This data is available from UCSC at:

<http://genome-preview.ucsc.edu/cgi-bin/hgTrackUi?hgsid=2563118&c=chrX&g=wgEncodeBroadHmm>

A similar segmentation based on the CD4T cell line is also provided in the file:

Files:

[map_allstates.bed.txt.gz](#) – segmentation based on CD4T cell line
[chromatinMarks.tar.gz](#)

Files in archive:

[wgEncodeBroadHmmGm12878HMM.bed](#) – modelling in GM12878 cells
[wgEncodeBroadHmmH1hescHMM.bed](#) – modelling in H1-hESC cells
[wgEncodeBroadHmmHmecHMM.bed](#) – modelling in HMEC cells
[wgEncodeBroadHmmHsmmHMM.bed](#) – modelling in HSMM cells
[wgEncodeBroadHmmHuvecHMM.bed](#) – modelling in HUVEC cells
[wgEncodeBroadHmmHepg2HMM.bed](#) – modelling in HepG2 cells
[wgEncodeBroadHmmNhekHMM.bed](#) – modelling in NHEK cells
[wgEncodeBroadHmmK562HMM.bed](#) – modelling in K562 cells
[wgEncodeBroadHmmNhlhHMM.bed](#) – modelling in NHLF cells

File format: BED

Contact: Jason Ernst jernst@mit.edu

References:

Ernst J and Kellis M. [Discovery and characterization of chromatin states for systematic annotation of the human genome](#). *Nature Biotechnology* 2010 Jul 25;28:817-825.

J. Ernst, P. Kheradpour, T.S. Mikkelsen, N. Shores, L.D. Ward, C.B. Epstein, X. Zhang, L. Wang, R. Issner, M. Coyne, M. Ku, T. Durham, M. Kellis, B.E. Bernstein
 Mapping and analysis of chromatin state dynamics in nine human cell types.
Nature 473: 43-49, 2011.

Accounting for conserved elements

Summary: A list of each conserved element (omega), the chromatin state it resides in, and, if applicable, any genic annotation it overlaps, as well as any overlapping motif instances.

File: elementPartition.txt.gz

File Format:

Column 1: 'ELEMENT:'

Column 2-6: 'Chromosome Start End Lods-score Branch-length'

Column 3-4*: 'chromatin: 'cell-type:chromatin-state-' for 9 cell types. *if a conserved element overlaps multiple chromatin states, then the consecutive states appear in consecutive columns

Column 5-6 (if applicable): 'gencode:' annotation

Column 7-8 (if applicable); 'motifs:' overlapping motifs for this element

Chromatin state numbers and candidate annotations:

State 1 – Active Promoter

State 2 – Weak Promoter

State 3 – Inactive/Poised Promoter

State 4,5 – Strong enhancer

State 6,7 – Weak enhancer

State 8 – Insulator

State 9 – Transcriptional transition

State 10 – Transcriptional elongation

State 11 – Weak transcribed

State 12 – Polycomb-repressed

State 13 – Heterochromatin; low signal

State 14,15 – Repetitive/Copy Number Variation

Contact: Evan Mauceli evan@broadinstitute.org

Associated GWAs SNPs overlapping constraint

Summary: SNPs and data from the NHGRI GWAS catalog, 5/30/11

File: conserved_gwas.xlsx

File format: Microsoft Excel Workbook

Contact: Luke Ward lucas.d.ward@gmail.com

Positively selected codons

Summary: Main data files and backing data for the analysis identifying positively selected codons. This data and updates are available for download from here:

<http://www.ebi.ac.uk/goldman-srv/mammals/>

File: PositivelySelectedCodons.tar.gz

Files in archive:

Pol_sel_score.bed - one region per Ensembl gene analyzed, with a score corresponding to the negative log of an overall p-value for positive selection at that gene based on mammalian alignments.

Overall dN_dS.bed - one region per Ensembl gene analyzed, with a score corresponding to the overall dN/dS at that gene based on mammalian alignments.

Sites.bedGraph - one value per codon analyzed, with a score corresponding to the signed sitewise likelihood ratio statistic for non-neutral selection. Values above zero indicate evidence for positive selection, values below zero indicate evidence for negative selection. The statistic is approximately chi-square distributed when the data is neutrally evolving. The bedGraph file looks best when displayed on UCSC the following track parameters:

```
type=bigWig
lineMark=0
lineOnOff=on
autoScale=off
viewLimits=-20:10
minLimit=-20
maxLimit=10
visibility=full
```

mammals_e57_sitewise_tables.Rdata - the .Rdata file which contains the main results tables (genes and sites)

web/ - directory containing the 'complete archive' package which contains alignments, PDFs, and other data used during the analysis

File formats: BED, bedGraph, Rdata

Format note: coordinates are on hg18

Contact: Gregory Jordan greg@ebi.ac.uk

Exapted repeats

Summary: List of exapted elements identified as described in the following reference.

File: exaptedElements.bed.gz

File format: BED

Format note: coordinates are on hg18

Contact: Craig Lowe craiglowe@gmail.com

Reference: Lowe, C. B. & Haussler, D. 29 mammalian genomes reveal novel exaptations of mobile elements for likely regulatory functions in the human genome. *In preparation* (2011).

Human and Primate Accelerated Regions

Summary: Lists of human accelerated regions (HARs) and primate accelerated regions (PARs). Regions with accelerated substitution rates in either lineage were identified by first defining candidate elements using the phastCons program (not including the lineage of interest) and then scoring those elements for accelerated substitution rates in the subtree (human or primate) of interest.

Files:

2xHARs.bed

2xPARs.bed

Format: BED

Format note: coordinates are on hg18

Contact: Katherine Pollard <katherine.pollard@gladstone.ucsf.edu>

Supplementary References.

- 1 Margulies, E. H. *et al.* An initial strategy for the systematic identification of functional elements in the human genome by low-redundancy comparative sequencing. *Proc Natl Acad Sci U S A* **102**, 4795-4800, doi:0409882102 [pii] 10.1073/pnas.0409882102 (2005).
- 2 Mikkelsen, T. S. *et al.* Genome of the marsupial *Monodelphis domestica* reveals innovation in non-coding sequences. *Nature* **447**, 167-177, doi:nature05805 [pii] 10.1038/nature05805 (2007).
- 3 Jaffe, D. B. *et al.* Whole-genome sequence assembly for mammalian genomes: Arachne 2. *Genome Res* **13**, 91-96, doi:10.1101/gr.828403 (2003).
- 4 Gnerre, S., Lander, E. S., Lindblad-Toh, K. & Jaffe, D. B. Assisted assembly: how to improve a de novo genome assembly by using related species. *Genome Biol* **10**, R88, doi:gb-2009-10-8-r88 [pii] 10.1186/gb-2009-10-8-r88 (2009).
- 5 Hubisz, M. J., Lin, M. F., Kellis, M. & Siepel, A. Error and Error Mitigation in Low-Coverage Genome Assemblies. *PLoS ONE* (2011).
- 6 Paten, B., Herrero, J., Beal, K., Fitzgerald, S. & Birney, E. Enredo and Pecan: genome-wide mammalian consistency-based multiple alignment with paralogs. *Genome Res* **18**, 1814-1828, doi:gr.076554.108 [pii] 10.1101/gr.076554.108 (2008).
- 7 Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol* **24**, 1586-1591, doi:msm088 [pii] 10.1093/molbev/msm088 (2007).
- 8 Garber, M. *et al.* Identifying novel constrained elements by exploiting biased substitution patterns. *Bioinformatics* **25**, i54-62, doi:btp190 [pii] 10.1093/bioinformatics/btp190 (2009).
- 9 Keinan, A., Mullikin, J. C., Patterson, N. & Reich, D. Measurement of the human allele frequency spectrum demonstrates greater genetic drift in East Asians than in Europeans. *Nat Genet* **39**, 1251-1255, doi:ng2116 [pii] 10.1038/ng2116 (2007).
- 10 Lin, M. F. *et al.* Revisiting the protein-coding gene catalog of *Drosophila melanogaster* using 12 fly genomes. *Genome Res* **17**, 1823-1836, doi:gr.6679507 [pii] 10.1101/gr.6679507 (2007).
- 11 Lin, M. F., Deoras, A. N., Rasmussen, M. D. & Kellis, M. Performance and scalability of discriminative metrics for comparative gene identification in 12 *Drosophila* genomes. *PLoS Comput Biol* **4**, e1000067, doi:10.1371/journal.pcbi.1000067 (2008).
- 12 DeCaprio, D. *et al.* Conrad: gene prediction using conditional random fields. *Genome Res* **17**, 1389-1398, doi:gr.6558107 [pii] 10.1101/gr.6558107 (2007).
- 13 Birney, E. *et al.* Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447**, 799-816, doi:10.1038/nature05874 (2007).
- 14 Guttman, M. *et al.* Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nat Biotechnol* **28**, 503-510, doi:nbt.1633 [pii]

- 10.1038/nbt.1633 (2010).
- 15 Finn, R. D. *et al.* The Pfam protein families database. *Nucleic Acids Res* **38**, D211-222, doi:gkp985 [pii]
- 10.1093/nar/gkp985 (2010).
- 16 Ernst, J. & Kellis, M. Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nat Biotechnol* **28**, 817-825, doi:nbt.1662 [pii]
- 10.1038/nbt.1662 (2010).
- 17 Ernst, J. *et al.* Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* **473**, 43-49, doi:nature09906 [pii]
- 10.1038/nature09906 (2011).
- 18 Lin, M. F., Jungreis, I. & Kellis, M. PhyloCSF: a comparative genomics method to distinguish protein-coding and non-coding regions. *Nature Precedings*, doi:<http://hdl.handle.net/10101/npre.2010.4784.1> (2010).
- 19 Lin, M. F. *et al.* Locating protein-coding sequences under selection for additional, overlapping functions in 29 mammalian genomes. *Genome Research* (2011).
- 20 Siepel, A. *et al.* Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* **15**, 1034-1050, doi:gr.3715005 [pii]
- 10.1101/gr.3715005 (2005).
- 21 Pedersen, J. S. *et al.* Identification and classification of conserved RNA secondary structures in the human genome. *PLoS Comput Biol* **2**, e33, doi:10.1371/journal.pcbi.0020033 (2006).
- 22 Parker, B. J. *et al.* New families of human regulatory RNA structures identified by comparative analysis of vertebrate genomes. *Genome Research* (2011).
- 23 Nawrocki, E. P., Kolbe, D. L. & Eddy, S. R. Infernal 1.0: inference of RNA alignments. *Bioinformatics* **25**, 1335-1337, doi:btp157 [pii]
- 10.1093/bioinformatics/btp157 (2009).
- 24 Hartuv, E. & Shamir, R. A clustering algorithm based on graph connectivity. *Information Processing Letters* **76**, 175-181 (2000).
- 25 Alexa, A., Rahnenfuhrer, J. & Lengauer, T. Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. *Bioinformatics* **22**, 1600-1607, doi:bt1140 [pii]
- 10.1093/bioinformatics/bt1140 (2006).
- 26 Korb, M. *et al.* The Innate Immune Database (IIDB). *BMC Immunol* **9**, 7, doi:1471-2172-9-7 [pii]
- 10.1186/1471-2172-9-7 (2008).
- 27 Pollard, K. S. *et al.* An RNA gene expressed during cortical development evolved rapidly in humans. *Nature* **443**, 167-172, doi:nature05113 [pii]
- 10.1038/nature05113 (2006).
- 28 Washietl, S., Hofacker, I. L. & Stadler, P. F. Fast and reliable prediction of noncoding RNAs. *Proc Natl Acad Sci U S A* **102**, 2454-2459, doi:0409169102 [pii]
- 10.1073/pnas.0409169102 (2005).
- 29 Gruber, A. R., Findeiss, S., Washietl, S., Hofacker, I. L. & Stadler, P. F. Rnaz 2.0: Improved Noncoding Rna Detection. *Pac Symp Biocomput* **15**, 69-79, doi:9789814295291_0009 [pii] (2010).

- 30 Washietl, S. & Hofacker, I. L. Consensus folding of aligned sequences as a new measure for the detection of functional RNAs by comparative genomics. *J Mol Biol* **342**, 19-30, doi:10.1016/j.jmb.2004.07.018
S0022-2836(04)00828-9 [pii] (2004).
- 31 Gesell, T. & Washietl, S. Dinucleotide controlled null models for comparative RNA gene prediction. *BMC Bioinformatics* **9**, 248, doi:1471-2105-9-248 [pii]
10.1186/1471-2105-9-248 (2008).
- 32 Kheradpour, P., Stark, A., Roy, S. & Kellis, M. Reliable prediction of regulator targets using 12 *Drosophila* genomes. *Genome Res* **17**, 1919-1931, doi:gr.7090407 [pii]
10.1101/gr.7090407 (2007).
- 33 Matys, V. *et al.* TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res* **31**, 374-378 (2003).
- 34 Sandelin, A., Alkema, W., Engstrom, P., Wasserman, W. W. & Lenhard, B. JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res* **32**, D91-94, doi:10.1093/nar/gkh012
32/suppl_1/D91 [pii] (2004).
- 35 Berger, M. F. *et al.* Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. *Nat Biotechnol* **24**, 1429-1435, doi:nbt1246 [pii]
10.1038/nbt1246 (2006).
- 36 Badis, G. *et al.* Diversity and complexity in DNA recognition by transcription factors. *Science* **324**, 1720-1723, doi:1162327 [pii]
10.1126/science.1162327 (2009).
- 37 Berger, M. F. *et al.* Variation in homeodomain DNA binding revealed by high-resolution analysis of sequence preferences. *Cell* **133**, 1266-1276, doi:S0092-8674(08)00683-1 [pii]
10.1016/j.cell.2008.05.024 (2008).
- 38 Touzet, H. & Varre, J. S. Efficient and accurate P-value computation for Position Weight Matrices. *Algorithms Mol Biol* **2**, 15, doi:1748-7188-2-15 [pii]
10.1186/1748-7188-2-15 (2007).
- 39 Hindorff, L. A. *et al.* Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci U S A* **106**, 9362-9367, doi:0903103106 [pii]
10.1073/pnas.0903103106 (2009).
- 40 Harrow, J. *et al.* GENCODE: producing a reference annotation for ENCODE. *Genome Biol* **7 Suppl 1**, S4 1-9, doi:gb-2006-7-s1-s4 [pii]
10.1186/gb-2006-7-s1-s4 (2006).
- 41 Altshuler, D. M. *et al.* Integrating common and rare genetic variation in diverse human populations. *Nature* **467**, 52-58, doi:nature09298 [pii]
10.1038/nature09298 (2010).
- 42 Masingham, T. & Goldman, N. Detecting amino acid sites under positive selection and purifying selection. *Genetics* **169**, 1753-1762, doi:genetics.104.032144 [pii]
10.1534/genetics.104.032144 (2005).
- 43 Vilella, A. J. *et al.* EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. *Genome Res* **19**, 327-335, doi:gr.073585.107 [pii]
10.1101/gr.073585.107 (2009).
- 44 Hubbard, T. J. *et al.* Ensembl 2007. *Nucleic Acids Res* **35**, D610-617, doi:gkl996 [pii]

- 10.1093/nar/gkl996 (2007).
- 45 Pollard, K. S., Hubisz, M. J., Rosenbloom, K. R. & Siepel, A. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res* **20**, 110-121, doi:gr.097857.109 [pii]
- 10.1101/gr.097857.109 (2010).
- 46 Zhang, J., Nielsen, R. & Yang, Z. Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. *Mol Biol Evol* **22**, 2472-2479, doi:msi237 [pii]
- 10.1093/molbev/msi237 (2005).
- 47 Lowe, C. B. & Haussler, D. 29 mammalian genomes reveal novel exaptations of mobile elements for likely regulatory functions in the human genome. *In preparation* (2011).
- 48 Visel, A., Rubin, E. M. & Pennacchio, L. A. Genomic views of distant-acting enhancers. *Nature* **461**, 199-205, doi:nature08451 [pii]
- 10.1038/nature08451 (2009).
- 49 Hubisz, M. J., Pollard, K. S. & Siepel, A. PHAST and RPHAST: phylogenetic analysis with space/time models. *Briefings in Bioinformatics* (2011).
- 50 Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc B* **57**, 289-300 (1995).
- 51 Visel, A. *et al.* ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature* **457**, 854-858, doi:nature07730 [pii]
- 10.1038/nature07730 (2009).
- 52 Heintzman, N. D. *et al.* Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature* **459**, 108-112, doi:nature07829 [pii]
- 10.1038/nature07829 (2009).
- 53 Rahmann, S., Muller, T. & Vingron, M. On the power of profiles for transcription factor binding site detection. *Stat Appl Genet Mol Biol* **2**, Article7, doi:10.2202/1544-6115.1032 (2003).

29 Mammals
Table S1

Table S1. Information on the 29 mammalian assemblies

Common name	Scientific name	Assembly	Sample source	Data generation center	GenBank accession number	Coverage (x-fold)	Total contig length (Gb)	N50 contig (kb)	N50 scaffold (gapped, kb)
Human	<i>Homo sapiens</i>	hg18	n/a	multiple	NCBI36	FINISHED	2.83	38,500	N/A
Chimpanze	<i>Pan troglodytus</i>	panTro2	n/a	multiple	AADA01000000	6.0	2.97	29	9,700
Rhesus Macaque	<i>Macaca mulatta</i>	rheMac2	n/a	multiple	AANU01000000	5.1	2.87	26	5,870
Tarsier	<i>Tarsier syrichta</i>	tarSyr1	Duke University Primate Center	WUSG	ABRT000000000	2.1	2.77	2.9	12
Mouse Lemur	<i>Microcebus murinus</i>	micMur1	Caltech	Broad	ABDC01000000	1.9	1.86	3.5	140
Bushbaby (Northern Greater Galago)	<i>Otolemur garnetti</i>	otoGar1	Duke University Primate Center	Broad	AAQR01000000	1.9	1.97	3.1	137
Tree Shrew	<i>Tupaia belangeri</i>	tupBel1	German Primate Center, Goettingen	Broad	AAPY01000000	1.9	2.14	3	124
Mouse	<i>Mus musculus</i>	mm9	n/a	multiple	Build 37	FINISHED	2.72	39,294	N/A
Rat	<i>Rattus norvegicus</i>	rn4	n/a	BCM	Baylor3.4	7.3	2.49	18,985	N/A
Kangaroo Rat	<i>Dipodomys ordii</i>	dipOrd1	Museum of Vertebrate Zoology, University of California, Berkeley,	BCM	ABRO01000000	1.8	1.84	4.3	33
Guinea Pig	<i>Cavia porcellus</i>	cavPor2	Covance Research	Broad	AAKN01000000	1.9	1.95	81	28,000
Squirrel, Thirteen-lined Ground	<i>Spermophilus tridecemlineatus</i>	speTri1	National Institute of Neurological Disorders and Stroke	Broad	AAQQ01000000	1.9	1.91	2.7	95
Rabbit	<i>Oryctolagus cuniculus</i>	oryCun1	Covance Research	Broad	AAGW02000000	2.0	2.08	3.3	55
Pika	<i>Ochotona princeps</i>	ochPri2	New Mexico Museum of Natural History	Broad	AAYZ01000000	1.9	1.92	3.3	88
Alpaca	<i>Vicugna pacos</i>	vicPac1	Binghamton University	WUSC	ABRR01000000	2.9	1.92	3.9	22
Dolphin, Bottlenosed	<i>Tursiops truncatus</i>	turTru1	Portland State University	BCM	ABRN01000000	2.8	2.30	9.7	16
Cow	<i>Bos taurus</i>	bosTau4	n/a	BCM	AAF030000000	5.4	2.87	49	1920
Horse	<i>Equus caballus</i>	equCab2	n/a	Broad	AAWR02000000	6.8	2.43	112	47,000
Cat, Domestic	<i>Felis catus</i>	felCat3	n/a	Agencourt	AANG01000000	1.9	1.64	2.4	113
Dog, Domestic	<i>Canis familiaris</i>	canFam2	n/a	Broad	AACN01000000	7.6	2.39	180	45,000
Little Brown Bat (Microbat)	<i>Myotis lucifugus</i>	myoLuc1	Center for Ecology and Conservation Biology, Boston University	Broad	AAPE01000000	1.8	1.67	3.1	93
Fruit Bat (Megabat, Flying Fox)	<i>Pteropus vampyrus</i>	pteVam1	Lubee Bat Conservancy, Gainesville	BCM	ABRP01000000	2.9	1.84	8.5	121
Hedgehog, European	<i>Erinaceus europeaus</i>	eriEur1	University of East Anglia, Norwich, UK	Broad	AANN01000000	1.9	2.13	2.8	33
Shrew, Common	<i>Sorex araneus</i>	sorAra1	Cornell University, Ithaca	Broad	AALT00000000	1.9	1.83	3.2	48
Elephant, African Savannah	<i>Loxodonta africana</i>	loxAfr2	San Diego Zoo's CRES	Broad	AAGU02000000	1.9	2.45	2.9	64
Hyrax, Rock	<i>Procavia capensis</i>	proCap1	The Dallas Zoo	BCM	ABRQ01000000	2.4	2.40	3.4	24
Tenrec	<i>Echinops telfari</i>	echTel1	Institute of Anatomy, University of Munich, Germany	Broad	AAIY01000000	1.9	2.11	3.1	48
Nine-banded Armadillo	<i>Dasypus novemcinctus</i>	dasNov2	School of Veterinary Medicine, Louisiana State University	Broad	AAGV020000000	2.3	2.37	2.4	55
Sloth, Two-toed	<i>Choloepus hoffmanni</i>	choHof1	San Diego Zoo's CRES	WUSG	ABVD01000000	2.2	2.06	2.3	10

Table S2. Constraint estimation and detection statistics

Stat	29way ω 12mers	29way ω -lods 12mers	29way π -lods 12mers
#called (kmers) fdr=0.1	44,931,962	81,923,925	110,250,366
frac called (kmers)	1.46%	2.66%	3.58%
# elements	2,759,895	3,621,583	4,463,319
# bases	80,384,823	128,766,046	165,088,933
frac. called (bases)	2.61%	4.18%	5.36%
frac. constrained (kmers)	5.34%	XX	11.99%
frac. constrained (kmers, chr-shifted)	5.44%	XX	12.33%

Stat	29way ω 12mers (BL-corr.)	29way ω -lods 12mers (BL-corr.)	29way π -lods 12mers (BL-corr.)
#called (kmers) fdr=0.1	38,744,951	14,409,822	5,411,300
frac called (kmers)	1.26%	0.47%	0.18%
# elements	2,760,769	1,232,883	471,773
# bases	74,366,747	30,098,017	11,803,440
frac. called (bases)	2.41%	0.98%	0.38%
frac. constrained (kmers)	4.18%	XX	56.33%
frac. constrained (kmers, chr-shifted)	4.22%	XX	5.63%

Stat	HMRD ω 12mers	HMRD ω -lods 12mers	HMRD π -lods 12mers
#called (kmers) fdr=0.1	965,648	12,163,613	21,385,019
frac called (kmers)	0.03%	0.39%	0.69%
# elements	295,134	1,583,803	2,546,071
# bases	4,225,299	30,086,694	54,526,055
frac. called (bases)	0.14%	0.98%	1.77%
frac. constrained (kmers)	XX	XX	10.40%
frac. constrained (kmers, chr-shifted)	XX	XX	10.52%

Stat	HMRD ω 50mers	HMRD ω -lods 50mers	HMRD π -lods 50mers
#called (kmers) fdr=0.1	42,521,200	68,775,126	82,962,293
frac called (kmers)	1.38%	2.23%	2.69%
# elements	789,570	817,021	35,057
# bases	87,972,128	114,540,295	4,646,148
frac. called (bases)	2.86%	3.72%	0.15%
frac. constrained (kmers)	5.34%	XX	11.76%
frac. constrained (kmers, chr-shifted)	5.36%	XX	11.94%

XX = Unreliable estimate due to noisy curve or lack of distinction from Gaussian

Genome len: 3,080,436,051

Table S3. Comparison of SiPhy and Phastcons elements to HMRD and vertebrate Siepel elements

a. Element description

	Elements	Total bases (Mb)	Median element size (bp)	Mean element size (bp)	Maximum element size (bp)	Minimum element size (bp)
A. 29 mammals SiPhy-omega (10%FDR)	3,621,583	129	19	36	4,182	12
B. 29 mammals SiPhy-omega (5%FDR)	1,985,021	69	20	35	2,503	12
C. 29 mammals SiPhy-pi (10%FDR)	4,463,319	165	18	37	17,137	12
D. 29 mammals SiPhy-pi (5%FDR)	2,261,877	99	21	44	5,832	12
E. 29 mammals Phastcons	2,040,420	112	29	55	3,895	1
F. MRH.D SiPhy - 50 bp	488,089	60	89	123	6,275	49
G. 5 vertebrate Phastcons - syntenic elements (Siepel 2005)	1,183,203	123	69	104	4,922	1
H. Union of HMRD and 5 vertebrates (F + G)	1,188,513	137	77	115	6,280	1

b. Overlap of sets of elements

	A	B	C	D	E	F	G	H
A. 29 mammals SiPhy-omega (10%FDR)	0	1,292,279	2,939,856	1,924,733	2,084,383	688,122	1,581,042	1,666,254
B. 29 mammals SiPhy-omega (5%FDR)	1,958,119	0	1,904,076	1,879,787	1,878,647	801,002	1,452,392	1,508,727
C. 29 mammals SiPhy-pi (10%FDR)	2,378,558	998,510	0	1,598,747	1,594,852	504,997	1,347,895	1,411,653
D. 29 mammals SiPhy-pi (5%FDR)	1,994,844	1,302,081	2,198,146	0	1,646,898	582,943	1,275,799	1,333,626
E. 29 mammals Phastcons	1,739,288	1,161,052	1,668,545	1,391,162	0	570,218	1,151,100	1,200,484
F. MRH.D SiPhy - 50 bp	421,408	384,403	411,641	403,453	426,654	0	421,603	488,089
G. 5 vertebrate Phastcons - syntenic elements (Siepel 2005)	850,816	657,568	869,285	739,129	806,837	415,833	0	1,183,203
H. Union of HMRD and 5 vertebrates (F + G)	819,674	623,443	837,631	703,506	778,697	421,163	1,122,066	0

c. Intersection between sets in basepairs

	A	B	C	D	E	F	G	H
A. 29 mammals SiPhy-omega (10%FDR)	0	67,627,303	113,238,367	89,400,914	92,714,916	44,242,045	78,301,827	82,952,319
B. 29 mammals SiPhy-omega (5%FDR)	67,627,303	0	65,981,370	66,004,495	64,535,015	35,334,387	54,961,790	57,415,163
C. 29 mammals SiPhy-pi (10%FDR)	113,238,367	65,981,370	0	95,843,873	89,646,311	46,473,954	83,096,502	89,168,007
D. 29 mammals SiPhy-pi (5%FDR)	89,400,914	66,004,495	95,843,873	0	76,958,066	43,091,056	68,617,675	73,290,379
E. 29 mammals Phastcons	92,714,916	64,535,015	89,646,311	76,958,066	0	43,205,177	75,124,448	79,480,139
F. MRH.D SiPhy - 50 bp	44,242,045	35,334,387	46,473,954	43,091,056	43,205,177	0	45,823,368	59,825,356
G. 5 vertebrate Phastcons - syntenic elements (Siepel 2005)	78,301,827	54,961,790	83,096,502	68,617,675	75,124,448	45,823,368	0	122,814,031
H. Union of HMRD and 5 vertebrates (F + G)	82,952,319	57,415,163	89,168,007	73,290,379	79,480,139	59,825,356	122,814,031	0

Table S4. Statistics on new exon predictions by CONGO and supporting evidence

Total new predictions	3789	100.00%	All new predictions, not overlapping protein-coding exons or pseudogenes in Gencode/Ensembl, RefSeq, or UCSC known genes
Breakdown by genomic territory:	1375	36.29%	Intergenic
	728	19.21%	Intron
	482	12.72%	UTR
	481	12.69%	NCExon
	76	2.01%	NCIntron
	114	3.01%	AntisenseCDS
	313	8.26%	AntisenseIntron
	103	2.72%	AntisenseUTR
	45	1.19%	AntisenseNCExon
71	1.87%	AntisenseNCIntron	
Predictions supported by different lines of evidence (not mutually exclusive):			
	394	10.40%	Pfam domain
	895	23.62%	Mrna (cDNA)
	1318	34.78%	IntronEst
	1640	43.28%	Scripture transcripts (Illumina human body map)

Table S5. Predicted new protein-coding exons by CONGO

id	chrom	start	end	strand	phase	length	nt	CONGO score	territory	seqdup	Pfam domains	mrna	IntronEst	ScriptureTU	ScriptureLinkage	ScriptureLinkageTo	
CONGO_chr1_000849675_517_p0	chr1	849675	850191	+	0	517	1048.16	UTR			SAND		1	TU689	SpliceToPCG	ENSG00000187634	
CONGO_chr1_000906021_252_m0	chr1	906021	906272	-	0	252	512.23	Intron						TU690	SpliceToPCG	ENSG00000187642	
CONGO_chr1_000959799_148_p0	chr1	959799	959946	+	0	148	704.49	NCExon					1	TU691	SpliceToPCG	ENSG00000188157	
CONGO_chr1_002351817_112_p0	chr1	2351817	2351928	+	0	112	56.39	NCExon					1	TU694	SpliceToPCG	ENSG00000149527	
CONGO_chr1_002706469_119_m1	chr1	2706469	2706587	-	1	119	550.24	Intergenic									
CONGO_chr1_002706737_513_m0	chr1	2706737	2707249	-	0	513	2692.97	Intergenic									
CONGO_chr1_006048712_212_p0	chr1	6048712	6048923	+	0	212	1139.1	Intron									
CONGO_chr1_006437081_645_p0	chr1	6437081	6437275	+	0	645	2611.29	Intron									
CONGO_chr1_006437885_960_p0	chr1	6437885	6438844	+	0	960	3025.77	Intron									
CONGO_chr1_006488673_133_m0	chr1	6488673	6488805	-	0	133	663.25	Intron									
CONGO_chr1_007743350_179_p0	chr1	7743350	7743428	+	0	179	5.26	Intron						TU695	SpliceToPCG	ENSG00000049245	
CONGO_chr1_008596809_126_p1	chr1	8596809	8597034	+	1	126	132.74	AntisenseIntron									
CONGO_chr1_008853471_59_m2	chr1	8853471	8853529	-	2	59	464.44	UTR					1	TU696	SpliceToPCG	ENSG00000074800	
CONGO_chr1_009933876_67_m0	chr1	9933876	9933942	-	0	67	469.19	AntisenseIntron									
CONGO_chr1_009935031_84_m2	chr1	9935031	9935114	-	2	84	407.43	AntisenseIntron									
CONGO_chr1_009938466_80_m1	chr1	9938466	9938545	-	1	80	394.73	AntisenseIntron						TU697	NovelMultiExon		
CONGO_chr1_009943359_74_m0	chr1	9943359	9943432	-	0	74	263.51	AntisenseIntron									
CONGO_chr1_010089867_153_p1	chr1	10089867	10090019	+	1	153	100.13	NCExon					1	TU698	SpliceToPCG	ENSG00000130939	
CONGO_chr1_010326735_201_p2	chr1	10326735	10326935	+	2	201	326.92	Intron									
CONGO_chr1_010624667_223_m0	chr1	10624667	10624889	-	0	223	111.23	Intron						TU699	SpliceToPCG	ENSG00000130940	
CONGO_chr1_010703959_110_m2	chr1	10703959	10704068	-	2	110	34.12	Intron									
CONGO_chr1_010928863_172_m1	chr1	10928863	10929034	-	1	172	1.34	Intergenic									
CONGO_chr1_012938056_264_m1	chr1	12938056	12938319	-	1	264	237.3	Intron					1				
CONGO_chr1_016409222_118_m0	chr1	16409222	16409339	-	0	118	191.28	Intron									
CONGO_chr1_021746484_119_p0	chr1	21746484	21746602	+	0	119	501.95	Intron									
CONGO_chr1_022067965_84_m0	chr1	22067965	22068048	-	0	84	243.46	Intron						TU124	SpliceToPCG	ENSG00000142798	
CONGO_chr1_022068546_198_m0	chr1	22068546	22068743	-	0	198	285.39	Intron						TU124	SpliceToPCG	ENSG00000142798	
CONGO_chr1_022069914_39_m0	chr1	22069914	22069952	-	0	39	25.31	Intron						TU124	SpliceToPCG	ENSG00000142798	
CONGO_chr1_022149815_165_m0	chr1	22149815	22149979	-	0	165	748.93	Intergenic									
CONGO_chr1_022150126_154_m0	chr1	22150126	22150279	-	0	154	51.12	Intergenic									
CONGO_chr1_022150402_125_m0	chr1	22150402	22150526	-	0	125	105.62	Intergenic									
CONGO_chr1_023293410_51_m1	chr1	23293410	23293460	-	1	51	32.32	UTR					1	TU700	SpliceToPCG	ENSG00000169641	
CONGO_chr1_024160045_167_p0	chr1	24160045	24160211	+	0	167	122.13	UTR					1	TU701	SpliceToPCG	ENSG00000189266	
CONGO_chr1_026363239_295_m1	chr1	26363239	26363533	-	1	295	891.37	Intergenic									
CONGO_chr1_026364904_98_m0	chr1	26364904	26365001	-	0	98	452.66	Intergenic									
CONGO_chr1_026365113_84_m0	chr1	26365113	26365196	-	0	84	382.41	Intergenic									
CONGO_chr1_026367494_86_m2	chr1	26367494	26367579	-	2	86	352.18	Intergenic									
CONGO_chr1_026370267_147_m0	chr1	26370267	26370413	-	0	147	737.94	AntisenseUTR									
CONGO_chr1_026370695_45_m0	chr1	26370695	26370739	-	0	45	175.02	AntisenseUTR									
CONGO_chr1_026553738_126_m0	chr1	26553738	26553863	-	0	126	15.66	Intergenic									
CONGO_chr1_026555537_64_m1	chr1	26555537	26555600	-	1	64	62.88	Intergenic									
CONGO_chr1_026671455_115_p1	chr1	26671455	26671569	+	1	115	32.19	UTR									
CONGO_chr1_026922591_97_m2	chr1	26922591	26922687	-	2	97	55.5	AntisenseIntron						1	TU702	SpliceToPCG	ENSG00000198830
CONGO_chr1_027050207_42_m0	chr1	27050207	27050248	-	0	42	51.02	AntisenseCDS									
CONGO_chr1_027770933_92_m2	chr1	27770933	27771024	-	2	92	63.34	Intron						TU703	SpliceToPCG	ENSG00000126705	
CONGO_chr1_027824198_52_m1	chr1	27824198	27824249	-	1	52	3.73	UTR						1	TU704	SpliceToPCG	ENSG00000000938
CONGO_chr1_028791611_233_p0	chr1	28791611	28791843	+	0	233	886.32	UTR			Ras			1	TU705	SpliceToPCG	ENSG00000188060
CONGO_chr1_028823045_72_m0	chr1	28823045	28823116	-	0	72	132.53	NCExon						1	TU706	SpliceToPCG	ENSG00000120656
CONGO_chr1_029086139_171_p0	chr1	29086139	29086309	+	0	171	71.17	UTR						1	TU125	SpliceToPCG	ENSG00000159023
CONGO_chr1_029267832_51_p0	chr1	29267832	29267882	+	0	51	30.96	Intron						TU125	SpliceToPCG	ENSG00000159023	
CONGO_chr1_029268265_129_p0	chr1	29268265	29268393	+	0	129	463.29	Intron						TU125	SpliceToPCG	ENSG00000159023	
CONGO_chr1_029269571_144_p0	chr1	29269571	29269714	+	0	144	102.51	Intron									
CONGO_chr1_029365874_92_m1	chr1	29365874	29365965	-	1	92	107.11	NCExon						1	TU707	SpliceToPCG	ENSG00000116350
CONGO_chr1_032157755_92_m0	chr1	32157755	32157846	-	0	92	34.1	UTR						1	TU708	SpliceToPCG	ENSG00000184007
CONGO_chr1_03282949_60_m0	chr1	3282949	3283008	-	0	60	97.99	NCIntron									
CONGO_chr1_032565594_53_p1	chr1	32565594	32565646	+	1	53	55.09	NCExon						1	TU709	SpliceToPCG	ENSG00000116478
CONGO_chr1_035422839_65_p1	chr1	35422839	35422903	+	1	65	5.09	AntisenseIntron									
CONGO_chr1_035422908_94_m0	chr1	35422908	35423001	-	0	94	200.1	Intron									
CONGO_chr1_036323353_115_m1	chr1	36323353	36323467	-	1	115	965	AntisenseCDS									
CONGO_chr1_036532238_84_p1	chr1	36532238	36532321	+	1	84	46.39	NCExon						TU712	SpliceToPCG	ENSG000000054118	
CONGO_chr1_037216344_96_p2	chr1	37216344	37216439	+	2	96	59.27	AntisenseIntron									
CONGO_chr1_037479084_87_p0	chr1	37479084	37479170	+	0	87	36.33	Intergenic									
CONGO_chr1_038509057_81_m0	chr1	38509057	38509137	-	0	81	8.95	Intergenic									
CONGO_chr1_038619746_54_m2	chr1	38619746	38619799	-	2	54	66.46	Intergenic									
CONGO_chr1_038785616_97_m1	chr1	38785616	38785712	-	1	97	26.48	Intergenic									
CONGO_chr1_038800099_92_p0	chr1	38800099	38800190	+	0	92	11.71	Intergenic									
CONGO_chr1_039397228_52_p0	chr1	39397228	39397279	+	0	52	204.48	Intron									
CONGO_chr1_039490105_96_p0	chr1	39490105	39490200	+	0	96	75.1	Intron						TU714	SpliceToPCG	ENSG00000127603	
CONGO_chr1_039507395_115_m2	chr1	39507395	39507509	-	2	115	430.7	AntisenseCDS									
CONGO_chr1_039537191_73_p2	chr1	39537191	39537263	+	2	73	7.64	Intron									
CONGO_chr1_040860353_141_m0	chr1	40860353	40860493	-	0	141	6	UTR									
CONGO_chr1_040947438_31_p2	chr1	40947438	40947468	+	2	31	1.24	Intron									
CONGO_chr1_040947526_137_m2	chr1	40947526	40947662	-	2	137	56.83	AntisenseUTR									
CONGO_chr1_041867007_68_m0	chr1	41867007	41867074	-	0	68	391.27	UTR									
CONGO_chr1_044512393_176_p2	chr1	44512393	44512568	+	2	176	38.82	AntisenseIntron									
CONGO_chr1_044762974_108_p0	chr1	44762974	44763081	+	0	108	19.53	Intron									
CONGO_chr1_046771316_478_m1	chr1	46771316	467														

CONGO_chr12_022559160_65_m1	chr12	22559160	22559224	-	1	65	12.07	Intron				TU442	SpliceToPCG	ENSG00000111731		
CONGO_chr12_024125654_75_m0	chr12	24125654	24125728	-	0	75	54.83	Intron								
CONGO_chr12_024372630_75_m0	chr12	24372630	24372704	-	0	75	48.99	Intron								
CONGO_chr12_024804615_63_m0	chr12	24804615	24804677	-	0	63	32.14	Intergenic								
CONGO_chr12_025048570_150_p0	chr12	25048570	25048719	+	0	150	234.43	Intergenic								
CONGO_chr12_025061335_159_p1	chr12	25061335	25061493	+	1	159	552.9	Intergenic				1				
CONGO_chr12_025074479_57_p0	chr12	25074479	25074535	+	0	57	171.14	Intergenic								
CONGO_chr12_025076506_65_p0	chr12	25076506	25076570	+	0	65	69.1	Intergenic								
CONGO_chr12_025078049_100_p1	chr12	25078049	25078148	+	1	100	18.92	Intergenic								
CONGO_chr12_025097062_96_p2	chr12	25097062	25097157	+	2	96	31.46	UTR			1	1	TU443	SpliceToPCG	ENSG00000118308	
CONGO_chr12_029827650_321_m1	chr12	29827650	29827970	+	1	321	110.16	Intron			1	1	TU444	SpliceToPCG	ENSG00000133687	
CONGO_chr12_030350265_91_p0	chr12	30350265	30350355	+	0	91	388.79	Intergenic								
CONGO_chr12_031081488_159_m0	chr12	31081488	31081646	+	0	159	161.91	NCIntron								
CONGO_chr12_032483995_107_p0	chr12	32483995	32484098	+	0	104	54.76	Intergenic				1				
CONGO_chr12_032608390_101_p1	chr12	32608390	32608490	+	0	101	29.88	UTR				1	TU445	SpliceToPCG	ENSG00000139132	
CONGO_chr12_046470256_170_m1	chr12	46470256	46470245	+	1	170	30.9	NCExon			1	1	TU446	SpliceToPCG	ENSG00000061273	
CONGO_chr12_047035877_141_m1	chr12	47035877	47036017	+	1	141	207.81	Intergenic								
CONGO_chr12_047036542_96_m0	chr12	47036542	47036637	+	0	96	75.03	Intergenic								
CONGO_chr12_047065476_200_m2	chr12	47065476	47065675	+	2	200	202.03	Intergenic								
CONGO_chr12_047065908_171_m0	chr12	47065908	47066078	+	0	171	34.08	Intergenic								
CONGO_chr12_047396728_55_p1	chr12	47396728	47396782	+	1	55	180.65	AntisenseUTR					TU447	NovelMultiExon		
CONGO_chr12_047667083_94_p1	chr12	47667083	47667176	+	1	94	29.3	Intergenic								
CONGO_chr12_048246790_102_m0	chr12	48246790	48246891	+	0	102	92.3	UTR			1	1	TU448	SpliceToPCG	ENSG00000187778	
CONGO_chr12_048338301_85_m0	chr12	48338301	48338385	+	0	85	88.82	Intron					TU449	SpliceToPCG	ENSG00000161791	
CONGO_chr12_048753635_301_p0	chr12	48753635	48753935	+	0	301	622.65	Intron								
CONGO_chr12_048753953_342_p0	chr12	48753953	48754294	+	0	342	236.15	Intron								
CONGO_chr12_049877820_254_m2	chr12	49877820	49878073	+	2	254	186.24	UTR			1	1	TU45	SpliceToPCG	ENSG00000170545	
CONGO_chr12_049878600_226_m0	chr12	49878600	49878825	+	0	226	610.19	UTR			1	1	TU45	SpliceToPCG	ENSG00000170545	
CONGO_chr12_049879797_135_m0	chr12	49879797	49879931	+	0	135	769.35	UTR			1	1	TU45	SpliceToPCG	ENSG00000170545	
CONGO_chr12_049884223_196_m0	chr12	49884223	49884418	+	0	196	1051.32	UTR			1	1	TU45	SpliceToPCG	ENSG00000170545	
CONGO_chr12_049886840_48_m0	chr12	49886840	49886887	+	0	48	243	UTR			1	1	TU45	SpliceToPCG	ENSG00000170545	
CONGO_chr12_050494264_681_m0	chr12	50494264	50494444	+	0	681	943.99	Intergenic								
CONGO_chr12_050946597_152_p0	chr12	50946597	50946798	+	0	152	138.78	Intergenic								
CONGO_chr12_050947014_117_p0	chr12	50947014	50947130	+	0	117	46.56	Intergenic								
CONGO_chr12_050947396_197_p0	chr12	50947396	50947592	+	0	197	141.3	Intergenic								
CONGO_chr12_050955872_105_p0	chr12	50955872	50955974	+	0	105	420.33	UTR				1	TU450	SpliceToPCG	ENSG00000170442	
CONGO_chr12_051021373_295_p1	chr12	51021373	51021667	+	1	295	914.71	Intergenic								
CONGO_chr12_051024272_209_p0	chr12	51024272	51024480	+	0	209	489.46	Intergenic								
CONGO_chr12_051026223_61_p1	chr12	51026223	51026283	+	1	61	494.58	Intergenic								
CONGO_chr12_051027314_45_p0	chr12	51027314	51027358	+	0	45	97.12	Intergenic								
CONGO_chr12_051028052_87_p0	chr12	51028052	51028138	+	0	87	347.69	Intergenic								
CONGO_chr12_051028156_45_m0	chr12	51028156	51028200	+	0	45	18.01	Intergenic				1				
CONGO_chr12_051028438_126_p0	chr12	51028438	51028563	+	0	126	14.96	Intergenic								
CONGO_chr12_051030164_221_p0	chr12	51030164	51030384	+	0	221	1312.66	Intergenic								
CONGO_chr12_051033172_32_p1	chr12	51033172	51033203	+	1	32	261.82	Intergenic								
CONGO_chr12_051091804_186_m2	chr12	51091804	51091989	+	2	186	172.87	Intergenic								
CONGO_chr12_051094012_152_m0	chr12	51094012	51094163	+	0	152	541.77	Intergenic					TU451	SpliceToPCG	ENSG00000170421	
CONGO_chr12_051096025_143_m0	chr12	51096025	51096167	+	0	143	342.04	Intergenic								
CONGO_chr12_051098402_78_m0	chr12	51098402	51098479	+	0	78	8.32	Intergenic				1				
CONGO_chr12_051101710_209_m0	chr12	51101710	51101918	+	0	209	581.48	Intergenic								
CONGO_chr12_051313513_74_m0	chr12	51313513	51313586	+	0	74	7.01	Intergenic								
CONGO_chr12_051395628_161_m0	chr12	51395628	51395788	+	0	161	520.56	Intergenic								
CONGO_chr12_051396516_126_m0	chr12	51396516	51396641	+	0	126	406.56	Intergenic								
CONGO_chr12_051397787_84_m0	chr12	51397787	51397870	+	0	84	133.36	Intergenic								
CONGO_chr12_051398219_100_m0	chr12	51398219	51398318	+	0	100	75.15	Intergenic								
CONGO_chr12_051399020_61_m1	chr12	51399020	51399080	+	1	61	417.62	Intergenic								
CONGO_chr12_051401216_291_m0	chr12	51401216	51401506	+	0	291	44.18	Intergenic				1				
CONGO_chr12_051428282_83_m2	chr12	51428282	51428364	+	2	83	1.42	Intergenic								
CONGO_chr12_051433360_152_m0	chr12	51433360	51433511	+	0	152	295.53	Intergenic								
CONGO_chr12_051724637_75_p0	chr12	51724637	51724711	+	0	75	160.67	AntisenseUTR								
CONGO_chr12_051921974_277_p0	chr12	51921974	51922250	+	0	277	423.03	Intergenic							NovelUnspliced	
CONGO_chr12_052132158_270_p2	chr12	52132158	52132427	+	2	270	64.29	UTR				1	1	TU184	SpliceToPCG	ENSG00000197111
CONGO_chr12_052154920_349_p1	chr12	52154920	52155268	+	1	349	369.52	Intron				1	1	TU184	SpliceToPCG	ENSG00000197111
CONGO_chr12_052696814_142_p2	chr12	52696814	52696955	+	2	142	66.26	UTR					TU185	SpliceToPCG	ENSG00000172789	
CONGO_chr12_052708503_56_p1	chr12	52708503	52708558	+	1	56	72.88	UTR				1	TU185	SpliceToPCG	ENSG00000172789	
CONGO_chr12_053032379_91_p1	chr12	53032379	53032469	+	1	91	59.77	Intergenic								
CONGO_chr12_053796384_142_p1	chr12	53796384	53796525	+	1	142	468.75	Intergenic								
CONGO_chr12_053874319_234_m0	chr12	53874319	53874552	+	0	234	1242.35	Intergenic							71m_1	
CONGO_chr12_053874620_110_m2	chr12	53874620	53874729	+	2	110	183.54	Intergenic							71m_1	
CONGO_chr12_053874761_107_m0	chr12	53874761	53874867	+	0	107	85.87	Intergenic								
CONGO_chr12_054022914_408_p2	chr12	54022914	54023321	+	2	408	2157.8	Intergenic							71m_1	
CONGO_chr12_054023337_80_p1	chr12	54023337	54023416	+	1	80	146.33	Intergenic								
CONGO_chr12_054023472_351_p2	chr12	54023472	54023822	+	2	351	2032.79	Intergenic							71m_1	
CONGO_chr12_054291674_281_p0	chr12	54291674	54291954	+	0	281	1016.17	Intergenic							71m_1	
CONGO_chr12_054291998_563_p2	chr12	54291998	54292560	+	2	563	2207.98	Intergenic							71m_1	
CONGO_chr12_054431554_33_m0	chr12	54431554	54431586	+	0	33	30.85	AntisenseUTR					TU453	SpliceToPCG	ENSG00000135392	
CONGO_chr12_054698189_40_p0	chr12	54698189	54698228	+	0	40	37.03	Intergenic				1	1	TU455	SpliceToPCG	ENSG00000123411
CONGO_chr12_054701432_63_m1	chr12	54701432	54701494	+	1	63	182.36	AntisenseCDS								
CONGO_chr12_054701520_133_m2	chr12	54701520	54701652	+	2	133	55.7	AntisenseCDS								
CONGO_chr12_054798284_86_p1	chr12	54798284	54798369	+	1	86	98.92	UTR				1	TU187	SpliceToPCG	ENSG00000092841	
CONGO_chr12_054799139_64_p2	chr12	54799139	54799202	+	2	64	105.49	UTR				1	TU187	SpliceToPCG	ENSG00000092841	
CONGO_chr12_054919829_86_p2	chr12	54919829	54919914	+	2	86	14.63	AntisenseUTR					TU456	UnsplicedMergeWithPCG	ENSG00000139645	
CONGO_chr12_055368185_122_p1	chr12	55368185	55368306	+	1	122	151.04	AntisenseUTR								
CONGO_chr12_055663545_114_m1	chr12	55663545	55663658	+	1	114	366.79	Intergenic								
CONGO_chr12_055665733_97_m2	chr12	55665733	55665829	+	2	97	5.66	Intergenic								
CONGO_chr12_055665923_89_m0	chr12	55665923	55666011	+	0	89	56.74	Intergenic								
CONGO_chr12_055848235_51_p2	chr12	55848235	55848285	+	2	51	274.36	NCExon				1	TU457	SpliceToPCG	ENSG00000123384	
CONGO_chr12_056209294_73_m2	chr12	56209294	56209366	+	2	73	115.21	AntisenseUTR					TU458	SpliceToPCG	ENSG00000175203	
CONGO_chr12_056299620_309_p2	chr12	56299620	56299928	+	2	309	573.21	NCExon					TU459	SpliceToPCG	ENSG00000135502	
CONGO_chr12_056304781_63_p0	chr12	5630														

CONGO chr14_076561307_91_m1	chr14	76561307	76561397	-	1	91	69.7	UTR				UnsplicedMergeWithPCG	ENSG00000119669
CONGO chr14_07779185_314_p2	chr14	77779185	77779498	+	2	314	2330.42	NCExon	Laminin_G_2	1	TU105	SpliceToPCG	ENSG00000021645
CONGO chr14_07779526_373_p0	chr14	77779526	77779898	+	0	373	3531.17	NCExon	EGF_Laminin_G_2	1	TU105	SpliceToPCG	ENSG00000021645
CONGO chr14_078344705_60_m2	chr14	78344705	78344764	-	2	60	132.59	AntisenseIntron					
CONGO chr14_079389635_110_p2	chr14	79389635	79389744	+	2	110	222.03	NCExon		1	TU105	SpliceToPCG	ENSG00000021645
CONGO chr14_080804511_84_m0	chr14	80804511	80804594	-	0	84	522.82	Intron			TU521	SpliceToPCG	ENSG00000140022
CONGO chr14_090141685_171_m2	chr14	90141685	90141855	-	2	171	337.01	Intron			TU522	SpliceToPCG	ENSG00000165914
CONGO chr14_090821767_133_m1	chr14	90821767	90821899	-	1	133	176.05	Intron			TU523	SpliceToPCG	ENSG00000015133
CONGO chr14_092966755_243_m0	chr14	92966755	92966997	-	0	243	187.27	AntisenseIntron					
CONGO chr14_093002896_103_p0	chr14	93002896	93002998	+	0	103	13.61	Intron					
CONGO chr14_093156991_150_p2	chr14	93156991	93157140	+	2	150	119.54	Intron			TU524	SpliceToPCG	ENSG00000133958
CONGO chr14_093544708_62_p0	chr14	93544708	93544769	+	0	62	26.75	UTR		1	TU525	SpliceToPCG	ENSG00000175699
CONGO chr14_093883636_185_m0	chr14	93883636	93883820	-	0	185	162.97	Intergenic					
CONGO chr14_094729409_97_m1	chr14	94729409	94729505	+	1	97	4.3	Intron					
CONGO chr14_095458871_166_p0	chr14	95458871	95459056	+	0	186	1914.25	NCExon		1	TU526	SpliceToPCG	ENSG00000165959
CONGO chr14_097109420_144_p0	chr14	97109420	97109563	+	0	144	61.88	Intergenic					
CONGO chr14_097169388_59_p2	chr14	97169388	97169445	+	2	59	7.07	AntisenseUTR					
CONGO chr14_098169608_77_p1	chr14	98169608	98169684	+	1	77	48.14	Intergenic					
CONGO chr14_098346292_100_p2	chr14	98346292	98346391	+	2	100	29.42	Intergenic					
CONGO chr14_099428229_57_p2	chr14	99428229	99428285	+	2	57	402.85	Intron			TU527	SpliceToPCG	ENSG00000066629
CONGO chr14_099814047_128_m0	chr14	99814047	99814174	-	0	128	9.68	AntisenseUTR		1	TU528	SpliceToPCG	ENSG00000197119
CONGO chr14_100086510_90_m1	chr14	100086510	100086599	-	1	90	16.51	Intron					
CONGO chr14_100113442_332_m1	chr14	100113442	100113773	-	1	332	208.79	Intergenic					
CONGO chr14_102624175_157_p0	chr14	102624175	102624331	+	0	157	590.21	Intergenic			TU529	NovelMultiExon	
CONGO chr14_103476896_298_m1	chr14	103476896	103477193	+	1	298	1465.18	NCExon		1	TU198	SpliceToPCG	ENSG00000156411
CONGO chr14_103477487_281_m0	chr14	103477487	103477767	-	0	281	1776.66	NCExon		1	TU198	SpliceToPCG	ENSG00000156411
CONGO chr14_103672555_138_p0	chr14	103672555	103672692	+	0	138	70.21	Intergenic					
CONGO chr14_105159294_84_m0	chr14	105159294	105159377	-	0	84	114.19	Intron		1	TU199	NovelMultiExon	
CONGO chr14_105160450_131_m2	chr14	105160450	105160580	-	2	131	431.45	Intron		1	TU199	NovelMultiExon	
CONGO chr14_105178347_84_m0	chr14	105178347	105178430	-	0	84	13.58	NCExon		1	TU200	NovelMultiExon	
CONGO chr14_105179161_131_m2	chr14	105179161	105179291	-	2	131	89.81	NCExon		1	TU200	NovelMultiExon	
CONGO chr15_023124120_92_p2	chr15	23124120	23124211	+	2	92	7.72	NCIntron					
CONGO chr15_026662711_132_p2	chr15	26662711	26662842	+	2	132	131.44	Intergenic		1	TU106	NovelMultiExon	
CONGO chr15_02673453_136_p0	chr15	2673453	2673598	+	0	136	215.75	Intergenic		1	TU106	NovelMultiExon	
CONGO chr15_02676824_171_p2	chr15	2676824	2676994	+	1	171	201.81	Intergenic	Cyl-b5	1	TU106	SpliceToPCG	ENSG00000206149
CONGO chr15_026775918_126_p2	chr15	26775918	26776043	+	2	126	189.36	Intergenic		1	TU107	SpliceToPCG	ENSG00000206149
CONGO chr15_026717867_102_p1	chr15	26717867	26717968	+	1	102	148.11	Intron		1	TU107	SpliceToPCG	ENSG00000206149
CONGO chr15_026718075_140_p1	chr15	26718075	26718214	+	1	140	111.94	UTR	ZZ	1	TU107	SpliceToPCG	ENSG00000206149
CONGO chr15_028035727_133_m1	chr15	28035727	28035859	-	1	133	1555.61	NCExon		1	TU531	SpliceToPCG	ENSG00000104067
CONGO chr15_028152115_97_m0	chr15	28152115	28152211	-	0	97	296.33	NCExon		1			
CONGO chr15_029240401_54_m0	chr15	29240401	29240454	-	0	54	406.53	Intergenic		1			
CONGO chr15_031198742_120_m2	chr15	31198742	31198861	-	2	120	146.26	Intron					
CONGO chr15_033333263_160_p1	chr15	33333263	33333422	+	1	160	1.33	Intergenic					
CONGO chr15_033333721_75_m2	chr15	33333721	33333795	-	2	75	30.38	Intergenic					
CONGO chr15_033645385_147_m0	chr15	33645385	33645531	-	0	147	33.87	AntisenseNCIntron					
CONGO chr15_033832569_85_m1	chr15	33832569	33832653	-	1	85	28	AntisenseNCIntron					
CONGO chr15_033918414_64_m2	chr15	33918414	33918477	-	2	64	8.09	AntisenseNCIntron					
CONGO chr15_034575049_220_m1	chr15	34575049	34575268	-	1	220	74.45	Intergenic					
CONGO chr15_034973571_115_m1	chr15	34973571	34973685	-	1	115	11.18	Intron			TU532	SpliceToPCG	ENSG00000134138
CONGO chr15_035028469_68_p1	chr15	35028469	35028536	+	1	68	194.81	AntisenseIntron					
CONGO chr15_035028585_181_m1	chr15	35028585	35028765	+	1	181	55.34	Intron					
CONGO chr15_035028891_108_p0	chr15	35028891	35028998	+	0	108	69.59	AntisenseIntron					
CONGO chr15_035136929_67_p2	chr15	35136929	35136995	+	2	67	20.31	AntisenseIntron					
CONGO chr15_035161635_99_m1	chr15	35161635	35161733	-	1	99	106.97	Intron					
CONGO chr15_035301017_92_p1	chr15	35301017	35301108	+	1	92	11.43	Intergenic					
CONGO chr15_035438962_113_m2	chr15	35438962	35439074	+	2	113	15.25	Intergenic					
CONGO chr15_035441177_147_m0	chr15	35441177	35441323	-	0	147	38.3	Intergenic					
CONGO chr15_035959557_104_m2	chr15	35959557	35959660	+	2	104	19.64	Intergenic					
CONGO chr15_035983897_138_m0	chr15	35983897	35984034	-	0	138	4.84	Intergenic					
CONGO chr15_038403549_462_p0	chr15	38403549	38404010	-	0	462	270.58	NCExon		1	TU533	NovelMultiExon	
CONGO chr15_038992229_61_m2	chr15	38992229	38992289	-	2	61	40.42	Intergenic					
CONGO chr15_039299320_69_m2	chr15	39299320	39299388	-	2	69	456.72	Intron		1	TU201	SpliceToPCG	ENSG00000178997
CONGO chr15_039305933_133_m0	chr15	39305933	39306065	-	0	133	399.96	Intron		1	TU201	SpliceToPCG	ENSG00000178997
CONGO chr15_040453777_114_p0	chr15	40453777	40453890	-	0	114	463.59	Intron					
CONGO chr15_041648676_150_m0	chr15	41648676	41648825	-	0	150	242.66	NCExon		1			
CONGO chr15_041958989_172_m0	chr15	41958989	41959160	-	0	172	758.96	Intron			TU534	SpliceToPCG	ENSG00000171877
CONGO chr15_045444453_40_m0	chr15	45444453	45444492	-	0	40	45.02	AntisenseIntron					
CONGO chr15_046305368_69_p2	chr15	46305368	46305436	+	2	69	0.87	UTR		1	TU535	SpliceToPCG	ENSG00000074803
CONGO chr15_048580190_63_m1	chr15	48580190	48580252	-	1	63	1.05	UTR		1	TU537	SpliceToPCG	ENSG00000092439
CONGO chr15_050003529_105_m0	chr15	50003529	50003633	-	0	105	182.79	AntisenseNCIntron			TU65	SpliceToPCG	ENSG00000166477
CONGO chr15_050005234_135_m1	chr15	50005234	50005368	-	1	135	771.48	AntisenseNCIntron	Leo1		TU65	SpliceToPCG	ENSG00000166477
CONGO chr15_050006267_95_m0	chr15	50006267	50006361	-	0	95	414.95	AntisenseNCIntron	Leo1		TU65	SpliceToPCG	ENSG00000166477
CONGO chr15_050008822_146_m0	chr15	50008822	50008967	-	0	146	11.78	AntisenseNCExon			TU65	SpliceToPCG	ENSG00000166477
CONGO chr15_050125319_70_p1	chr15	50125319	50125388	+	1	70	141.78	UTR		1	TU538	SpliceToPCG	ENSG00000069956
CONGO chr15_051056351_63_p0	chr15	51056351	51056413	-	0	63	27.49	Intergenic					
CONGO chr15_053397118_165_m0	chr15	53397118	53397282	-	0	165	151.15	NCExon		1	TU539	SpliceToPCG	ENSG00000069974
CONGO chr15_053699351_136_m1	chr15	53699351	53699486	-	1	136	169.14	UTR		1	TU540	SpliceToPCG	ENSG00000166450
CONGO chr15_055820716_111_m0	chr15	55820716	55820826	-	0	111	4.75	AntisenseNCIntron					
CONGO chr15_058283898_63_m0	chr15	58283898	58283960	-	0	63	3.55	Intergenic					
CONGO chr15_058343875_73_p1	chr15	58343875	58343947	+	1	73	92.83	Intergenic					
CONGO chr15_058604967_48_m2	chr15	58604967	58605014	-	2	48	100.9	Intron					
CONGO chr15_058756642_82_m1	chr15	58756642	58756723	-	1	82	170.87	Intron		1			
CONGO chr15_061679949_234_p1	chr15	61679949	61680182	+	1	234	485.73	UTR		1	TU541	SpliceToPCG	ENSG00000140455
CONGO chr15_062002122_474_m0	chr15	62002122	62002595	-	0	474	2567.48	Intron			TU542	SpliceToPCG	ENSG00000035664
CONGO chr15_063991474_119_m2	chr15	63991474	63991592	-	2	119	131.13	NCExon		1	TU543	SpliceToPCG	ENSG00000157890
CONGO chr15_064986651_101_m2	chr15	64986651	64986761	-	2	101	63.54	Intergenic					
CONGO chr15_065800214_74_p0	chr15	65800214	65800287	+	0	74	0.63	Intron					
CONGO chr15_067504675_42_p1	chr15	67504675	67504716	+	1	42	33.8	NCExon		1	TU544	SpliceToPCG	ENSG00000137807
CONGO chr15_070576186_68_p2	chr15	70576186	70576253	+	2	68	210.51	Intron		1	TU545	SpliceToPCG	ENSG00000166233
CONGO chr15_073305452_107_m2	chr15	73305452	73305558	-	2	107	14.79	Intergenic		1			
CONGO chr15_075975765_1034_p1	chr15	75975765	75976798	+	1	1034	2757.52	Intergenic	Laminin_G_2				
CONGO chr15_075976862_887_p0	chr15	75976862	75977748	+	0	887	2682.95	Intergenic		1			
CONGO chr15_075977774_162_p0	chr15	759777											

CONGO_chr17_024430875_1270_m1	chr17	24430875	24432144	-	1	1270	9466.29	Intron		1	TU67	SpliceToPCG	ENSG00000196535	
CONGO_chr17_024477037_75_m0	chr17	24477037	24477111	-	0	75	29.42	Intron		1	TU67	SpliceToPCG	ENSG00000196535	
CONGO_chr17_024479219_432_m1	chr17	24479219	24479650	-	1	432	1582.74	UTR		1	TU67	SpliceToPCG	ENSG00000196535	
CONGO_chr17_024482527_60_m0	chr17	24482527	24482586	-	0	60	190.23	UTR		1	TU67	SpliceToPCG	ENSG00000196535	
CONGO_chr17_024638477_59_p2	chr17	24638477	24638535	+	2	59	147	AntisenseCDS		1	TU600	NovelMultiExon		
CONGO_chr17_027573891_144_p1	chr17	27573891	27574034	+	1	144	310.79	Intron		1	TU601	SpliceToPCG	ENSG00000126858	
CONGO_chr17_027729047_66_p0	chr17	27729047	27729112	+	0	66	32.37	UTR		1	TU602	SpliceToPCG	ENSG00000102444	
CONGO_chr17_031975581_80_p2	chr17	31975581	31975660	+	2	80	89.7	UTR		1	TU603	SpliceToPCG	ENSG00000050955	
CONGO_chr17_032384498_110_m2	chr17	32384498	32384607	-	2	110	646.1	AntisenseCDS		1	TU604	NovelMultiExon		
CONGO_chr17_032410584_167_m0	chr17	32410584	32410750	-	0	167	129.82	AntisenseIntron						
CONGO_chr17_032424328_138_m0	chr17	32424328	32424465	-	0	138	29.64	AntisenseIntron						
CONGO_chr17_032534210_106_p1	chr17	32534210	32534315	-	1	106	93.11	AntisenseIntron						
CONGO_chr17_032540993_93_p0	chr17	32540993	32541085	+	0	93	1.52	AntisenseIntron						
CONGO_chr17_03261474_81_p1	chr17	3261474	32614854	+	1	81	80.62	AntisenseIntron						
CONGO_chr17_033618861_122_m2	chr17	33618861	33618982	+	2	122	193.3	NCExon		1	TU114	NovelMultiExon		
CONGO_chr17_033628903_78_m2	chr17	33628903	33628980	-	2	78	186.96	NCExon	1	Peptidase_M1	1	TU114	NovelMultiExon	
CONGO_chr17_033652033_85_m0	chr17	33652033	33652117	-	0	85	213.3	NCExon	1		1	TU114	NovelMultiExon	
CONGO_chr17_033955035_186_m0	chr17	33955035	33955220	-	0	186	341.88	Intron		1	TU605	SpliceToPCG	ENSG00000017373	
CONGO_chr17_034005846_63_m2	chr17	34005846	34005908	-	2	63	217.42	Intron						
CONGO_chr17_034335473_78_p0	chr17	34335473	34335550	+	0	78	248.26	NCExon		1	TU606	NovelMultiExon		
CONGO_chr17_034647610_83_p0	chr17	34647610	34647692	+	0	83	106.68	Intron		1	TU607	SpliceToPCG	ENSG00000223399	
CONGO_chr17_034972746_181_m1	chr17	34972746	34972926	-	1	181	18.48	Intergenic						
CONGO_chr17_035017615_157_p1	chr17	35017615	35017771	+	1	157	34.76	AntisenseUTR						
CONGO_chr17_035109664_137_m1	chr17	35109664	35109800	-	1	137	27.23	AntisenseUTR						
CONGO_chr17_035541016_62_m0	chr17	35541016	35541077	-	0	62	208.01	AntisenseUTR				NovelUnspliced		
CONGO_chr17_035722243_100_p1	chr17	35722243	35722349	+	1	100	35.72	Intron						
CONGO_chr17_036569744_196_m1	chr17	36569744	36569939	+	1	196	97.88	UTR	Keratin_B2	1				
CONGO_chr17_036622729_208_p2	chr17	36622729	36622936	+	2	208	293.82	Intron						
CONGO_chr17_037036568_50_m1	chr17	37036568	37036617	-	1	50	155.08	NCExon	1		TU223	SpliceToNCG	ENSG00000214514	
CONGO_chr17_037043650_99_m0	chr17	37043650	37043748	-	0	99	53.54	NCIntron						
CONGO_chr17_037422966_222_m2	chr17	37422966	37423187	-	2	222	91.55	UTR		1	TU609	SpliceToPCG	ENSG00000168259	
CONGO_chr17_037445249_771_m0	chr17	37445249	37446019	-	0	771	2854.17	Intron						
CONGO_chr17_037446047_687_m0	chr17	37446047	37446733	-	0	687	1428.35	Intron						
CONGO_chr17_037448378_190_m1	chr17	37448378	37448567	-	1	190	288.18	Intron						
CONGO_chr17_037448361_310_m0	chr17	37448361	37448920	-	0	310	766.92	Intron						
CONGO_chr17_037908338_75_p0	chr17	37908338	37908412	+	0	75	85.07	Intron						
CONGO_chr17_038277948_286_p2	chr17	38277948	38278328	+	2	286	779.8	Intergenic	Cu_amine_oxid	1	TU610	SpliceToPCG	ENSG00000033627	
CONGO_chr17_038278936_130_p1	chr17	38278936	38279065	+	1	130	347.49	Intergenic	Cu_amine_oxid	1	TU115	SpliceToPCG	ENSG00000131467	
CONGO_chr17_038279637_189_p0	chr17	38279637	38279825	+	0	189	629.9	Intergenic	Cu_amine_oxid	1	TU115	SpliceToPCG	ENSG00000131467	
CONGO_chr17_038736027_105_m0	chr17	38736027	38736131	-	0	105	423.66	NCExon		1	TU611	SpliceToNCG	ENSG00000236383	
CONGO_chr17_038975865_76_m0	chr17	38975865	38975940	-	0	76	19.76	Intron						
CONGO_chr17_039371122_170_p0	chr17	39371122	39371291	+	0	170	150.21	Intergenic			1	TU612	NovelMultiExon	
CONGO_chr17_039499410_134_p1	chr17	39499410	39499543	+	1	134	48.61	AntisenseCDS				UnsplicedMergeWithPCG	ENSG00000161654	
CONGO_chr17_039871408_202_m0	chr17	39871408	39871609	-	0	202	27.45	UTR		1	TU613	SpliceToPCG	ENSG00000186566	
CONGO_chr17_040113493_45_p0	chr17	40113493	40113537	-	0	45	25.83	AntisenseCDS			1			
CONGO_chr17_040843381_57_m0	chr17	40843381	40843437	-	0	57	64.55	UTR			1	TU615	SpliceToPCG	ENSG00000159314
CONGO_chr17_040948349_153_m2	chr17	40948349	40948501	-	2	153	69.02	UTR	1		1	TU616	SpliceToPCG	ENSG00000214420
CONGO_chr17_043414788_99_p0	chr17	43414788	43414886	+	0	99	20.16	Intergenic						
CONGO_chr17_043486618_163_p1	chr17	43486618	43486780	+	1	163	26.48	Intron			1	TU618	SpliceToPCG	ENSG00000082641
CONGO_chr17_043581913_102_p2	chr17	43581913	43582014	+	2	102	60.5	AntisenseIntron						
CONGO_chr17_043585696_130_m0	chr17	43585696	43585825	-	0	130	102.22	Intron						
CONGO_chr17_043593092_83_p2	chr17	43593092	43593174	+	2	83	23.16	AntisenseIntron						
CONGO_chr17_043610956_63_m2	chr17	43610956	43611018	-	2	63	80.51	Intron						
CONGO_chr17_043726700_118_m1	chr17	43726700	43726817	-	1	118	39.46	AntisenseNCExon						
CONGO_chr17_043987892_88_m1	chr17	43987892	43987979	-	1	88	16.13	UTR		1	TU619	SpliceToPCG	ENSG00000108511	
CONGO_chr17_044287194_75_p0	chr17	44287194	44287268	+	0	75	34.35	UTR			TU620	SpliceToPCG	ENSG00000136436	
CONGO_chr17_046183648_114_m0	chr17	46183648	46183761	-	0	114	39.86	AntisenseUTR		1	TU621	UnsplicedMergeWithPCG	ENSG00000108484	
CONGO_chr17_046500292_81_m0	chr17	46500292	46500372	-	0	81	277.65	Intron			1	TU621	SpliceToPCG	ENSG00000080294
CONGO_chr17_046618385_184_m0	chr17	46618385	46618568	-	0	184	103.73	Intron		1	TU622	SpliceToPCG	ENSG00000011258	
CONGO_chr17_050943419_276_m0	chr17	50943419	50943694	-	0	276	1635.86	Intergenic						
CONGO_chr17_051394422_110_p0	chr17	51394422	51394531	+	0	110	464.13	Intergenic						
CONGO_chr17_051397939_98_p1	chr17	51397939	51398036	+	1	98	262.11	Intergenic						
CONGO_chr17_051478599_61_p0	chr17	51478599	51478659	+	0	61	213.51	Intergenic						
CONGO_chr17_053437396_54_p1	chr17	53437396	53437449	+	1	54	36.92	AntisenseUTR		1	TU623	NovelMultiExon		
CONGO_chr17_054785465_33_m0	chr17	54785465	54785497	-	0	33	104.01	AntisenseUTR						
CONGO_chr17_056264448_96_m2	chr17	56264448	56264543	-	2	96	24.12	AntisenseIntron						
CONGO_chr17_056577258_164_m0	chr17	56577258	56577421	-	0	164	136.56	AntisenseIntron						
CONGO_chr17_056768309_118_m0	chr17	56768309	56768426	-	0	118	23.61	AntisenseIntron						
CONGO_chr17_056799096_104_m0	chr17	56799096	56799199	-	0	104	28.94	AntisenseIntron						
CONGO_chr17_057679149_148_m0	chr17	57679149	57679296	-	0	148	70.99	Intergenic	1		TU624	NovelMultiExon		
CONGO_chr17_057740518_54_p2	chr17	57740518	57740571	+	2	54	56.19	Intergenic						
CONGO_chr17_057741035_105_p2	chr17	57741035	57741139	+	2	105	162.78	Intergenic						
CONGO_chr17_057772215_105_p2	chr17	57772215	57772319	+	2	105	178	Intergenic						
CONGO_chr17_057772981_54_p2	chr17	57772981	57773034	+	2	54	61.64	Intergenic						
CONGO_chr17_057781764_105_p2	chr17	57781764	57781868	+	2	105	184.26	Intergenic						
CONGO_chr17_057800263_54_p2	chr17	57800263	57800316	+	2	54	151.72	Intergenic						
CONGO_chr17_058181296_254_m0	chr17	58181296	58181549	-	0	254	1496.45	Intron			TU225	SpliceToPCG	ENSG00000173838	
CONGO_chr17_058181559_71_m2	chr17	58181559	58181629	-	2	71	117.5	Intron			TU225	SpliceToPCG	ENSG00000173838	
CONGO_chr17_058530342_109_m0	chr17	58530342	58530450	-	0	109	299.85	Intron			1	TU226	SpliceToPCG	ENSG00000170921
CONGO_chr17_058582363_111_p2	chr17	58582363	58582473	-	2	111	256.63	Intron		1	TU226	SpliceToPCG	ENSG00000170921	
CONGO_chr17_059927361_95_p2	chr17	59927361	59927455	+	2	95	48.82	AntisenseIntron			1	TU116	UnsplicedMergeWithPCG	ENSG00000108654
CONGO_chr17_059927883_80_p2	chr17	59927883	59927962	+	2	80	145.04	AntisenseIntron			1	TU116	NovelMultiExon	
CONGO_chr17_059928347_163_p1	chr17	59928347	59928509	+	1	163	3.95	AntisenseIntron			1	TU116	NovelMultiExon	
CONGO_chr17_060176496_93_m0	chr17	60176496	60176588	-	0	93	27.63	UTR		1	TU668	SpliceToPCG	ENSG00000215769	
CONGO_chr17_060178387_150_m0	chr17	60178387	60178536	-	0	150	589.47	UTR		1	TU668	SpliceToPCG	ENSG00000215769	
CONGO_chr17_060188496_132_m1	chr17	60188496	60188627	-	1	132	25.95	UTR		1	TU668	SpliceToPCG	ENSG00000215769	
CONGO_chr17_060188935_356_m0	chr17	60188935	60189290	-	0	356	495.76	UTR		1	TU668	SpliceToPCG	ENSG00000215769	
CONGO_chr17_060380257_397_p0	chr17	60380257	60380653	+	0	397	1109.12	Intergenic			1			
CONGO_chr17_060394898_64_m0	chr17	60394898	60394961	-	0	64	34.78	NCExon		1	TU333	SpliceToNCG	ENSG00000214174	
CONGO_chr17_060395895_121_m1	chr17	60395895	60396015	-	1	121	402.53	NCIntron		1	TU333	SpliceToNCG	ENSG00000214174	
CONGO_chr17_060399114_129_m2	chr17	60399114	60399242	-	2	129	332.07	NCExon		1	TU333	SpliceToNCG	ENSG00000214174	
CONGO_chr17_060399334_155_m1	chr17	60399334	60399488	-										

CONGO_chr2_179225625_84_m2	chr2	179225625	179225708	-	2	84	259.74	Intron			1	TU77	NovelMultiExon	
CONGO_chr2_179225820_84_m2	chr2	179225820	179225903	-	2	84	109.73	Intron				TU77	NovelMultiExon	
CONGO_chr2_179226013_84_m2	chr2	179226013	179226096	-	2	84	78.88	Intron				TU77	NovelMultiExon	
CONGO_chr2_179248643_84_m2	chr2	179248643	179248726	-	2	84	57.55	Intron				TU77	NovelMultiExon	
CONGO_chr2_179250173_87_m2	chr2	179250173	179250259	-	2	87	205.49	NCExon			1	TU882	NovelMultiExon	
CONGO_chr2_179325146_122_m0	chr2	179325146	179325267	-	0	122	142.47	Intron				TU883	NovelMultiExon	
CONGO_chr2_181899893_69_p0	chr2	181899893	181899961	+	0	69	67.87	NCIntron						
CONGO_chr2_182247579_129_p0	chr2	182247579	182247707	+	0	129	77.03	AntisenseIntron						
CONGO_chr2_183576705_80_m0	chr2	183576705	183576784	-	0	80	15.48	Intron				TU884	SpliceToPCG	ENSG00000061676
CONGO_chr2_187709879_71_m0	chr2	187709879	187709949	-	0	71	26.72	AntisenseNCIntron						
CONGO_chr2_190234492_90_p0	chr2	190234492	190234581	+	0	90	271.35	UTR			1	TU78	SpliceToPCG	ENSG00000138381
CONGO_chr2_190236872_54_p0	chr2	190236872	190236925	+	0	54	296.55	UTR			1	TU78	SpliceToPCG	ENSG00000138381
CONGO_chr2_190238347_76_p0	chr2	190238347	190238422	+	0	76	465.28	UTR			1	TU78	SpliceToPCG	ENSG00000138381
CONGO_chr2_190239012_71_p2	chr2	190239012	190239081	+	2	71	60.48	Intron			1	TU78	SpliceToPCG	ENSG00000138381
CONGO_chr2_190633791_57_m0	chr2	190633791	190633847	+	0	57	96.28	Intron						
CONGO_chr2_197563649_248_m2	chr2	197563649	197563896	-	2	248	734.73	NCExon			1	TU885	SpliceToPCG	ENSG00000065413
CONGO_chr2_198358025_170_m2	chr2	198358025	198358194	-	2	170	82.52	UTR			1	TU886	SpliceToPCG	ENSG00000152430
CONGO_chr2_199236062_108_p0	chr2	199236062	199236169	-	0	108	24.4	AntisenseNCIntron						
CONGO_chr2_200030307_105_m2	chr2	200030307	200030411	-	2	105	87.41	UTR			1	TU887	SpliceToPCG	ENSG00000119042
CONGO_chr2_200392540_186_m2	chr2	200392540	200392725	-	2	186	175.68	NCExon			1	TU888	SpliceToPCG	ENSG00000226124
CONGO_chr2_200424932_65_p0	chr2	200424932	200424996	+	0	65	102	Intergenic						
CONGO_chr2_201433125_242_p2	chr2	201433125	201433366	+	2	242	43.98	AntisenseCDS			1			
CONGO_chr2_202775397_139_p1	chr2	202775397	202775535	+	1	139	228.14	Intergenic				TU889	SpliceToPCG	ENSG00000182329
CONGO_chr2_203670515_87_p0	chr2	203670515	203670601	-	0	87	207.15	NCExon				TU890	SpliceToPCG	ENSG00000144426
CONGO_chr2_204654859_87_m1	chr2	204654859	204654945	-	1	87	58.62	Intergenic						
CONGO_chr2_206255590_72_m1	chr2	206255590	206255661	-	1	72	66.01	AntisenseUTR					UnsplicedMergeWithPCG	ENSG00000118257
CONGO_chr2_206279169_87_m1	chr2	206279169	206279255	-	1	87	23.15	AntisenseIntron						
CONGO_chr2_206400859_93_m0	chr2	206400859	206400951	-	0	93	11.66	Intergenic						
CONGO_chr2_206465115_81_p2	chr2	206465115	206465195	+	2	81	14.25	Intergenic						
CONGO_chr2_207243456_75_m0	chr2	207243456	207243530	-	0	75	527.62	Intron						
CONGO_chr2_207686605_107_m2	chr2	207686605	207686711	-	2	107	11.36	Intron						
CONGO_chr2_208102784_204_p1	chr2	208102784	208102987	+	1	204	37.19	UTR			1	TU891	SpliceToPCG	ENSG00000118260
CONGO_chr2_208543541_92_p2	chr2	208543541	208543632	+	2	92	15.36	AntisenseIntron						
CONGO_chr2_210260688_243_p2	chr2	210260688	210260930	+	2	243	472.86	Intron						
CONGO_chr2_210467176_160_p0	chr2	210467176	210467335	-	0	160	287.68	Intron				TU893	SpliceToPCG	ENSG00000144406
CONGO_chr2_210543943_82_m1	chr2	210543943	210544023	-	1	82	37.33	AntisenseIntron						
CONGO_chr2_213720440_67_m0	chr2	213720440	213720506	-	0	67	109.99	UTR						
CONGO_chr2_215953801_141_p1	chr2	215953801	215953941	+	1	141	261.17	AntisenseCDS			1	TU894	NovelMultiExon	
CONGO_chr2_218583068_117_m0	chr2	218583068	218583184	-	0	117	146.38	Intergenic				TU895	SpliceToPCG	ENSG00000079308
CONGO_chr2_220019233_67_p1	chr2	220019233	220019299	+	1	67	28.03	Intron						
CONGO_chr2_221998896_55_p1	chr2	221998896	221998950	+	1	55	40.71	AntisenseUTR			1		UnsplicedMergeWithPCG	ENSG00000116106
CONGO_chr2_222841904_108_p0	chr2	222841904	222842011	-	0	108	6.03	AntisenseIntron						
CONGO_chr2_222891991_876_m0	chr2	222891991	222892866	-	0	876	1443.63	AntisenseNCExon	zf-C3HC4					
CONGO_chr2_225410272_46_m0	chr2	225410272	225410317	-	0	46	66.45	Intron						
CONGO_chr2_225555168_162_m0	chr2	225555168	225555329	-	0	162	721.06	NCExon			1	TU897	SpliceToPCG	ENSG00000135905
CONGO_chr2_227960643_267_m0	chr2	227960643	227960909	-	0	267	356.09	Intergenic						
CONGO_chr2_228171850_182_m0	chr2	228171850	228172031	-	0	182	415.13	Intergenic						
CONGO_chr2_228172038_178_p1	chr2	228172038	228172215	+	1	178	49.16	Intergenic						
CONGO_chr2_228181744_140_p0	chr2	228181744	228181883	-	0	140	11.2	Intergenic						
CONGO_chr2_228187498_303_m0	chr2	228187498	228187800	-	0	303	366.54	Intron						
CONGO_chr2_228190458_318_p0	chr2	228190458	228190775	-	0	318	502.69	AntisenseNCExon			1		UnsplicedMergeWithNCG	ENSG00000236116
CONGO_chr2_228221234_162_m0	chr2	228221234	228221395	-	0	162	387.7	Intergenic	Folate carrier			TU898	SpliceToPCG	ENSG00000135917
CONGO_chr2_228297400_318_m0	chr2	228297400	228297717	-	0	318	424.86	Intergenic						
CONGO_chr2_228301295_291_p0	chr2	228301295	228301585	+	0	291	474.53	Intergenic						
CONGO_chr2_228318854_327_p0	chr2	228318854	228319180	+	0	327	450.46	Intergenic						
CONGO_chr2_230711598_139_p2	chr2	230711598	230711736	+	2	139	468.22	NCIntron						
CONGO_chr2_231563652_421_p0	chr2	231563652	231564072	+	0	421	707.36	AntisenseNCIntron				TU899	SpliceToPCG	ENSG00000173699
CONGO_chr2_232064788_239_m1	chr2	232064788	232067026	+	1	239	3.62	Intergenic					NovelUnspliced	
CONGO_chr2_233093553_97_p0	chr2	233093553	233093649	+	0	97	25.52	Intergenic			1			
CONGO_chr2_233094275_108_p2	chr2	233094275	233094382	+	2	108	495.38	Intergenic			1			
CONGO_chr2_233094770_51_p2	chr2	233094770	233094820	+	2	51	212.61	Intergenic			1			
CONGO_chr2_233094925_190_p2	chr2	233094925	233095114	+	2	190	1116.33	Intergenic	Trypsin		1			
CONGO_chr2_233095469_100_p1	chr2	233095469	233095568	+	1	100	461.34	Intergenic			1			
CONGO_chr2_233095656_160_p0	chr2	233095656	233095815	+	0	160	739.93	Intergenic	Trypsin		1			
CONGO_chr2_233096014_143_p2	chr2	233096014	233096156	+	2	143	1105.28	Intergenic	Trypsin		1			
CONGO_chr2_233096370_163_p0	chr2	233096370	233096532	+	0	163	1224.14	Intergenic	Trypsin		1			
CONGO_chr2_238303162_67_p0	chr2	238303162	238303228	+	0	67	162.54	Intron				TU901	SpliceToPCG	ENSG00000124831
CONGO_chr2_241156812_393_p0	chr2	241156812	241157004	-	0	393	335.56	UTR			1	TU266	SpliceToPCG	ENSG00000142327
CONGO_chr2_241160513_141_p0	chr2	241160513	241160625	-	0	141	978.68	UTR	Peptidase_M1		1	TU266	SpliceToPCG	ENSG00000142327
CONGO_chr2_241494604_78_p0	chr2	241494604	241494681	+	0	78	93.38	Intergenic						
CONGO_chr2_00503362_119_m0	chr20	503362	503480	-	0	119	28.48	Intergenic						
CONGO_chr2_005520484_49_m2	chr20	5520484	5520532	-	2	49	31.87	Intron			1	TU783	SpliceToPCG	ENSG00000125772
CONGO_chr20_007445455_117_m0	chr20	7445455	7445571	-	0	117	59.72	Intergenic						
CONGO_chr20_10224831_75_p1	chr20	10224831	10224905	+	1	75	38.49	NCExon			1		UnsplicedMergeWithPCG	ENSG00000132639
CONGO_chr20_10342417_119_m0	chr20	10342417	10342535	-	0	119	246.35	UTR			1	TU784	SpliceToPCG	ENSG00000125863
CONGO_chr20_13444987_87_m0	chr20	13444987	13445075	-	0	87	84.93	Intron						
CONGO_chr20_15821572_84_p2	chr20	15821572	15821655	+	2	84	327.57	AntisenseUTR						
CONGO_chr20_16576582_62_m2	chr20	16576582	16576643	-	2	62	13.07	Intergenic						
CONGO_chr20_20421347_105_p0	chr20	20421347	20421451	-	0	105	83.33	AntisenseIntron						
CONGO_chr20_20527323_141_m2	chr20	20527323	20527463	-	2	141	659.14	NCExon			1	TU785	SpliceToPCG	ENSG00000188559
CONGO_chr20_21019080_126_p0	chr20	21019080	21019205	-	0	126	9.39	Intergenic			1			
CONGO_chr20_29472994_149_m2	chr20	29472994	29473142	-	2	149	10.25	NCExon	Defensin_beta		1			
CONGO_chr20_31257288_64_p0	chr20	31257288	31257351	-	0	64	111.47	NCExon						
CONGO_chr20_33328931_219_m0	chr20	33328931	33329149	-	0	219	404.58	NCExon			1	TU787	SpliceToPCG	ENSG00000088298
CONGO_chr20_37768313_82_m0	chr20	37768313	37768394	-	0	82	23.89	Intergenic						
CONGO_chr20_38774627_62_m2	chr20	38774627	38774688	-	2	62	122.72	Intergenic						
CONGO_chr20_40144036_55_p0	chr20	40144036	40144090	-	0	55	387.2	AntisenseCDS			1	TU789	NovelMultiExon	
CONGO_chr20_43200573_77_m2	chr20	43200573	43200649	-	2	77	182.22	Intergenic				TU248	NovelMultiExon	
CONGO_chr20_43200666_64_m2	chr20	43200666	43200729	-	2	64	120.29	Intergenic				TU248	NovelMultiExon	
CONGO_chr20_43996754_97_p0	chr20	43996754	43996850	+	0	97	51.17	AntisenseCDS			1	TU791	SpliceToPCG	ENSG00000100982
CONGO_chr20_50203254_134_p0	chr20	50203254	50203387	+	0	134	399.46	AntisenseCDS					UnsplicedMergeWithPCG	ENSG00000020256

CONGO_chr3_152984310_188_p0	chr3	152984310	152984497	+	0	188	227.44	NCExon			1	TU283	NovelMultiExon	
CONGO_chr3_152984614_138_p1	chr3	152984614	152984751	+	1	138	56.39	NCExon			1	1		
CONGO_chr3_153499884_55_p1	chr3	153499884	153499938	+	1	55	3.28	UTR			1	TU947	SpliceToPCG	ENSG00000152601
CONGO_chr3_156491427_95_m2	chr3	156491427	156491521	-	2	95	37.39	NCExon			1	TU948	SpliceToNCG	OTTHUMG00000158471
CONGO_chr3_156904629_79_m0	chr3	156904629	156904707	-	0	79	301.74	UTR			1	TU949	SpliceToPCG	OTTHUMG00000158477
CONGO_chr3_159632443_70_p0	chr3	159632443	159632512	+	0	70	96.03	Intron						
CONGO_chr3_159632604_157_m2	chr3	159632604	159632760	-	2	157	128.54	AntisenseIntron						
CONGO_chr3_160041634_184_m1	chr3	160041634	160041817	-	1	184	435.36	Intergenic						
CONGO_chr3_160041906_383_m0	chr3	160041906	160042288	-	0	383	618.53	Intergenic	7m_1					
CONGO_chr3_160932818_80_p2	chr3	160932818	160932897	+	2	80	49.22	Intron			1	TU950	SpliceToPCG	ENSG00000151967
CONGO_chr3_169113268_119_p2	chr3	169113268	169113386	+	2	119	223.65	NCExon			1	TU136	SpliceToNCG	OTTHUMG00000158502
CONGO_chr3_169115543_143_p0	chr3	169115543	169115685	+	0	143	73.71	NCExon			1	TU136	SpliceToNCG	OTTHUMG00000158503
CONGO_chr3_169117781_198_p1	chr3	169117781	169117978	+	1	198	437.94	NCIntron						
CONGO_chr3_169413563_57_m1	chr3	169413563	169413619	-	1	57	3.41	Intergenic						
CONGO_chr3_170316314_104_p1	chr3	170316314	170316417	+	1	104	513.95	AntisenseCDS						
CONGO_chr3_170316802_113_p2	chr3	170316802	170316914	+	2	113	432.05	AntisenseCDS						
CONGO_chr3_170442969_101_p1	chr3	170442969	170443069	+	1	101	2.81	AntisenseIntron					UnsplicedMergeWithPCG	OTTHUMG00000158596
CONGO_chr3_170565195_63_p0	chr3	170565195	170565257	+	0	63	87.64	AntisenseIntron						
CONGO_chr3_170600778_66_m0	chr3	170600778	170600843	-	0	66	16.89	Intron						
CONGO_chr3_170600869_81_m0	chr3	170600869	170600949	-	0	81	0.09	Intron						
CONGO_chr3_170676968_98_p2	chr3	170676968	170677065	+	2	98	23.61	AntisenseIntron					NovelUnspliced	
CONGO_chr3_170678512_105_m0	chr3	170678512	170678616	-	0	105	89.54	Intron						
CONGO_chr3_170775007_116_m1	chr3	170775007	170775122	-	1	116	116.81	Intron						
CONGO_chr3_170782017_74_p1	chr3	170782017	170782090	+	1	74	134.18	AntisenseIntron						
CONGO_chr3_174805774_75_m0	chr3	174805774	174805848	-	0	75	5.45	AntisenseIntron						
CONGO_chr3_181944661_265_m0	chr3	181944661	181944925	-	0	265	3	Intron					NovelUnspliced	
CONGO_chr3_182375145_140_m2	chr3	182375145	182375284	-	2	140	90.42	AntisenseNCIntron						
CONGO_chr3_182375431_58_m2	chr3	182375431	182375488	-	2	58	18.14	AntisenseNCIntron						
CONGO_chr3_182927354_224_p0	chr3	182927354	182927577	+	0	224	61.04	NCIntron						
CONGO_chr3_182952927_68_p1	chr3	182952927	182952994	+	1	68	46.01	NCIntron						
CONGO_chr3_183224362_106_m1	chr3	183224362	183224467	+	1	106	39.93	Intergenic						
CONGO_chr3_183461555_80_m1	chr3	183461555	183461634	-	1	80	43.6	Intergenic						
CONGO_chr3_183472039_171_p0	chr3	183472039	183472209	+	0	171	25.16	Intergenic						
CONGO_chr3_184333299_140_m2	chr3	184333299	184333438	-	2	140	122.09	Intron	Lamp		TU953	SpliceToPCG	OTTHUMG00000158355	
CONGO_chr3_185700227_245_m0	chr3	185700227	185700471	-	0	245	486.83	AntisenseIntron						
CONGO_chr3_185700502_135_m0	chr3	185700502	185700636	-	0	135	45.2	AntisenseIntron						
CONGO_chr3_185700779_74_m0	chr3	185700779	185700852	-	0	74	9.4	AntisenseIntron					NovelUnspliced	
CONGO_chr3_188922291_73_m0	chr3	188922291	188922363	-	0	73	56.13	UTR			1	TU956	SpliceToPCG	OTTHUMG00000156441
CONGO_chr3_189552868_62_p0	chr3	189552868	189552929	+	0	62	30.42	Intron						
CONGO_chr3_197119963_495_m0	chr3	197119963	197120457	-	0	495	1488.68	UTR			1	TU957	SpliceToPCG	OTTHUMG00000155737
CONGO_chr4_000558694_1158_m0	chr4	558694	559581	-	0	1158	1870.43	Intergenic					NovelUnspliced	
CONGO_chr4_001386720_884_m2	chr4	1386720	1387603	-	2	884	2104.43	Intergenic	Homeobox		1			
CONGO_chr4_001389768_463_m0	chr4	1389768	1390230	-	0	463	937.92	Intergenic						
CONGO_chr4_001913923_66_m0	chr4	1913923	1913988	-	0	66	79.54	AntisenseCDS			1		UnsplicedMergeWithPCG	OTTHUMG00000121147
CONGO_chr4_002416729_105_p0	chr4	2416729	2416833	+	0	105	725.92	UTR			1	TU17	SpliceToPCG	ENSG00000206113
CONGO_chr4_002422472_72_p0	chr4	2422472	2422543	+	0	72	263.02	Intergenic			1	TU17	SpliceToPCG	ENSG00000206113
CONGO_chr4_002422789_89_p0	chr4	2422789	2422877	+	0	89	446.13	Intergenic			1	TU17	SpliceToPCG	ENSG00000206113
CONGO_chr4_002423667_205_p1	chr4	2423667	2423871	+	1	205	1414.57	Intergenic			1	TU17	SpliceToPCG	ENSG00000206113
CONGO_chr4_002430248_142_p0	chr4	2430248	2430389	+	0	142	823.73	Intergenic			1	TU17	SpliceToPCG	ENSG00000206113
CONGO_chr4_002430632_152_p2	chr4	2430632	2430783	+	2	152	670.57	Intergenic			1	TU17	SpliceToPCG	ENSG00000206113
CONGO_chr4_002431562_206_p0	chr4	2431562	2431767	+	0	206	2709.21	Intergenic			1	TU17	SpliceToPCG	ENSG00000206113
CONGO_chr4_002433968_163_p1	chr4	2433968	2434130	+	1	163	855.92	Intergenic			1	TU17	SpliceToPCG	ENSG00000206113
CONGO_chr4_002567609_246_p0	chr4	2567609	2567854	+	0	246	1995.85	UTR			1	TU137	SpliceToPCG	ENSG00000125386
CONGO_chr4_002596787_134_p0	chr4	2596787	2596920	+	0	134	1465.04	UTR			1	TU137	SpliceToPCG	ENSG00000125386
CONGO_chr4_002597935_168_p1	chr4	2597935	2598102	+	1	168	2002.55	UTR			1	TU137	SpliceToPCG	ENSG00000125386
CONGO_chr4_003379014_85_p0	chr4	3379014	3379098	-	0	85	27.07	Intron			1	TU958	SpliceToPCG	ENSG00000125386
CONGO_chr4_004574583_122_p2	chr4	4574583	4574704	-	2	122	21.25	AntisenseIntron						
CONGO_chr4_006041332_494_m2	chr4	6041332	6041825	-	2	494	1093.29	NCExon			1		UnsplicedMergeWithNCG	OTTHUMG00000125490
CONGO_chr4_006060624_537_m0	chr4	6060624	6061160	-	0	537	942.73	Intergenic						
CONGO_chr4_006064458_139_m1	chr4	6064458	6064596	-	1	139	612.73	Intergenic						
CONGO_chr4_006070773_287_m0	chr4	6070773	6071059	-	0	287	1292.75	Intergenic						
CONGO_chr4_006122685_106_m0	chr4	6122685	6122790	-	0	106	616.15	Intron						
CONGO_chr4_008063057_120_m2	chr4	8063057	8063176	-	2	120	187.25	Intron			1	TU959	SpliceToPCG	ENSG00000163995
CONGO_chr4_008493450_96_p0	chr4	8493450	8493545	+	0	96	263.27	UTR			1	TU284	SpliceToPCG	OTTHUMG00000128484
CONGO_chr4_008493783_286_p0	chr4	8493783	8494068	+	0	286	372.38	UTR			1	TU284	SpliceToPCG	ENSG00000205959
CONGO_chr4_014613795_209_m2	chr4	14613795	14614003	-	2	209	23.47	Intergenic					NovelUnspliced	
CONGO_chr4_015807379_32_m2	chr4	15807379	15807410	-	2	32	5.77	Intron			1	TU960	SpliceToPCG	ENSG00000169762
CONGO_chr4_017485898_120_m0	chr4	17485898	17486010	-	0	120	455.85	Intron					UnsplicedMergeWithPCG	ENSG00000215273
CONGO_chr4_017486231_253_m1	chr4	17486231	17486483	-	1	253	745.33	Intron					UnsplicedMergeWithPCG	ENSG00000215273
CONGO_chr4_017486750_88_m1	chr4	17486750	17486837	-	1	88	234.64	Intron						
CONGO_chr4_017487100_299_m1	chr4	17487100	17487398	-	1	299	189.54	Intron					NovelUnspliced	
CONGO_chr4_017487805_318_m2	chr4	17487805	17488122	-	2	318	668.68	Intron					NovelUnspliced	
CONGO_chr4_020091095_78_m1	chr4	20091095	20091172	-	1	78	45.74	AntisenseIntron						
CONGO_chr4_023846359_116_p0	chr4	23846359	23846474	-	0	116	46	Intergenic						
CONGO_chr4_024083096_90_m0	chr4	24083096	24083185	-	0	90	7.1	Intergenic			1		NovelUnspliced	
CONGO_chr4_024147975_99_m0	chr4	24147975	24148073	-	0	99	163.48	UTR			1	TU961	SpliceToPCG	ENSG00000109606
CONGO_chr4_024341118_63_p2	chr4	24341118	24341180	-	2	63	21.54	Intergenic						
CONGO_chr4_025774303_32_p0	chr4	25774303	25774334	-	0	32	95.2	Intergenic			1	TU962	SpliceToPCG	OTTHUMG00000097793
CONGO_chr4_030397447_70_p0	chr4	30397447	30397516	-	0	70	1.02	Intron					SpliceToPCG	ENSG00000169851
CONGO_chr4_030560840_89_p2	chr4	30560840	30560928	-	2	89	116.14	Intron			1	TU964	SpliceToPCG	ENSG00000169851
CONGO_chr4_035959776_208_p0	chr4	35959776	35959983	-	0	208	501.72	UTR			1		UnsplicedMergeWithPCG	ENSG00000197057
CONGO_chr4_038351775_81_m0	chr4	38351775	38351855	-	0	81	2.73	AntisenseIntron						
CONGO_chr4_041444983_107_m0	chr4	41444983	41445089	-	0	107	57.78	Intron						
CONGO_chr4_041554195_143_m0	chr4	41554195	41554337	-	0	143	15.92	Intergenic						
CONGO_chr4_041554379_152_p2	chr4	41554379	41554530	-	2	152	61.53	Intergenic						
CONGO_chr4_041569824_152_m2	chr4	41569824	41569975	-	2	152	19.76	Intergenic				TU965	NovelMultiExon	
CONGO_chr4_046643817_83_p2	chr4	46643817	46643899	+	2	83	23.26	AntisenseIntron						
CONGO_chr4_048091319_109_p0	chr4	48091319	48091427	+	0	109	149.94	Intron			1	TU966	SpliceToPCG	ENSG00000109171
CONGO_chr4_048295603_45_m0	chr4	48295603	48295647	-	0	45	175.21	Intron				TU967	SpliceToPCG	ENSG00000075539
CONGO_chr4_052555772_67_p2	chr4	52555772	52555838	-	2	67	217.43	AntisenseCDS					UnsplicedMergeWithPCG	ENSG00000188993
CONGO_chr4_056380950_121_p0	chr4	56380950	56381070	+	0	121	641.96	UTR			1	TU968	Nov	

CONGO_chr7_150894633_86_m0	chr7	150894633	150894718	-	0	86	77.11	Intron				TU1140	SpliceToPCG	ENSG00000106617	
CONGO_chr7_152739285_450_m0	chr7	152739285	152739734	-	0	450	294.36	NCExon	1	rve					
CONGO_chr7_153512591_53_p2	chr7	153512591	153512643	+	2	53	47.28	Intron							
CONGO_chr7_155031287_113_m2	chr7	155031287	155031399	-	2	113	452.98	Intergenic							
CONGO_chr7_156856219_105_p2	chr7	156856219	156856323	+	2	105	140.46	NCExon				1			
CONGO_chr7_156857602_174_p2	chr7	156857602	156857775	+	2	174	724.21	NCExon					UnsplicedMergeWithPCG	ENSG00000105993	
CONGO_chr7_158242784_102_m0	chr7	158242784	158242885	-	0	102	39.02	Intron				TU1141	SpliceToPCG	ENSG00000117868	
CONGO_chr8_001196258_33_p2	chr8	1196258	1196290	+	2	33	312.86	Intergenic				1	TU318	SpliceToPCG	ENSG000001198010
CONGO_chr8_001436939_66_p2	chr8	1436939	1437004	+	2	66	347.56	UTR				1	TU318	SpliceToPCG	ENSG000001198010
CONGO_chr8_008614331_120_m2	chr8	8614331	8614450	-	2	120	56.74	Intergenic							
CONGO_chr8_009510405_66_p0	chr8	9510405	9510470	+	0	66	73.87	Intron							
CONGO_chr8_010369779_198_m1	chr8	10369779	10369976	-	1	198	487.16	Intergenic		Trypsin					
CONGO_chr8_010391281_106_m0	chr8	10391281	10391386	-	0	106	604.47	NCIntron							
CONGO_chr8_010391555_123_m0	chr8	10391555	10391677	-	0	123	368.63	NCIntron							
CONGO_chr8_010392632_260_m1	chr8	10392632	10392891	-	1	260	1162.39	NCExon							
CONGO_chr8_010393536_193_m2	chr8	10393536	10393728	-	2	193	1205.48	NCExon							
CONGO_chr8_010505352_97_p1	chr8	10505352	10505448	+	1	97	516.64	AntisenseCDS							
CONGO_chr8_010506112_73_p0	chr8	10506112	10506184	-	0	73	39.78	AntisenseCDS							
CONGO_chr8_011257814_76_m2	chr8	11257814	11257889	-	2	76	357.84	AntisenseCDS							
CONGO_chr8_011260502_140_p2	chr8	11260502	11260641	+	2	140	995.78	UTR				1		UnsplicedMergeWithPCG	ENSG00000154316
CONGO_chr8_011614151_40_p2	chr8	11614151	11614190	+	2	40	62.4							NovelUnspliced	
CONGO_chr8_018985996_92_m2	chr8	18985996	18986087	-	2	92	174.51	Intergenic				1			
CONGO_chr8_019750953_78_p1	chr8	19750953	19751030	+	1	78	227.55	Intron				1	TU1142	SpliceToPCG	ENSG00000104613
CONGO_chr8_021902180_111_p0	chr8	21902180	21902290	+	0	111	99.06	Intron				1	TU1143	SpliceToPCG	ENSG00000130227
CONGO_chr8_022752260_122_p2	chr8	22752260	22752381	+	2	122	29.28	AntisenseIntron							
CONGO_chr8_024266319_63_p0	chr8	24266319	24266381	+	0	63	316.83	Intron						NovelUnspliced	
CONGO_chr8_025497994_80_m2	chr8	25497994	25498073	-	2	80	52.32	AntisenseIntron							
CONGO_chr8_025559110_91_m1	chr8	25559110	25559200	-	1	91	104.07	AntisenseIntron							
CONGO_chr8_025845963_186_m0	chr8	25845963	25846148	-	0	186	86.35	Intron							
CONGO_chr8_025851506_150_p0	chr8	25851506	25851655	+	0	150	9.32	Intron							
CONGO_chr8_026205141_44_p0	chr8	26205141	26205184	+	0	44	57.02	UTR				1	TU1144	SpliceToPCG	ENSG00000221914
CONGO_chr8_027152660_54_m1	chr8	27152660	27152713	-	1	54	60.91	Intron				1	TU1145	SpliceToPCG	ENSG00000015592
CONGO_chr8_030533717_141_p0	chr8	30533717	30533857	+	0	141	21.37	Intron				1	TU1147	SpliceToPCG	ENSG00000157110
CONGO_chr8_030622369_288_p0	chr8	30622369	30622656	+	0	288	1986.82	AntisenseIntron				1	TU1148	NovelMultiExon	
CONGO_chr8_031008801_152_p0	chr8	31008801	31008952	+	0	152	120.78	AntisenseCDS							
CONGO_chr8_031163631_127_p0	chr8	31163631	31164489	+	0	127	136.99	Intergenic							
CONGO_chr8_032624263_147_p0	chr8	32624263	32624409	+	0	147	458.87	UTR							
CONGO_chr8_037050370_80_m2	chr8	37050370	37050449	-	2	80	32.8	Intergenic				1		UnsplicedMergeWithPCG	ENSG00000157168
CONGO_chr8_03724523_84_p0	chr8	3724523	3724606	+	0	84	41.08	Intergenic							
CONGO_chr8_037302641_173_m2	chr8	37302641	37302813	-	2	173	1.21	Intergenic						NovelUnspliced	
CONGO_chr8_037306477_132_p0	chr8	37306477	37306608	+	0	132	10.81	Intergenic							
CONGO_chr8_037456127_162_p0	chr8	37456127	37456288	+	0	162	21.37	Intergenic							
CONGO_chr8_037542570_148_m1	chr8	37542570	37542717	-	1	148	113.33	Intergenic							
CONGO_chr8_037614885_124_p1	chr8	37614885	37615008	+	1	124	71.69	Intergenic							
CONGO_chr8_039513190_167_p0	chr8	39513190	39513356	+	0	167	178.17	Intergenic							
CONGO_chr8_041286119_164_m1	chr8	41286119	41286282	-	1	164	29.06	UTR				1	TU1149	SpliceToPCG	ENSG00000104332
CONGO_chr8_041744017_77_m2	chr8	41744017	41744093	-	2	77	10.25	Intron							
CONGO_chr8_041744310_102_m2	chr8	41744310	41744411	-	2	102	9.78	Intron							
CONGO_chr8_049067001_310_m1	chr8	49067001	49067310	-	1	310	89.86	Intergenic							
CONGO_chr8_049067385_281_m2	chr8	49067385	49067665	-	2	281	473.33	Intergenic		rve					
CONGO_chr8_053300789_105_m0	chr8	53300789	53300893	-	0	105	37	Intron							
CONGO_chr8_053324471_173_m0	chr8	53324471	53324643	-	0	173	91.06	Intron							
CONGO_chr8_053329363_114_m2	chr8	53329363	53329476	-	2	114	36.55	UTR				1			
CONGO_chr8_053329654_120_p0	chr8	53329654	53329773	+	0	120	0.15	AntisenseIntron							
CONGO_chr8_053330042_106_p1	chr8	53330042	53330147	+	1	106	93.77	AntisenseIntron							
CONGO_chr8_053484476_90_m0	chr8	53484476	53484565	-	0	90	67.87	UTR				1	TU1150	SpliceToPCG	ENSG00000147488
CONGO_chr8_053404323_75_m0	chr8	53404323	53404397	-	0	75	399.15	UTR				1		UnsplicedMergeWithPCG	ENSG00000082556
CONGO_chr8_055253685_68_m0	chr8	55253685	55253752	-	0	68	38.12	Intergenic							
CONGO_chr8_055858661_164_p2	chr8	55858661	55858824	+	2	164	808.67	Intergenic				1	TU320	NovelMultiExon	
CONGO_chr8_055912564_220_p0	chr8	55912564	55912783	+	0	220	261.26	Intergenic				1	TU320	NovelMultiExon	
CONGO_chr8_055927730_113_p2	chr8	55927730	55927842	+	2	113	308.47	Intergenic				1	TU83	NovelMultiExon	
CONGO_chr8_055932142_79_p0	chr8	55932142	55932200	+	0	79	214.52	Intergenic				1	TU83	NovelMultiExon	
CONGO_chr8_055940949_82_p2	chr8	55940949	55941030	+	2	82	244.63	Intergenic				1	TU83	NovelMultiExon	
CONGO_chr8_055944957_56_p1	chr8	55944957	55945012	+	1	56	226.4	Intergenic				1	TU83	NovelMultiExon	
CONGO_chr8_057237437_133_m1	chr8	57237437	57237569	-	1	133	63.98	UTR				1		UnsplicedMergeWithPCG	ENSG00000181690
CONGO_chr8_061867981_292_m2	chr8	61867981	61868272	-	2	292	116.12	AntisenseIntron							
CONGO_chr8_062092023_116_m2	chr8	62092023	62092138	-	2	116	84.2	Intergenic							
CONGO_chr8_064029015_71_p2	chr8	64029015	64029085	+	2	71	112.7	Intron							
CONGO_chr8_065342115_109_p0	chr8	65342115	65342223	+	0	109	73.16	Intergenic					TU1151	NovelMultiExon	
CONGO_chr8_065345457_121_p1	chr8	65345457	65345577	+	1	121	130.95	Intergenic							
CONGO_chr8_065680628_94_p1	chr8	65680628	65680721	+	1	94	3.11	AntisenseIntron							
CONGO_chr8_066220109_150_m0	chr8	66220109	66220258	-	0	150	9.74	Intergenic							
CONGO_chr8_066263134_160_p1	chr8	66263134	66263293	+	1	160	59.21	Intergenic							
CONGO_chr8_068022840_114_m0	chr8	68022840	68022953	-	0	114	295.9	Intergenic				1		NovelUnspliced	
CONGO_chr8_069993288_60_m0	chr8	69993288	69993347	-	0	60	15.47	NCIntron							
CONGO_chr8_073195300_290_p2	chr8	73195300	73195589	+	2	290	41.07	Intergenic							
CONGO_chr8_076728512_69_p2	chr8	76728512	76728580	+	2	69	78.46	Intergenic		Gag_p30					
CONGO_chr8_076912656_46_p0	chr8	76912656	76912701	+	0	46	20	Intergenic							
CONGO_chr8_077603344_78_p0	chr8	77603344	77603421	+	0	78	20.73	Intergenic							
CONGO_chr8_077654112_81_p1	chr8	77654112	77654192	+	1	81	40.7	Intergenic							
CONGO_chr8_07765496_176_p2	chr8	7765496	77656671	+	2	176	31.87	Intergenic							
CONGO_chr8_077745032_36_p1	chr8	77745032	77745067	+	1	36	26.99	AntisenseNCIntron							
CONGO_chr8_077747617_148_m1	chr8	77747617	77747764	-	1	148	155	NCExon				1	TU1152	NovelMultiExon	
CONGO_chr8_077758530_216_p1	chr8	77758530	77758745	+	1	216	93.6	Intron							
CONGO_chr8_077759816_93_p0	chr8	77759816	77759908	-	0	93	69.24	Intron							
CONGO_chr8_077761991_74_m0	chr8	77761991	77762064	-	0	74	116.91	AntisenseIntron							
CONGO_chr8_077773689_51_p2	chr8	77773689	77773739	+	2	51	15.66	Intron							
CONGO_chr8_077778818_211_m1	chr8	77778818	77779028	-	1	211	244.19	AntisenseCDS							
CONGO_chr8_077876958_180_p1	chr8	77876958	77877137	+	1	180	24.59	Intron							
CONGO_chr8_077940137_85_p2	chr8	77940137	77940221	+	2	85	35.33	UTR				1	TU1153	SpliceToPCG	ENSG00000091656
CONGO_chr8_078984487_143_m2	chr8	78984487	78984629	-	2	143	28.08	Intergenic							
CONGO_chr8_079721867_168_p0	chr8	79721867	79722034	+	0	168	293.66	Intergenic							
CONGO_chr8_081948334_469_m1	chr8	81948334	81948802	-	1	469	521.59	Intron							

Table S6. Basic statistics on motif instances

Confidence	# of motifs reaching confidence	Total # of instances	# of examined bases covered	# of TFs with a motif reaching confidence	Total # of instances (best motif per TF)
0.0	630	55,021,406	80.6	335	35,366,716
0.1	540	15,817,545	45.3	294	11,181,918
0.2	492	8,385,913	26.0	270	6,068,955
0.3	435	4,697,272	14.3	252	3,495,271
0.4	375	2,675,802	7.7	225	2,050,302
0.5	293	1,449,752	3.9	188	1,175,237
0.6	216	707,141	1.7	151	595,984
0.7	129	269,944	0.6	101	240,849
0.8	56	90,464	0.2	45	80,138
0.9	16	33,822	0.1	14	29,080

Table S7. Listing of data sets and motifs used in analysis

Factor	Cell type	Technology	Num peaks	Motif used	Citation
CTCF	CD4+ T cells Embryonic stem (mouse)	Sequencing	21,544 8,546 (+mouse)	Jaspar MA0139.1	17512414 bern, unpub (mouse)
ER	MCF-7 breast cancer	Paired-end Tags	1,229	Transfac M00191	17542648
Fos	K562 CML	Sequencing	18,963	Transfac M00926	20139302
FOXA2	Liver	Promoter array	143 19 (+mouse)	Jaspar MA0047.2	17529977
HNF1	Liver	Promoter array	246 23 (+mouse)	Jaspar MA0046.1	17529977
HNF4	Liver	Promoter array	1,231 99 (+mouse)	Transfac M01036	17529977
HNF6	Liver	Promoter array	149 20 (+mouse)	Transfac M00639	17529977
Myc	K562 CML Embryonic stem (mouse)	Sequencing Promoter array (mouse)	15,749 2,399 (+mouse)	Transfac M00187	20139302 18358816 (mouse)
NF-κB	GM12878 B-Lymphocyte	Sequencing	38,559	Jaspar MA0061.1	20299548
NRSF	Jurkat T cell line	Sequencing	1,931	Transfac M00325	17540862
p53	HCT116 colon cancer	Paired-end Tags	62,939	Transfac M00034	16413492
STAT1	HeLa S3 cells	Sequencing	41,530	Transfac M00224	17558387
YY1	NT2/D1	Sequencing	11,018	Transfac M00651	Encode, unpub

Table S8. Coincidence of GWAS results with TSS-distal noncoding conserved elements

		Noncoding TSS-distal genome	Noncoding TSS-distal Hapmap CEU SNPs
All callable positions	<i>n</i>	1996107874	3464198
Conserved by SiPhy- ω	<i>n</i>	88530563	138260
GWAS hit in at least one study	<i>n</i>	3402	3390
Conserved GWAS	<i>n</i>	187	186
	fold enrichment	1.24	1.37
	<i>p</i>	1.6×10^{-3}	9.8×10^{-6}

Table S9. Summary of sitewise selective pressures

Data set	N sites	dN/dS <0.5	dN/dS <1	dN/dS >1	dN/dS >1.5	Domain instances	Domain types	Genes	Positive domain instances	Positive domain types	Positive genes	Positive sites	Fraction positive sites
Mammals	6,050,046	84%	94%	6%	2%	26,667	3,117	12,871	1,451	871	4,431	15,381	0.25423%
Primates	6,386,902	86%	91%	9%	6%	27,441	3,164	12,968	7	7	17	19	0.00030%
Glires	6,335,182	88%	95%	5%	2%	27,671	3,168	13,068	8	6	23	27	0.00043%
Laurasiatheria	6,313,057	85%	94%	6%	3%	27,120	3,146	12,860	89	77	369	468	0.00741%

Table S10. Simulation results for the power of site-wise analysis using three different mammalian trees

Tree	Indel rate	Spearman's rank correlation between inferred/true dN/dS	TPR at FDR <0.1	TP at FDR <0.1	TP at FDR <0.05	TPR at FDR <0.05	FPR at SLR default threshold	TP at SLR default threshold	FP at SLR default threshold	FDR at SLR default threshold
29 Eutherian	0	0.942	0.66400	3760	2990	0.528	0.00137	2900	0.00137	0.0423
9 high-coverage	0	0.849	0.13700	782	429	0.075	0.00248	1419	0.00248	0.1400
HMRD	0	0.749	0.00942	50	10	0.002	0.00538	802	0.00538	0.3860
29 Eutherian	0.05	0.831	0.43800	2433	1686	0.304	0.00176	2006	0.00176	0.0764
9 high-coverage	0.05	0.696	0.01760	98	86	0.015	0.00340	938	0.00340	0.2530
HMRD	0.05	0.570	0.00000	NA	NA	NA	0.00464	412	0.00464	0.5160

Table S11. Summaries of site-wise data in regions of high and low recombination rate, GC content, and evidence for non-neutral evolution

Variable	Quantile range	Low value	High value	Number of sites	Mean dN/dS	Mean signed LRT	Mean male recombination	Mean female recombination	Mean sex-averaged recombination	Fraction of sites under positive selection	GC content (10-kb window)
LRT Statistic											
signed_lrt	0-0.01	-188.187	-44.199	57,981	0.006	-50.660	1.522	1.812	1.681	0.00000	0.512
signed_lrt	0.01-0.25	-44.199	-19.594	1,391,549	0.015	-26.806	1.128	1.734	1.445	0.00000	0.475
signed_lrt	0.25-0.5	-19.594	-12.030	1,449,530	0.040	-15.607	0.971	1.657	1.333	0.00000	0.455
signed_lrt	0.5-0.75	-12.030	-4.926	1,449,521	0.128	-8.575	0.929	1.613	1.294	0.00000	0.451
signed_lrt	0.75-0.99	-4.926	3.648	1,391,529	0.694	-1.505	0.921	1.619	1.293	0.01338	0.454
signed_lrt	0.99-1	3.648	108.850	57,970	2.984	8.208	0.879	1.614	1.272	1.00000	0.454
Male Recombination											
recomb_m	0-0.01	-0.010	0.000	743,415	0.265	-12.265	0.000	0.991	0.648	0.01539	0.468
recomb_m	0.01-0.25	0.000	0.214	707,759	0.249	-12.617	0.102	1.192	0.647	0.01364	0.450
recomb_m	0.25-0.5	0.214	0.591	1,448,727	0.243	-12.865	0.414	1.469	0.942	0.01375	0.446
recomb_m	0.5-0.75	0.591	1.275	1,448,935	0.236	-13.168	0.882	1.809	1.345	0.01278	0.447
recomb_m	0.75-0.99	1.275	6.107	1,393,462	0.233	-14.509	2.402	2.288	2.345	0.01187	0.483
recomb_m	0.99-1	6.107	10.372	55,782	0.206	-16.616	8.010	1.562	4.787	0.00932	0.540
GC Content											
gc	0-0.01	0.071	0.332	58,204	0.233	-11.474	0.563	1.269	0.929	0.01347	0.322
gc	0.01-0.25	0.332	0.395	1,392,861	0.233	-11.936	0.750	1.492	1.141	0.01260	0.369
gc	0.25-0.5	0.395	0.447	1,449,024	0.256	-12.137	0.873	1.698	1.310	0.01461	0.419
gc	0.5-0.75	0.447	0.521	1,449,980	0.254	-13.575	0.987	1.778	1.402	0.01438	0.483
gc	0.75-0.99	0.521	0.644	1,390,064	0.225	-15.376	1.328	1.704	1.532	0.01125	0.564
gc	0.99-1	0.644	0.706	57,947	0.225	-16.780	2.147	0.775	1.466	0.01034	0.660
High GC, Male Recombination											
recomb_m	0-0.01	-0.010	0.000	98,083	0.230	-15.273	0.000	0.583	0.379	0.01139	0.608
recomb_m	0.01-0.25	0.000	0.242	47,884	0.207	-15.997	0.084	1.313	0.700	0.00892	0.593
recomb_m	0.25-0.5	0.242	0.987	151,389	0.196	-16.185	0.648	1.544	1.096	0.00816	0.599
recomb_m	0.5-0.75	0.987	2.081	139,459	0.209	-16.277	1.532	1.825	1.678	0.00994	0.604
recomb_m	0.75-0.99	2.081	8.735	141,879	0.202	-17.217	4.427	1.632	3.030	0.00841	0.610
recomb_m	0.99-1	8.735	10.372	2,811	0.348	-11.262	9.690	2.961	6.330	0.03380	0.585
High GC, Female Recombination											
recomb_f	0-0.01	-0.010	0.000	87,934	0.249	-15.971	0.698	0.000	0.349	0.01273	0.618
recomb_f	0.01-0.25	0.000	0.415	57,495	0.205	-16.672	2.991	0.211	1.602	0.00990	0.609
recomb_f	0.25-0.5	0.415	1.323	145,697	0.198	-16.101	1.632	0.771	1.202	0.00800	0.608
recomb_f	0.5-0.75	1.323	2.339	145,206	0.207	-15.733	1.171	1.784	1.529	0.00916	0.596
recomb_f	0.75-0.99	2.339	4.801	139,831	0.197	-16.924	2.309	3.095	2.705	0.00891	0.597
recomb_f	0.99-1	4.801	8.041	5,342	0.164	-18.475	1.319	5.943	3.734	0.00449	0.590

Table S12. GO enrichments for codon-wise and gene-wise positively selected genes

GO.ID	Term	GO term enrichments						
		Annotated	Significant	Expected	Rank in	pval.elim	pval.fis	pval.elim
ENRICHMENTS FOR CODON-WISE POSITIVE SELECTION								
GO:0007018	microtubule-based movement	288	89	44.54	1	0	0	0
GO:0007026	negative regulation of microtubule depol...	60	29	9.28	2	0	0	0
GO:0006265	DNA topological change	65	29	10.05	3	0	0	0
GO:0000723	telomere maintenance	53	23	8.2	4	0	0	0.01
GO:0090161	Golgi ribbon formation	7	7	1.08	5	0	0	0.01
GO:0043044	ATP-dependent chromatin remodeling	17	11	2.63	6	0	0	0.03
GO:0055114	oxidation reduction	1553	303	240.18	7	0	0	0.04
GO:0055072	iron ion homeostasis	69	27	10.67	50	0	0	0.01
GO:0006974	response to DNA damage stimulus	1023	212	158.21	66	0	0.01	0.02
GO:0006259	DNA metabolic process	1523	311	235.54	153	0	0.03	0
GO:0002460	adaptive immune response based on somati...	208	66	32.17	347	0	0.1	0
GO:0002250	adaptive immune response	211	66	32.63	449	0	0.13	0
GO:0032886	regulation of microtubule-based process	124	45	19.18	915	0	0.29	0
GO:0002455	humoral immune response mediated by circ...	78	29	12.06	918	0	0.3	0.01
GO:0002443	leukocyte mediated immunity	269	70	41.6	991	0	0.32	0.03
GO:0070507	regulation of microtubule cytoskeleton o...	110	42	17.01	1172	0	0.4	0
GO:0007017	microtubule-based process	621	166	96.04	1269	0	0.43	0
GO:0043242	negative regulation of protein complex d...	97	34	15	2085	0	0.69	0.01
GO:0031110	regulation of microtubule polymerization...	75	31	11.6	2086	0	0.7	0
GO:0031109	microtubule polymerization or depolymeri...	86	31	13.3	3065	0	0.93	0.01
ENRICHMENTS FOR GENE-WISE POSITIVE SELECTION								
GO:0045132	meiotic chromosome segregation	29	23	1.11	1	9.10E-028	9.10E-028	5.19E-024
GO:0006355	regulation of transcription, DNA-depende...	4628	312	177.86	2	5.40E-024	5.40E-024	3.08E-020
GO:0042742	defense response to bacterium	186	34	7.15	3	0	0	0
GO:0006955	immune response	2043	168	78.52	4	8.30E-021	0	4.73E-017
GO:0050909	sensory perception of taste	55	20	2.11	5	6.10E-015	0	0
GO:0007217	tachykinin receptor signaling pathway	21	11	0.81	6	0	0	0
GO:0031103	axon regeneration	37	13	1.42	7	0	0	0
GO:0006526	activation of transmembrane receptor pro...	16	9	0.61	8	0	0	0
GO:0031640	killing of cells of another organism	33	12	1.27	9	0	0	0
GO:0042523	positive regulation of tyrosine phosphor...	12	8	0.46	10	0	0	0
GO:0007566	embryo implantation	38	12	1.46	11	0	0	0
GO:0070257	positive regulation of mucus secretion	7	6	0.27	12	0	0	0
GO:0006952	defense response	1246	138	47.89	13	4.20E-029	0	2.39E-025
GO:0031424	keratinization	82	16	3.15	14	0	0	0
GO:0046080	dUTP metabolic process	5	5	0.19	15	0	0	0
GO:0006968	cellular defense response	115	18	4.42	16	0	0	0
GO:0006935	chemotaxis	314	36	12.07	17	0	0	0
GO:0007171	activation of transmembrane receptor pro...	15	7	0.58	18	0	0	0
GO:0045651	positive regulation of macrophage differ...	11	6	0.42	19	0	0	0.01
GO:0045086	positive regulation of interleukin-2 bio...	17	7	0.65	20	0	0	0.01

Notes: All terms were first sorted by 'pval.fis.bonf' (the Bonferroni-corrected p-value for enrichment) and a threshold of $p < 0.05$ was applied. Terms were subsequently sorted by 'pval.elim' for display purposes. Terms mentioned in the text are in bold, and only the top 20 terms for each type of enrichment are shown.

29 Mammals
Table S13

Table S13. PFAM domain enrichment site-wise and gene-wise positively selected genes

Pfam ID	Pfam link	Domain name	Notes / description	Mammals mean dNdS								
				N sites	N genes	Pos. sites	Neg. sites	Neutral sites	Fraction positive	Fraction negative	Fraction neutral	Mean dN/dS
TOP DOMAINS BY FRACTION OF POSITIVELY SELECTED SITES												
PF03765	PF03765	CRAL/TRIO, n-terminus	Comprises retinal-binding proteins and various transfer proteins	690	9	18	449	223	0.03	0.65	0.32	0.36
PF00031	PF00031	Cystatin	Proteinase inhibitors, play role in protein degradation, bone remodeling, possibly antigen presentation	1404	11	31	816	557	0.02	0.58	0.4	0.41
PF06467	PF06467	MYM-type zinc finger	Zinc finger in proteins associated with chromosomal translocations and myeloproliferative disorders	1594	6	31	909	654	0.02	0.57	0.41	0.42
PF00048	PF00048	IL8	Immune response	1218	17	23	633	562	0.02	0.52	0.46	0.46
PF00957	PF00957	Synaptobrevin	Membrane protein of neuronal synaptic vesicles	600	9	11	294	295	0.02	0.49	0.49	0.41
PF04103	PF04103	CD20-like family	IgE Fc receptor subunit	1649	12	30	845	774	0.02	0.51	0.47	0.46
PF03645	PF03645	Tctex-1	Binds to rhodopsin, possible role in retinitis pigmentosa	596	6	10	375	211	0.02	0.63	0.35	0.59
PF01105	PF01105	Emp24 / GOLD / p24 family	Major membrane components of COPI and COPII coated vesicles	1319	15	21	971	327	0.02	0.74	0.25	0.28
PF05287	PF05287	PMG	Keratin-associated proteins	1024	6	16	203	805	0.02	0.2	0.79	0.74
PF00025	PF00025	ADP-ribosylation factor	Regulators of vesicle biogenesis	908	24	14	565	329	0.02	0.62	0.36	0.38
TOP DOMAINS BY MEAN DN/DS												
PF05287	PF05287	PMG	Keratin-associated proteins	1024	6	16	203	805	0.02	0.2	0.79	0.74
PF03645	PF03645	Tctex-1	Binds to rhodopsin, possible role in retinitis pigmentosa	596	6	10	375	211	0.02	0.63	0.35	0.59
PF00048	PF00048	IL8	Immune response	1218	17	23	633	562	0.02	0.52	0.46	0.46
PF04103	PF04103	CD20		1649	12	30	845	774	0.02	0.51	0.47	0.46
PF00230	PF00230	MIP		1734	10	26	991	717	0.01	0.57	0.41	0.42
PF00095	PF00095	WAP	Milk whey component, protease-inhibitor	564	8	2	309	253	0	0.55	0.45	0.42
PF00711	PF00711	Beta defensin		687	13	7	399	281	0.01	0.58	0.41	0.42
PF06467	PF06467	MYM-type zinc finger		1594	6	31	909	654	0.02	0.57	0.41	0.42
PF01390	PF01390	SEA domain	Sea urchin sperm protein, proteolytic activity	1876	10	16	1012	848	0.01	0.54	0.45	0.41
PF00957	PF00957	Synaptobrevin		600	9	11	294	295	0.02	0.49	0.49	0.41

Notes: Rows were sorted by either mean dN/dS or fraction of positive sites, and the top 10 domains for each set were retained. Those domains showing up in the top 10 on both lists were grayed out. Domains mentioned in the text are in bold.

Table S14. Transcription factor binding sites

JASPAR family	Matrix ID	Min	Max	Mean
ETS class	MF0001	0	13	0.41
bZIP CREB/G-box-like subclass	MF0002	0	7	0.21
REL class	MF0003	0	17	0.33
Nuclear receptor class	MF0004	0	13	0.78
Forkhead class	MF0005	0	13	0.36
bZIP cEBP-like subclass	MF0006	0	11	0.4
bHLH(xip) class	MF0007	0	14	0.58
MADS class	MF0008	0	29	0.42
TRP(MYB) class	MF0009	0	13	0.66
Homeobox class	MF0010	0	21	0.91
HMG class	MF0011	0	14	0.89
TOTAL		0	29	0.54

Table S15. Gene ontology enrichments for human- and primate-accelerated regions

ID	DESCRIPTION	ONTOLOGY	Gene-avg Vert.	Gene-codon Vert.	Codons Vert.	Gene-avg Prim.	Gene-codon Prim.	Codons Prim.	PARS_100kb	PARS_1kb	HARS_100kb	HARS_1kb
GO:0006955	immune response	BP	2.E-36	1.E-08	1.E-228	2.E-15	8.E-07	2.E-102	1.E+00	1.E+00	1.E+00	1.E+00
GO:0005576	extracellular region	CC	3.E-20	1.E+00	6.E-68	8.E-07	1.E+00	1.E-22	1.E+00	1.E+00	1.E+00	1.E+00
GO:0005090	sensory perception of taste	BP	7.E-14	8.E-04	8.E-165	4.E-12	5.E-06	1.E-120	1.E+00	1.E+00	1.E+00	1.E+00
GO:0005125	cytokine activity	MF	9.E-14	1.E+00	1.E-32	4.E-07	1.E+00	7.E-16	1.E+00	1.E+00	1.E+00	1.E+00
GO:0008009	chemokine activity	MF	4.E-10	1.E+00	6.E-19	9.E-01	4.E-02	2.E-15	1.E+00	1.E+00	1.E+00	1.E+00
GO:0031424	keratinization	BP	9.E-10	5.E-02	2.E-200	2.E-01	4.E-05	7.E-136	1.E+00	1.E+00	1.E+00	1.E+00
GO:0005529	sugar binding	MF	1.E-09	1.E+00	1.E-41	5.E-01	1.E+00	1.E-23	1.E+00	1.E+00	1.E+00	1.E+00
GO:0005615	extracellular space	CC	3.E-07	1.E+00	1.E+00	2.E-01	1.E+00	1.E+00	2.E-09	4.E-46	1.E+00	1.E+00
GO:0008527	taste receptor activity	MF	4.E-07	2.E-01	1.E-61	1.E-05	7.E-02	1.E-38	1.E+00	1.E+00	1.E+00	1.E+00
GO:0007166	cell surface receptor linked signal transduction	BP	5.E-06	1.E+00	2.E-33	2.E-03	1.E+00	1.E-09	1.E+00	1.E+00	1.E+00	1.E+00
GO:0004888	transmembrane receptor activity	MF	7.E-06	1.E+00	2.E-40	1.E-01	1.E+00	3.E-04	1.E+00	1.E+00	1.E+00	1.E+00
GO:0042612	MHC class I protein complex	CC	8.E-06	1.E+00	2.E-122	1.E-07	2.E-09	3.E-44	1.E+00	1.E+00	1.E+00	1.E+00
GO:0006952	defense response	BP	9.E-06	1.E+00	2.E-96	2.E-04	1.E+00	4.E-30	1.E+00	1.E+00	1.E+00	1.E+00
GO:0006954	inflammatory response	BP	1.E-05	5.E-01	5.E-17	1.E+00	1.E+00	2.E-06	1.E+00	1.E+00	1.E+00	1.E+00
GO:0006935	chemotaxis	BP	2.E-05	1.E+00	7.E-26	1.E+00	3.E-03	2.E-02	2.E-91	2.E-144	1.E+00	1.E+00
GO:0019882	antigen processing and presentation	BP	3.E-05	1.E+00	6.E-112	3.E-05	7.E-05	2.E-36	1.E+00	1.E+00	1.E+00	1.E+00
GO:0009615	response to virus	BP	2.E-03	1.E+00	6.E-48	5.E-02	1.E+00	6.E-17	1.E+00	1.E+00	1.E+00	1.E+00
GO:0001533	cornified envelope	CC	3.E-03	1.E-04	4.E-51	3.E-02	1.E-06	4.E-21	1.E+00	1.E+00	1.E+00	1.E+00
GO:0004872	receptor activity	MF	4.E-03	5.E-20	4.E-39	1.E+00	1.E+00	1.E+00	1.E+00	8.E-13	1.E+00	1.E+00
GO:0042742	defense response to bacterium	BP	6.E-03	7.E-01	1.E-18	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00
GO:0016814	hydrolase activity, acting on carbon-nitrogen (but not MF	MF	7.E-03	1.E-01	1.E-07	6.E-03	1.E+00	3.E-39	1.E+00	1.E+00	1.E+00	1.E+00
GO:0004522	pancreatic ribonuclease activity	MF	7.E-03	1.E+00	4.E-05	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00
GO:0016021	integral to membrane	CC	7.E-03	1.E-11	6.E-13	1.E+00	1.E+00	1.E+00	1.E+00	2.E-16	1.E+00	1.E+00
GO:0006953	acute-phase response	BP	1.E-02	1.E+00	2.E-50	9.E-04	1.E+00	2.E-10	1.E+00	1.E+00	1.E+00	1.E+00
GO:0002504	antigen processing and presentation of peptide or polyBP	BP	2.E-02	4.E-03	8.E-11	1.E+00	1.E+00	1.E-09	1.E+00	1.E+00	2.E-02	2.E-07
GO:0042613	MHC class II protein complex	CC	2.E-02	4.E-03	8.E-11	1.E+00	1.E+00	1.E-09	1.E+00	1.E+00	2.E-02	2.E-07
GO:0005126	hematopoietin/interferon-class (D200-domain) cytokinMF	MF	2.E-02	1.E+00	3.E-38	2.E-02	1.E+00	1.E-12	1.E+00	1.E+00	1.E+00	1.E+00
GO:0030101	natural killer cell activation	BP	2.E-02	1.E+00	2.E-97	2.E-02	1.E+00	7.E-24	1.E+00	1.E+00	1.E+00	1.E+00
GO:0019864	IgG binding	MF	3.E-02	2.E-01	3.E-45	3.E-02	1.E+00	5.E-32	1.E+00	1.E+00	1.E+00	1.E+00
GO:0005882	intermediate filament	CC	6.E-02	7.E-01	4.E-04	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00
GO:0045087	innate immune response	BP	1.E-01	1.E+00	1.E-44	1.E+00	1.E+00	4.E-29	1.E+00	1.E+00	1.E+00	1.E+00
GO:0045410	positive regulation of interleukin-6 biosynthetic procesBP	BP	2.E-01	1.E+00	3.E-16	1.E+00	1.E+00	4.E-06	1.E+00	1.E+00	1.E+00	1.E+00
GO:0009897	external side of plasma membrane	CC	2.E-01	1.E+00	5.E-41	4.E-02	1.E+00	2.E-11	1.E+00	1.E+00	1.E+00	1.E+00
GO:0001580	detection of chemical stimulus involved in sensory per BP	BP	3.E-01	1.E+00	2.E-08	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00
GO:0004866	endopeptidase inhibitor activity	MF	3.E-01	1.E+00	7.E-48	1.E+00	1.E+00	4.E-08	1.E+00	1.E+00	1.E+00	1.E+00
GO:0030216	keratinocyte differentiation	BP	4.E-01	7.E-01	4.E-20	1.E+00	4.E-03	5.E-06	1.E+00	1.E+00	1.E+00	1.E+00
GO:0005179	hormone activity	MF	5.E-01	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00
GO:0042698	ovulation cycle	BP	5.E-01	1.E+00	1.E-17	5.E-01	1.E-02	3.E-19	1.E+00	1.E+00	1.E+00	1.E+00
GO:0004867	serine-type endopeptidase inhibitor activity	MF	5.E-01	1.E+00	5.E-45	1.E+00	1.E+00	2.E-07	1.E+00	1.E+00	1.E+00	1.E+00
GO:0007267	cell-cell signaling	BP	9.E-01	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00
GO:0004930	G-protein coupled receptor activity	MF	1.E+00	1.E-22	3.E-74	1.E+00	1.E+00	4.E-11	1.E+00	1.E+00	1.E+00	1.E+00
GO:0050896	response to stimulus	BP	1.E+00	3.E-28	3.E-43	1.E+00	1.E+00	5.E-10	1.E+00	1.E+00	1.E+00	1.E+00
GO:0007186	G-protein coupled receptor protein signaling pathway BP	BP	1.E+00	2.E-18	5.E-67	1.E+00	1.E+00	4.E-08	1.E+00	1.E+00	1.E+00	1.E+00
GO:0016020	membrane	CC	1.E+00	1.E-08	3.E-02	1.E+00	1.E+00	1.E+00	1.E+00	2.E-19	1.E+00	1.E+00
GO:0005149	interleukin-1 receptor binding	MF	1.E+00	1.E+00	2.E-15	5.E-02	1.E+00	2.E-08	1.E+00	1.E+00	1.E+00	1.E+00
GO:0001584	rhodopsin-like receptor activity	MF	1.E+00	5.E-21	5.E-36	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00
GO:0004984	olfactory receptor activity	MF	1.E+00	5.E-29	7.E-21	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00
GO:0007608	sensory perception of smell	BP	1.E+00	2.E-28	8.E-16	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00
GO:0004871	signal transducer activity	MF	1.E+00	1.E-11	6.E-23	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00
GO:0005694	chromosome	CC	1.E+00	1.E+00	4.E-02	1.E+00	1.E+00	7.E-04	1.E+00	7.E-01	1.E+00	1.E+00
GO:0000775	chromosome, pericentric region	CC	1.E+00	1.E+00	3.E-36	1.E+00	1.E+00	9.E-17	1.E+00	1.E+00	1.E+00	1.E+00
GO:0003950	NAD+ ADP-ribosyltransferase activity	MF	1.E+00	1.E+00	1.E-21	1.E+00	1.E+00	3.E-26	1.E+00	1.E+00	1.E+00	1.E+00
GO:0005198	structural molecule activity	MF	1.E+00	1.E+00	1.E-118	1.E+00	1.E+00	7.E-53	1.E+00	1.E+00	1.E+00	1.E+00
GO:0006281	DNA repair	BP	1.E+00	1.E+00	2.E-38	1.E+00	1.E+00	1.E-26	1.E+00	1.E+00	1.E+00	1.E+00
GO:0006282	regulation of DNA repair	BP	1.E+00	1.E+00	6.E-38	1.E+00	1.E+00	3.E-23	1.E+00	1.E+00	1.E+00	1.E+00
GO:0006471	protein amino acid ADP-ribosylation	BP	1.E+00	1.E+00	8.E-20	1.E+00	1.E+00	7.E-23	1.E+00	1.E+00	1.E+00	1.E+00
GO:0006817	phosphate transport	BP	1.E+00	1.E+00	1.E-30	1.E+00	1.E+00	7.E-21	1.E+00	1.E+00	1.E+00	1.E+00
GO:0006974	response to DNA damage stimulus	BP	1.E+00	1.E+00	5.E-37	1.E+00	1.E+00	1.E-27	1.E+00	1.E+00	1.E+00	1.E+00
GO:0007051	spindle organization and biogenesis	BP	1.E+00	1.E+00	2.E-46	1.E+00	1.E+00	3.E-18	1.E+00	1.E+00	1.E+00	1.E+00
GO:0007098	centrosome cycle	BP	1.E+00	1.E+00	1.E-20	1.E+00	1.E+00	2.E-18	1.E+00	1.E+00	1.E+00	1.E+00
GO:0008656	caspase activator activity	MF	1.E+00	1.E+00	2.E-74	1.E+00	1.E+00	2.E-19	1.E+00	1.E+00	1.E+00	1.E+00
GO:0017114	wide-spectrum protease inhibitor activity	MF	1.E+00	1.E+00	4.E-46	1.E+00	1.E+00	1.E-18	1.E+00	1.E+00	1.E+00	1.E+00
GO:0043159	acrosomal matrix	CC	1.E+00	1.E+00	2.E-77	1.E+00	1.E+00	5.E-19	1.E+00	1.E+00	1.E+00	1.E+00
GO:0045569	TRAIL binding	MF	1.E+00	1.E+00	2.E-16	1.E+00	1.E+00	2.E-18	1.E+00	1.E+00	1.E+00	1.E+00
GO:0042102	positive regulation of T cell proliferation	BP	1.E+00	1.E+00	1.E-21	1.E+00	1.E+00	1.E-14	1.E+00	1.E+00	1.E+00	1.E+00
GO:0006919	caspase activation	BP	1.E+00	1.E+00	2.E-58	1.E+00	1.E+00	1.E-12	1.E+00	1.E+00	1.E+00	1.E+00
GO:0007566	embryo implantation	BP	1.E+00	1.E+00	2.E-12	1.E+00	1.E+00	1.E-12	1.E+00	1.E+00	1.E+00	1.E+00

GO:0008625	induction of apoptosis via death domain receptors	BP	1.E+00	1.E+00	4.E-12	1.E+00	1.E+00	2.E-14	1.E+00	1.E+00	1.E+00	1.E+00
GO:0009048	dosage compensation, by inactivation of X chromosome	BP	1.E+00	1.E+00	1.E-23	1.E+00	1.E+00	1.E-11	1.E+00	1.E+00	1.E+00	1.E+00
GO:0045739	positive regulation of DNA repair	BP	1.E+00	1.E+00	6.E-33	1.E+00	1.E+00	5.E-11	1.E+00	1.E+00	1.E+00	1.E+00
GO:0006978	DNA damage response, signal transduction by p53 class	BP	1.E+00	1.E+00	9.E-33	1.E+00	1.E+00	3.E-10	1.E+00	1.E+00	1.E+00	1.E+00
GO:0042759	long-chain fatty acid biosynthetic process	BP	1.E+00	1.E+00	9.E-33	1.E+00	1.E+00	3.E-10	1.E+00	1.E+00	1.E+00	1.E+00
GO:0046600	negative regulation of centriole replication	BP	1.E+00	1.E+00	9.E-33	1.E+00	1.E+00	3.E-10	1.E+00	1.E+00	1.E+00	1.E+00
GO:0019899	enzyme binding	MF	1.E+00	1.E+00	2.E-41	1.E+00	1.E+00	3.E-10	1.E+00	1.E+00	1.E+00	1.E+00
GO:0005581	collagen	CC	1.E+00	1.E+00	4.E-10	1.E+00	1.E+00	9.E-25	1.E+00	1.E+00	1.E+00	1.E+00
GO:0007218	neuropeptide signaling pathway	BP	1.E+00	1.E+00	3.E-18	1.E+00	1.E+00	1.E-09	1.E+00	1.E+00	1.E+00	1.E+00
GO:0000940	outer kinetochore of condensed chromosome	CC	1.E+00	1.E+00	1.E-15	1.E+00	1.E+00	2.E-09	1.E+00	1.E+00	1.E+00	1.E+00
GO:0031436	BRCA1-BARD1 complex	CC	1.E+00	1.E+00	9.E-33	1.E+00	1.E+00	2.E-09	1.E+00	1.E+00	1.E+00	1.E+00
GO:0051298	centrosome duplication	BP	1.E+00	1.E+00	5.E-26	1.E+00	1.E+00	3.E-09	1.E+00	1.E+00	1.E+00	1.E+00
GO:0006958	complement activation, classical pathway	BP	1.E+00	1.E+00	1.E-10	1.E+00	1.E+00	5.E-09	1.E+00	1.E+00	1.E+00	1.E+00
GO:0031398	positive regulation of protein ubiquitination	BP	1.E+00	1.E+00	2.E-28	1.E+00	1.E+00	1.E-08	1.E+00	1.E+00	1.E+00	1.E+00
GO:0019959	interleukin-8 binding	MF	1.E+00	1.E+00	2.E-19	1.E+00	1.E+00	2.E-08	1.E+00	1.E+00	1.E+00	1.E+00
GO:0043120	tumor necrosis factor binding	MF	1.E+00	1.E+00	2.E-19	1.E+00	1.E+00	2.E-08	1.E+00	1.E+00	1.E+00	1.E+00
GO:0004905	interferon-alpha/beta receptor activity	MF	1.E+00	1.E+00	2.E-08	1.E+00	1.E+00	8.E-09	1.E+00	1.E+00	1.E+00	1.E+00
GO:0016046	detection of fungus	BP	1.E+00	1.E+00	4.E-19	1.E+00	1.E+00	6.E-08	1.E+00	1.E+00	1.E+00	1.E+00
GO:0045362	positive regulation of interleukin-1 biosynthetic process	BP	1.E+00	1.E+00	4.E-19	1.E+00	1.E+00	6.E-08	1.E+00	1.E+00	1.E+00	1.E+00
GO:0007250	activation of NF-kappaB-inducing kinase activity	BP	1.E+00	1.E+00	8.E-08	1.E+00	1.E+00	2.E-08	1.E+00	1.E+00	1.E+00	1.E+00
GO:0030431	sleep	BP	1.E+00	1.E+00	4.E-32	1.E+00	1.E+00	1.E-07	1.E+00	1.E+00	1.E+00	1.E+00
GO:0019966	interleukin-1 binding	MF	1.E+00	1.E+00	6.E-18	1.E+00	1.E+00	1.E-07	1.E+00	1.E+00	1.E+00	1.E+00
GO:0042088	T-helper 1 type immune response	BP	1.E+00	1.E+00	4.E-21	1.E+00	1.E+00	2.E-07	1.E+00	1.E+00	1.E+00	1.E+00
GO:0000089	mitotic metaphase	BP	1.E+00	1.E+00	5.E-17	1.E+00	1.E+00	2.E-07	1.E+00	1.E+00	1.E+00	1.E+00
GO:0008320	protein transmembrane transporter activity	MF	1.E+00	1.E+00	4.E-12	1.E+00	1.E+00	2.E-07	1.E+00	1.E+00	1.E+00	1.E+00
GO:0001675	acrosome formation	BP	1.E+00	1.E+00	5.E-35	1.E+00	1.E+00	8.E-07	1.E+00	1.E+00	1.E+00	1.E+00
GO:0003674	molecular_function	MF	1.E+00	1.E+00	3.E-12	1.E+00	1.E+00	1.E-06	1.E+00	1.E+00	1.E+00	1.E+00
GO:0030997	regulation of centriole-centriole cohesion	BP	1.E+00	1.E+00	8.E-23	1.E+00	1.E+00	1.E-06	1.E+00	1.E+00	1.E+00	1.E+00
GO:0051382	kinetochore assembly	BP	1.E+00	1.E+00	2.E-16	1.E+00	1.E+00	1.E-06	1.E+00	1.E+00	1.E+00	1.E+00
GO:0007079	mitotic chromosome movement towards spindle pole	BP	1.E+00	1.E+00	1.E-17	1.E+00	1.E+00	2.E-06	1.E+00	1.E+00	1.E+00	1.E+00
GO:0019835	cytolysis	BP	1.E+00	1.E+00	1.E-11	1.E+00	1.E+00	2.E-06	1.E+00	1.E+00	1.E+00	1.E+00
GO:0045576	mast cell activation	BP	1.E+00	1.E+00	4.E-17	1.E+00	1.E+00	2.E-06	1.E+00	1.E+00	1.E+00	1.E+00
GO:0051402	neuron apoptosis	BP	1.E+00	1.E+00	1.E-44	1.E+00	1.E+00	3.E-06	1.E+00	1.E+00	1.E+00	1.E+00
GO:0004865	type 1 serine/threonine specific protein phosphatase inhibition	MF	1.E+00	1.E+00	4.E-22	1.E+00	1.E+00	4.E-06	1.E+00	1.E+00	1.E+00	1.E+00
GO:0006957	complement activation, alternative pathway	BP	1.E+00	1.E+00	1.E-22	1.E+00	1.E+00	6.E-06	1.E+00	1.E+00	1.E+00	1.E+00
GO:0046696	lipopolysaccharide receptor complex	CC	1.E+00	1.E+00	4.E-16	1.E+00	1.E+00	6.E-06	1.E+00	1.E+00	1.E+00	1.E+00
GO:0008274	gamma-tubulin ring complex	CC	1.E+00	1.E+00	5.E-31	1.E+00	1.E+00	8.E-06	1.E+00	1.E+00	1.E+00	1.E+00
GO:0009898	internal side of plasma membrane	CC	1.E+00	1.E+00	2.E-32	1.E+00	1.E+00	1.E-05	1.E+00	1.E+00	1.E+00	1.E+00
GO:0007049	cell cycle	BP	1.E+00	1.E+00	1.E-32	1.E+00	1.E+00	2.E-05	1.E+00	1.E+00	1.E+00	1.E+00
GO:0042105	alpha-beta T cell receptor complex	CC	1.E+00	1.E+00	5.E-21	1.E+00	1.E+00	2.E-05	1.E+00	1.E+00	1.E+00	1.E+00
GO:0042923	neuropeptide binding	MF	1.E+00	1.E+00	2.E-30	1.E+00	1.E+00	2.E-05	1.E+00	1.E+00	1.E+00	1.E+00
GO:0050708	regulation of protein secretion	BP	1.E+00	1.E+00	4.E-06	1.E+00	1.E+00	2.E-05	1.E+00	1.E+00	1.E+00	1.E+00
GO:0001530	lipopolysaccharide binding	MF	1.E+00	1.E+00	6.E-15	1.E+00	1.E+00	2.E-05	1.E+00	1.E+00	1.E+00	1.E+00
GO:0007128	meiotic prophase I	BP	1.E+00	1.E+00	2.E-13	1.E+00	1.E+00	3.E-05	1.E+00	1.E+00	1.E+00	1.E+00
GO:0030414	protease inhibitor activity	MF	1.E+00	1.E+00	6.E-14	1.E+00	1.E+00	3.E-05	1.E+00	1.E+00	1.E+00	1.E+00
GO:0005876	spindle microtubule	CC	1.E+00	1.E+00	7.E-09	1.E+00	1.E+00	4.E-05	1.E+00	1.E+00	1.E+00	1.E+00
GO:0006959	humoral immune response	BP	1.E+00	1.E+00	4.E-05	1.E+00	1.E+00	1.E-08	1.E+00	1.E+00	1.E+00	1.E+00
GO:0005579	membrane attack complex	CC	1.E+00	1.E+00	1.E-12	1.E+00	1.E+00	4.E-05	1.E+00	1.E+00	1.E+00	1.E+00
GO:0007080	mitotic metaphase plate congression	BP	1.E+00	1.E+00	1.E-13	1.E+00	1.E+00	6.E-05	1.E+00	1.E+00	1.E+00	1.E+00
GO:0005813	centrosome	CC	1.E+00	1.E+00	4.E-08	1.E+00	1.E+00	7.E-05	1.E+00	1.E+00	1.E+00	1.E+00
GO:0005540	hyaluronic acid binding	MF	1.E+00	1.E+00	3.E-13	1.E+00	1.E+00	9.E-05	1.E+00	1.E+00	1.E+00	1.E+00
GO:0042981	regulation of apoptosis	BP	1.E+00	1.E+00	7.E-23	1.E+00	1.E+00	1.E-04	1.E+00	1.E+00	1.E+00	1.E+00
GO:0045368	positive regulation of interleukin-13 biosynthetic process	BP	1.E+00	1.E+00	2.E-12	1.E+00	1.E+00	1.E-04	1.E+00	1.E+00	1.E+00	1.E+00
GO:0030693	caspase activity	MF	1.E+00	1.E+00	9.E-05	1.E+00	1.E+00	3.E-05	1.E+00	1.E+00	1.E+00	1.E+00
GO:0005730	nucleolus	CC	1.E+00	1.E+00	5.E-06	1.E+00	1.E+00	2.E-04	1.E+00	1.E+00	1.E+00	1.E+00
GO:0007126	meiosis	BP	1.E+00	1.E+00	3.E-36	1.E+00	1.E+00	2.E-04	1.E+00	1.E+00	1.E+00	1.E+00
GO:0005590	collagen type VII	CC	1.E+00	1.E+00	7.E-10	1.E+00	1.E+00	3.E-04	1.E+00	1.E+00	1.E+00	1.E+00
GO:0000776	kinetochore	CC	1.E+00	1.E+00	2.E-20	1.E+00	1.E+00	4.E-04	1.E+00	1.E+00	1.E+00	1.E+00
GO:0005132	interferon-alpha/beta receptor binding	MF	1.E+00	1.E+00	3.E-21	1.E+00	1.E+00	4.E-04	1.E+00	1.E+00	1.E+00	1.E+00
GO:0045840	positive regulation of mitosis	BP	1.E+00	1.E+00	9.E-11	1.E+00	1.E+00	4.E-04	1.E+00	1.E+00	1.E+00	1.E+00
GO:0000281	cytokinesis after mitosis	BP	1.E+00	1.E+00	4.E-11	1.E+00	1.E+00	4.E-04	1.E+00	1.E+00	1.E+00	1.E+00
GO:0008157	protein phosphatase 1 binding	MF	1.E+00	1.E+00	6.E-04	1.E+00	1.E+00	2.E-04	1.E+00	1.E+00	1.E+00	1.E+00
GO:0008630	DNA damage response, signal transduction resulting in	BP	1.E+00	1.E+00	2.E-20	1.E+00	1.E+00	8.E-04	1.E+00	1.E+00	1.E+00	1.E+00
GO:0003684	damaged DNA binding	MF	1.E+00	1.E+00	9.E-22	1.E+00	1.E+00	8.E-04	1.E+00	1.E+00	1.E+00	1.E+00
GO:0045084	positive regulation of interleukin-12 biosynthetic process	BP	1.E+00	1.E+00	2.E-11	1.E+00	1.E+00	9.E-04	1.E+00	1.E+00	1.E+00	1.E+00
GO:0003880	C-terminal protein carboxyl methyltransferase activity	MF	1.E+00	1.E+00	1.E-03	1.E+00	1.E+00	2.E-07	1.E+00	1.E+00	1.E+00	1.E+00
GO:0006481	C-terminal protein amino acid methylation	BP	1.E+00	1.E+00	1.E-03	1.E+00	1.E+00	2.E-07	1.E+00	1.E+00	1.E+00	1.E+00
GO:0045089	positive regulation of innate immune response	BP	1.E+00	1.E+00	9.E-13	1.E+00	1.E+00	1.E-03	1.E+00	1.E+00	1.E+00	1.E+00
GO:0042116	macrophage activation	BP	1.E+00	1.E+00	2.E-09	1.E+00	1.E+00	1.E-03	1.E+00	1.E+00	1.E+00	1.E+00
GO:0000793	condensed chromosome	CC	1.E+00	1.E+00	4.E-11	1.E+00	1.E+00	1.E-03	1.E+00	1.E+00	1.E+00	1.E+00
GO:0019863	IgE binding	MF	1.E+00	1.E+00	3.E-14	1.E+00	1.E+00	1.E-03	1.E+00	1.E+00	1.E+00	1.E+00
GO:0019901	protein kinase binding	MF	1.E+00	1.E+00	2.E-03	1.E+00	1.E+00	2.E-05	1.E+00	1.E+00	1.E+00	1.E+00
GO:0007059	chromosome segregation	BP	1.E+00	1.E+00	2.E-07	1.E+00	1.E+00	2.E-03	1.E+00	1.E+00	1.E+00	1.E+00
GO:0045580	regulation of T cell differentiation	BP	1.E+00	1.E+00	1.E-10	1.E+00	1.E+00	2.E-03	1.E+00	1.E+00	1.E+00	1.E+00

GO:0006290	pyrimidine dimer repair	BP	1.E+00	1.E+00	2.E-03	1.E+00	1.E+00	2.E-11	1.E+00	1.E+00	1.E+00	1.E+00
GO:0015999	eta DNA polymerase activity	MF	1.E+00	1.E+00	2.E-03	1.E+00	1.E+00	2.E-11	1.E+00	1.E+00	1.E+00	1.E+00
GO:0016451	nu DNA polymerase activity	MF	1.E+00	1.E+00	2.E-03	1.E+00	1.E+00	2.E-11	1.E+00	1.E+00	1.E+00	1.E+00
GO:0030690	Noc1p-Noc2p complex	CC	1.E+00	1.E+00	6.E-12	1.E+00	1.E+00	2.E-03	1.E+00	1.E+00	1.E+00	1.E+00
GO:0015631	tubulin binding	MF	1.E+00	1.E+00	6.E-13	1.E+00	1.E+00	2.E-03	1.E+00	1.E+00	1.E+00	1.E+00
GO:0006956	complement activation	BP	1.E+00	1.E+00	8.E-08	1.E+00	1.E+00	3.E-03	1.E+00	1.E+00	1.E+00	1.E+00
GO:0008301	DNA bending activity	MF	1.E+00	1.E+00	3.E-03	1.E+00	1.E+00	4.E-15	1.E+00	1.E+00	1.E+00	1.E+00
GO:0030225	macrophage differentiation	BP	1.E+00	1.E+00	5.E-33	1.E+00	1.E+00	4.E-03	1.E+00	1.E+00	1.E+00	1.E+00
GO:0048306	calcium-dependent protein binding	MF	1.E+00	1.E+00	9.E-07	1.E+00	1.E+00	4.E-03	1.E+00	1.E+00	1.E+00	1.E+00
GO:0000781	chromosome, telomeric region	CC	1.E+00	1.E+00	4.E-03	1.E+00	1.E+00	3.E-04	1.E+00	1.E+00	1.E+00	1.E+00
GO:0001730	2'-5'-oligoadenylate synthetase activity	MF	1.E+00	1.E+00	7.E-04	1.E+00	1.E+00	4.E-03	1.E+00	1.E+00	1.E+00	1.E+00
GO:0004918	interleukin-8 receptor activity	MF	1.E+00	1.E+00	1.E-24	1.E+00	1.E+00	6.E-03	1.E+00	1.E+00	1.E+00	1.E+00
GO:0015711	organic anion transport	BP	1.E+00	1.E+00	1.E-11	1.E+00	1.E+00	6.E-03	1.E+00	1.E+00	1.E+00	1.E+00
GO:0040001	establishment of mitotic spindle localization	BP	1.E+00	1.E+00	4.E-11	1.E+00	1.E+00	7.E-03	1.E+00	1.E+00	1.E+00	1.E+00
GO:0042384	cilium biogenesis	BP	1.E+00	1.E+00	2.E-25	1.E+00	1.E+00	9.E-03	1.E+00	1.E+00	1.E+00	1.E+00
GO:0030887	positive regulation of myeloid dendritic cell activation	BP	1.E+00	1.E+00	3.E-09	1.E+00	1.E+00	1.E-02	1.E+00	1.E+00	1.E+00	1.E+00
GO:0019955	cytokine binding	MF	1.E+00	1.E+00	2.E-27	1.E+00	1.E+00	1.E-02	1.E+00	1.E+00	1.E+00	1.E+00
GO:0016410	N-acyltransferase activity	MF	1.E+00	1.E+00	1.E-08	1.E+00	1.E+00	1.E-02	1.E+00	1.E+00	1.E+00	1.E+00
GO:0047963	glycine N-choloyltransferase activity	MF	1.E+00	1.E+00	1.E-08	1.E+00	1.E+00	1.E-02	1.E+00	1.E+00	1.E+00	1.E+00
GO:0006915	apoptosis	BP	1.E+00	1.E+00	2.E-04	1.E+00	1.E+00	1.E-02	1.E+00	1.E+00	1.E+00	1.E+00
GO:0045132	meiotic chromosome segregation	BP	1.E+00	1.E+00	5.E-04	1.E+00	1.E+00	1.E-02	1.E+00	1.E+00	1.E+00	1.E+00
GO:0015299	solute:hydrogen antiporter activity	MF	1.E+00	1.E+00	1.E-02	1.E+00	1.E+00	2.E-07	1.E+00	1.E+00	1.E+00	1.E+00
GO:0042100	B cell proliferation	BP	1.E+00	1.E+00	2.E-02	1.E+00	1.E+00	2.E-08	1.E+00	1.E+00	1.E+00	1.E+00
GO:0048305	immunoglobulin secretion	BP	1.E+00	1.E+00	2.E-02	1.E+00	1.E+00	1.E-07	1.E+00	1.E+00	1.E+00	1.E+00
GO:0030139	endocytic vesicle	CC	1.E+00	1.E+00	2.E-05	1.E+00	1.E+00	2.E-02	1.E+00	1.E+00	1.E+00	1.E+00
GO:0005035	death receptor activity	MF	1.E+00	1.E+00	2.E-02	1.E+00	1.E+00	4.E-03	1.E+00	1.E+00	1.E+00	1.E+00
GO:0018675	(S)-limonene 6-monooxygenase activity	MF	1.E+00	1.E+00	5.E-04	1.E+00	1.E+00	2.E-02	1.E+00	1.E+00	1.E+00	1.E+00
GO:0018676	(S)-limonene 7-monooxygenase activity	MF	1.E+00	1.E+00	5.E-04	1.E+00	1.E+00	2.E-02	1.E+00	1.E+00	1.E+00	1.E+00
GO:0030271	chymase activity	MF	1.E+00	1.E+00	6.E-04	1.E+00	1.E+00	3.E-02	1.E+00	1.E+00	1.E+00	1.E+00
GO:0030890	positive regulation of B cell proliferation	BP	1.E+00	1.E+00	9.E-06	1.E+00	1.E+00	3.E-02	1.E+00	1.E+00	1.E+00	1.E+00
GO:0045351	interferon type I biosynthetic process	BP	1.E+00	1.E+00	3.E-02	1.E+00	1.E+00	3.E-06	1.E+00	1.E+00	1.E+00	1.E+00
GO:0043515	kinetochore binding	MF	1.E+00	1.E+00	6.E-29	1.E+00	1.E+00	3.E-02	1.E+00	1.E+00	1.E+00	1.E+00
GO:0005597	collagen type XVI	CC	1.E+00	1.E+00	5.E-09	1.E+00	1.E+00	4.E-02	1.E+00	1.E+00	1.E+00	1.E+00
GO:0005594	collagen type IX	CC	1.E+00	1.E+00	4.E-02	1.E+00	1.E+00	5.E-09	1.E+00	1.E+00	1.E+00	1.E+00
GO:0050906	detection of stimulus involved in sensory perception	BP	1.E+00	1.E+00	7.E-10	1.E+00	1.E+00	5.E-02	1.E+00	1.E+00	1.E+00	1.E+00
GO:0043229	intracellular organelle	CC	1.E+00	1.E+00	8.E-11	1.E+00	1.E+00	5.E-02	1.E+00	1.E+00	1.E+00	1.E+00
GO:0000278	mitotic cell cycle	BP	1.E+00	1.E+00	6.E-08	1.E+00	1.E+00	5.E-02	1.E+00	1.E+00	1.E+00	1.E+00
GO:0006325	establishment and/or maintenance of chromatin archi	BP	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	9.E-02	1.E-08	3.E-03	1.E+00	1.E+00
GO:0001656	metanephros development	BP	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E-60	8.E-82	1.E+00	1.E+00
GO:0001657	ureteric bud development	BP	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	3.E-69	7.E-87	1.E+00	1.E+00
GO:0003680	AT DNA binding	MF	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	2.E-43	2.E-16	1.E+00	1.E+00
GO:0004716	receptor signaling protein tyrosine kinase activity	MF	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	5.E-35	9.E-32	1.E+00	1.E+00
GO:0007156	homophilic cell adhesion	BP	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	5.E-52	4.E-69	1.E+00	1.E+00
GO:0007275	multicellular organismal development	BP	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	2.E-26	3.E-23	1.E+00	1.E+00
GO:0007399	nervous system development	BP	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	2.E-28	5.E-44	1.E+00	1.E+00
GO:0007411	axon guidance	BP	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	5.E-97	5.E-100	1.E+00	1.E+00
GO:0007420	brain development	BP	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	2.E-29	1.E-48	1.E+00	1.E+00
GO:0008046	axon guidance receptor activity	MF	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E-193	2.E-194	1.E+00	1.E+00
GO:0009986	cell surface	CC	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	6.E-69	1.E-105	1.E+00	1.E+00
GO:0017154	semaphorin receptor activity	MF	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	6.E-39	5.E-20	1.E+00	1.E+00
GO:0030154	cell differentiation	BP	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	4.E-38	3.E-44	1.E+00	1.E+00
GO:0030673	axolemma	CC	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	3.E-231	4.E-221	1.E+00	1.E+00
GO:0042802	identical protein binding	MF	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	5.E-46	1.E-83	1.E+00	1.E+00
GO:0050772	positive regulation of axonogenesis	BP	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	4.E-184	4.E-181	1.E+00	1.E+00
GO:0007155	cell adhesion	BP	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	5.E-10	8.E-25	1.E+00	1.E+00
GO:0005021	vascular endothelial growth factor receptor activity	MF	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	5.E-27	8.E-10	1.E+00	1.E+00
GO:0031032	actomyosin structure organization and biogenesis	BP	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	2.E-12	4.E-08	1.E+00	1.E+00
GO:0004714	transmembrane receptor protein tyrosine kinase activ	MF	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E-08	2.E-07	1.E+00	1.E+00
GO:0006487	protein amino acid N-linked glycosylation	BP	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	8.E-07	4.E-17	1.E+00	1.E+00
GO:0008113	peptide-methionine-(S)-S-oxide reductase activity	MF	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	4.E-12	5.E-06	1.E+00	1.E+00
GO:0008093	cytoskeletal adaptor activity	MF	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	7.E-07	2.E-05	1.E+00	1.E+00
GO:0007417	central nervous system development	BP	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	4.E-04	1.E-12	1.E+00	1.E+00
GO:0015247	aminophospholipid transporter activity	MF	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	4.E-04	9.E-06	1.E+00	1.E+00
GO:0019829	cation-transporting ATPase activity	MF	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	4.E-04	9.E-06	1.E+00	1.E+00
GO:0007169	transmembrane receptor protein tyrosine kinase signa	BP	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	2.E-03	2.E-07	1.E+00	1.E+00
GO:0015204	urea transmembrane transporter activity	MF	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	2.E-03	1.E-05	1.E+00	1.E+00
GO:0015840	urea transport	BP	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	2.E-03	1.E-05	1.E+00	1.E+00
GO:0008533	astacin activity	MF	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	2.E-08	2.E-03	1.E+00	1.E+00
GO:0005923	tight junction	CC	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E-05	2.E-03	1.E+00	1.E+00
GO:0030165	PDZ domain binding	MF	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	2.E-03	1.E-03	1.E+00	1.E+00
GO:0008510	sodium:bicarbonate symporter activity	MF	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	3.E-03	5.E-04	1.E+00	1.E+00
GO:0004152	dihydroorotate dehydrogenase activity	MF	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	4.E-04	1.E-02	1.E+00	1.E+00
GO:0004158	dihydroorotate oxidase activity	MF	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	4.E-04	1.E-02	1.E+00	1.E+00
GO:0007507	heart development	BP	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E-05	2.E-02	1.E+00	1.E+00

GO:0019229	regulation of vasoconstriction	BP	1.E+00	1.E+00	1.E-02	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00
GO:0016176	superoxide-generating NADPH oxidase activator activi	MF	1.E+00	1.E+00	1.E-02	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00
GO:0006506	GPI anchor biosynthetic process	BP	1.E+00	1.E+00	1.E-02	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00
GO:0009582	detection of abiotic stimulus	BP	1.E+00	1.E+00	1.E-02	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00
GO:0004919	interleukin-9 receptor activity	MF	1.E+00	1.E+00	1.E-02	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00
GO:0050715	positive regulation of cytokine secretion	BP	1.E+00	1.E+00	1.E-02	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00
GO:0004060	arylamine N-acetyltransferase activity	MF	1.E+00	1.E+00	1.E-02	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00
GO:0015198	oligopeptide transporter activity	MF	1.E+00	1.E+00	1.E-02	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00
GO:0050832	defense response to fungus	BP	1.E+00	1.E+00	2.E-02	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00
GO:0045263	proton-transporting ATP synthase complex, coupling fi	CC	1.E+00	1.E+00	2.E-02	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00
GO:0006183	GTP biosynthetic process	BP	1.E+00	1.E+00	2.E-02	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00
GO:0051260	protein homoooligomerization	BP	1.E+00	1.E+00	2.E-02	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00
GO:0000747	conjugation with cellular fusion	BP	1.E+00	1.E+00	2.E-02	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00
GO:0042289	MHC class II protein binding	MF	1.E+00	1.E+00	2.E-02	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00
GO:0019861	flagellum	CC	1.E+00	1.E+00	2.E-02	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00
GO:0004964	lutropin-choriogonadotropic hormone receptor activit	MF	1.E+00	1.E+00	2.E-02	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00
GO:0004238	mepirin A activity	MF	1.E+00	1.E+00	3.E-02	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00
GO:0043367	CD4-positive, alpha beta T cell differentiation	BP	1.E+00	1.E+00	3.E-02	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00
GO:0043374	CD8-positive, alpha-beta T cell differentiation	BP	1.E+00	1.E+00	3.E-02	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00
GO:0016712	oxidoreductase activity, acting on paired donors, with	MF	1.E+00	1.E+00	3.E-02	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00
GO:0004574	oligo-1,6-glucosidase activity	MF	1.E+00	1.E+00	3.E-02	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00
GO:0004575	sucrose alpha-glucosidase activity	MF	1.E+00	1.E+00	3.E-02	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00
GO:0015833	peptide transport	BP	1.E+00	1.E+00	3.E-02	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00
GO:0006633	fatty acid biosynthetic process	BP	1.E+00	1.E+00	4.E-02	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00
GO:0007586	digestion	BP	1.E+00	1.E+00	4.E-02	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00
GO:0006525	arginine metabolic process	BP	1.E+00	1.E+00	4.E-02	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00
GO:0005223	intracellular cGMP activated cation channel activity	MF	1.E+00	1.E+00	4.E-02	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00
GO:0006824	cobalt ion transport	BP	1.E+00	1.E+00	4.E-02	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00
GO:0015087	cobalt ion transmembrane transporter activity	MF	1.E+00	1.E+00	4.E-02	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00
GO:0030199	collagen fibril organization	BP	1.E+00	1.E+00	5.E-02	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00
GO:0001772	immunological synapse	CC	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	2.E-02	1.E+00	9.E-02	1.E+00	1.E+00	1.E+00
GO:0005588	collagen type V	CC	1.E+00	1.E+00	5.E-01	1.E+00	1.E+00	3.E-06	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00
GO:0045860	positive regulation of protein kinase activity	BP	1.E+00	1.E+00	6.E-01	1.E+00	1.E+00	2.E-09	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00
GO:0000724	double-strand break repair via homologous recombina	BP	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	5.E-19	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00
GO:0005587	collagen type IV	CC	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	2.E-33	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00
GO:0042062	long-term strengthening of neuromuscular junction	BP	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	3.E-27	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00
GO:0007090	regulation of S phase of mitotic cell cycle	BP	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	3.E-14	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00
GO:0016072	rRNA metabolic process	BP	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	3.E-13	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00
GO:0003697	single-stranded DNA binding	MF	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	5.E-13	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00
GO:0030957	Tat protein binding	MF	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	7.E-13	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00
GO:0042255	ribosome assembly	BP	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	7.E-13	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00
GO:0001915	negative regulation of T cell mediated cytotoxicity	BP	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	3.E-11	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00
GO:0001960	negative regulation of cytokine and chemokine mediat	BP	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	3.E-11	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00
GO:0050857	positive regulation of antigen receptor-mediated signa	BP	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	3.E-11	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00
GO:0004364	glutathione transferase activity	MF	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E-10	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00
GO:0005539	glycosaminoglycan binding	MF	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	3.E-08	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00
GO:0004126	cytidine deaminase activity	MF	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	6.E-08	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00
GO:0051209	release of sequestered calcium ion into cytosol	BP	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	9.E-08	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00
GO:0045869	negative regulation of retroviral genome replication	BP	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E-07	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00
GO:0030109	HLA-B specific inhibitory MHC class I receptor activity	MF	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E-07	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00
GO:0030529	ribonucleoprotein complex	CC	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	2.E-07	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00
GO:0000722	telomere maintenance via recombination	BP	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	2.E-07	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00
GO:0051607	defense response to virus	BP	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E-06	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00
GO:0050853	B cell receptor signaling pathway	BP	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	2.E-06	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00
GO:0004572	mannosyl-oligosaccharide 1,3-1,6-alpha-mannosidase	MF	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	2.E-06	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00
GO:0006642	triacylglycerol mobilization	BP	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	4.E-06	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00
GO:0003735	structural constituent of ribosome	MF	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	8.E-06	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00
GO:0007093	mitotic cell cycle checkpoint	BP	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	8.E-06	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00
GO:0015934	large ribosomal subunit	CC	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	9.E-06	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00
GO:0050852	T cell receptor signaling pathway	BP	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	9.E-06	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00
GO:0008037	cell recognition	BP	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	9.E-06	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00
GO:0016314	phosphatidylinositol-3,4,5-trisphosphate 3-phosphata:	MF	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	2.E-05	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00
GO:0006376	mRNA splice site selection	BP	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	4.E-05	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00
GO:0015643	toxin binding	MF	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	6.E-05	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00
GO:0030141	secretory granule	CC	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E-04	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00
GO:0016554	cytidine to uridine editing	BP	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	3.E-04	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00
GO:0005654	nucleoplasm	CC	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	4.E-04	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00
GO:0030217	T cell differentiation	BP	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	4.E-04	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00
GO:0007030	Golgi organization and biogenesis	BP	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	6.E-04	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00
GO:0007067	mitosis	BP	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	8.E-04	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00
GO:0003720	telomerase activity	MF	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E-03	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00
GO:0005697	telomerase holoenzyme complex	CC	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E-03	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00
GO:0007004	telomere maintenance via telomerase	BP	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	2.E-03	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00
GO:0051726	regulation of cell cycle	BP	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	2.E-03	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00

GO:0005795	Golgi stack	CC	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	2.E-03	1.E+00	1.E+00	1.E+00	1.E+00
GO:0005840	ribosome	CC	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	2.E-03	1.E+00	1.E+00	1.E+00	1.E+00
GO:0046329	negative regulation of JNK cascade	BP	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	3.E-03	1.E+00	1.E+00	1.E+00	1.E+00
GO:0016198	axon choice point recognition	BP	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	4.E-03	1.E+00	1.E+00	1.E+00	1.E+00
GO:0042627	chylomicron	CC	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	4.E-03	1.E+00	1.E+00	1.E+00	1.E+00
GO:0030546	receptor activator activity	MF	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	4.E-03	1.E+00	1.E+00	1.E+00	1.E+00
GO:0008396	oxysterol 7-alpha-hydroxylase activity	MF	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	5.E-03	1.E+00	1.E+00	1.E+00	1.E+00
GO:0004014	adenosylmethionine decarboxylase activity	MF	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	6.E-03	1.E+00	1.E+00	1.E+00	1.E+00
GO:0006597	spermine biosynthetic process	BP	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	6.E-03	1.E+00	1.E+00	1.E+00	1.E+00
GO:0045861	negative regulation of proteolysis	BP	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	7.E-03	1.E+00	1.E+00	1.E+00	1.E+00
GO:0016311	dephosphorylation	BP	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	7.E-03	1.E+00	1.E+00	1.E+00	1.E+00
GO:0047497	mitochondrion transport along microtubule	BP	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	8.E-03	1.E+00	1.E+00	1.E+00	1.E+00
GO:0045948	positive regulation of translational initiation	BP	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E-02	1.E+00	1.E+00	1.E+00	1.E+00
GO:0006622	protein targeting to lysosome	BP	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E-02	1.E+00	1.E+00	1.E+00	1.E+00
GO:0035162	embryonic hemopoiesis	BP	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E-02	1.E+00	1.E+00	1.E+00	1.E+00
GO:0000213	tRNA-intron endonuclease activity	MF	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	2.E-02	1.E+00	1.E+00	1.E+00	1.E+00
GO:0000214	tRNA-intron endonuclease complex	CC	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	2.E-02	1.E+00	1.E+00	1.E+00	1.E+00
GO:0006777	Mo-molybdopterin cofactor biosynthetic process	BP	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	2.E-02	1.E+00	1.E+00	1.E+00	1.E+00
GO:0001660	fever	BP	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	2.E-02	1.E+00	1.E+00	1.E+00	1.E+00
GO:0003711	transcription elongation regulator activity	MF	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	3.E-02	1.E+00	1.E+00	1.E+00	1.E+00
GO:0020037	heme binding	MF	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	3.E-02	1.E+00	1.E+00	1.E+00	1.E+00
GO:0030917	midbrain-hindbrain boundary development	BP	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	3.E-02	1.E+00	1.E+00	1.E+00	1.E+00
GO:00016363	nuclear matrix	CC	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	3.E-02	1.E+00	1.E+00	1.E+00	1.E+00
GO:0031575	G1/S transition checkpoint	BP	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	3.E-02	1.E+00	1.E+00	1.E+00	1.E+00
GO:0042405	nuclear inclusion body	CC	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	3.E-02	1.E+00	1.E+00	1.E+00	1.E+00
GO:0050868	negative regulation of T cell activation	BP	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	4.E-02	1.E+00	1.E+00	1.E+00	1.E+00
GO:0004277	granzyme A activity	MF	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	4.E-02	1.E+00	1.E+00	1.E+00	1.E+00
GO:0006922	cleavage of lamin	BP	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	4.E-02	1.E+00	1.E+00	1.E+00	1.E+00
GO:0005792	microsome	CC	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	4.E-02	1.E+00	1.E+00	1.E+00	1.E+00
GO:0006885	regulation of pH	BP	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	4.E-02	1.E+00	1.E+00	1.E+00	1.E+00
GO:0007286	spermatid development	BP	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	4.E-02	1.E+00	1.E+00	1.E+00	1.E+00
GO:0008023	transcription elongation factor complex	CC	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	4.E-02	1.E+00	1.E+00	1.E+00	1.E+00
GO:0003896	DNA primase activity	MF	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	5.E-02	1.E+00	1.E+00	1.E+00	1.E+00
GO:0015012	heparan sulfate proteoglycan biosynthetic process	BP	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	2.E-03	5.E-02	1.E+00	1.E+00	1.E+00
GO:0030507	spectrin binding	MF	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	7.E-04	6.E-02	1.E+00	1.E+00	1.E+00
GO:0045199	maintenance of epithelial cell polarity	BP	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	7.E-04	6.E-02	1.E+00	1.E+00	1.E+00
GO:0019221	cytokine and chemokine mediated signaling pathway	BP	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	5.E-05	7.E-02	1.E+00	1.E+00	1.E+00
GO:0016573	histone acetylation	BP	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	9.E-04	8.E-02	1.E+00	1.E+00	1.E+00
GO:0006207	'de novo' pyrimidine base biosynthetic process	BP	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	2.E-02	1.E-01	1.E+00	1.E+00	1.E+00
GO:0004614	phosphoglucomutase activity	MF	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	5.E-03	2.E-01	1.E+00	1.E+00	1.E+00
GO:0005020	stem cell factor receptor activity	MF	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	3.E-05	4.E-01	1.E+00	1.E+00	1.E+00
GO:0051539	4 iron, 4 sulfur cluster binding	MF	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E-03	4.E-01	1.E+00	1.E+00	1.E+00
GO:0004968	gonadotropin-releasing hormone receptor activity	MF	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	7.E-03	5.E-01	1.E+00	1.E+00	1.E+00
GO:0005018	platelet-derived growth factor alpha-receptor activity	MF	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	5.E-07	7.E-01	1.E+00	1.E+00	1.E+00
GO:0048407	platelet-derived growth factor binding	MF	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	5.E-07	7.E-01	1.E+00	1.E+00	1.E+00
GO:0004610	phosphoacetylglucosamine mutase activity	MF	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	6.E-05	7.E-01	1.E+00	1.E+00	1.E+00
GO:0006041	glucosamine metabolic process	BP	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	6.E-05	7.E-01	1.E+00	1.E+00	1.E+00
GO:0019255	glucose 1-phosphate metabolic process	BP	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	6.E-05	7.E-01	1.E+00	1.E+00	1.E+00
GO:0019834	phospholipase A2 inhibitor activity	MF	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	3.E-10	1.E+00	1.E+00	1.E+00	1.E+00
GO:0008486	diphosphoinositol-polyphosphate diphosphatase activ	MF	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	5.E-10	1.E+00	1.E+00	1.E+00	1.E+00
GO:0003727	single-stranded RNA binding	MF	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	3.E-08	1.E+00	1.E+00	1.E+00	1.E+00
GO:0007568	aging	BP	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	8.E-08	1.E+00	1.E+00	1.E+00	1.E+00
GO:0047536	2-aminoadipate transaminase activity	MF	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	3.E-07	1.E+00	1.E+00	1.E+00	1.E+00
GO:0004473	malate dehydrogenase (oxaloacetate-decarboxylating)	MF	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	3.E-06	1.E+00	1.E+00	1.E+00	1.E+00
GO:0006741	NADP biosynthetic process	BP	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	3.E-06	1.E+00	1.E+00	1.E+00	1.E+00
GO:0009743	response to carbohydrate stimulus	BP	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	3.E-06	1.E+00	1.E+00	1.E+00	1.E+00
GO:0006108	malate metabolic process	BP	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	5.E-06	1.E+00	1.E+00	1.E+00	1.E+00
GO:0004370	glycerol kinase activity	MF	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	2.E-05	1.E+00	1.E+00	1.E+00	1.E+00
GO:0000133	polarisome	CC	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	5.E-05	1.E+00	1.E+00	1.E+00	1.E+00
GO:0004700	atypical protein kinase C activity	MF	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	5.E-05	1.E+00	1.E+00	1.E+00	1.E+00
GO:0045216	intercellular junction assembly and maintenance	BP	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	5.E-05	1.E+00	1.E+00	1.E+00	1.E+00
GO:0005639	integral to nuclear inner membrane	CC	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	6.E-05	1.E+00	1.E+00	1.E+00	1.E+00
GO:0048665	neuron fate specification	BP	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	2.E-04	1.E+00	1.E+00	1.E+00	1.E+00
GO:0008449	N-acetylglucosamine-6-sulfatase activity	MF	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	2.E-04	1.E+00	1.E+00	1.E+00	1.E+00
GO:0004470	malic enzyme activity	MF	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	2.E-04	1.E+00	1.E+00	1.E+00	1.E+00
GO:0004692	cGMP-dependent protein kinase activity	MF	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	8.E-04	1.E+00	1.E+00	1.E+00	1.E+00
GO:0043325	phosphatidylinositol-3,4-bisphosphate binding	MF	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E-03	1.E+00	1.E+00	1.E+00	1.E+00
GO:0050509	N-acetylglucosaminyl-proteoglycan 4-beta-glucuronos	MF	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	3.E-03	1.E+00	1.E+00	1.E+00	1.E+00
GO:0005547	phosphatidylinositol-3,4,5-triphosphate binding	MF	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	3.E-03	1.E+00	1.E+00	1.E+00	1.E+00
GO:0010001	glial cell differentiation	BP	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	5.E-03	1.E+00	1.E+00	1.E+00	1.E+00
GO:0045663	positive regulation of myoblast differentiation	BP	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	6.E-03	1.E+00	1.E+00	1.E+00	1.E+00
GO:0016616	oxidoreductase activity, acting on the CH-OH group of	MF	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	6.E-03	1.E+00	1.E+00	1.E+00	1.E+00
GO:0006072	glycerol-3-phosphate metabolic process	BP	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	6.E-03	1.E+00	1.E+00	1.E+00	1.E+00
GO:0004653	polypeptide N-acetylglactosaminyltransferase activity	MF	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	6.E-03	1.E+00	1.E+00	1.E+00	1.E+00

GO:0008543	fibroblast growth factor receptor signaling pathway	BP	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E-02	1.E+00	1.E+00	1.E+00
GO:0030539	male genitalia development	BP	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	2.E-02	1.E+00	1.E+00	1.E+00
GO:0006024	glycosaminoglycan biosynthetic process	BP	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	3.E-02	1.E+00	1.E+00	1.E+00
GO:0004859	phospholipase inhibitor activity	MF	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	5.E-02	1.E+00	1.E+00	1.E+00
GO:0005487	nucleocytoplasmic transporter activity	MF	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E-01	1.E-02	1.E+00	1.E+00
GO:0004713	protein-tyrosine kinase activity	MF	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E-01	6.E-04	1.E+00	1.E+00
GO:0015917	aminophospholipid transport	BP	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	3.E-01	2.E-03	1.E+00	1.E+00
GO:0005887	integral to plasma membrane	CC	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	5.E-12	1.E+00	1.E+00
GO:0005515	protein binding	MF	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	3.E-09	1.E+00	1.E+00
GO:0006820	anion transport	BP	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	7.E-05	1.E+00	1.E+00
GO:0005834	heterotrimeric G-protein complex	CC	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	2.E-04	1.E+00	1.E+00
GO:0018107	peptidyl-threonine phosphorylation	BP	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E-03	1.E+00	1.E+00
GO:0043508	negative regulation of JNK activity	BP	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E-03	1.E+00	1.E+00
GO:0015380	anion exchanger activity	MF	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	2.E-03	1.E+00	1.E+00
GO:0005452	inorganic anion exchanger activity	MF	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	2.E-03	1.E+00	1.E+00
GO:0045749	negative regulation of S phase of mitotic cell cycle	BP	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	2.E-02	1.E+00	1.E+00
GO:0006904	vesicle docking during exocytosis	BP	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	2.E-02	1.E+00	1.E+00
GO:0030097	hemopoiesis	BP	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	3.E-02	1.E+00	1.E+00
GO:0004012	phospholipid-translocating ATPase activity	MF	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	5.E-02	1.E+00	1.E+00
GO:0050662	coenzyme binding	MF	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	1.E+00	2.E-01	3.E-03

Supplementary Figures

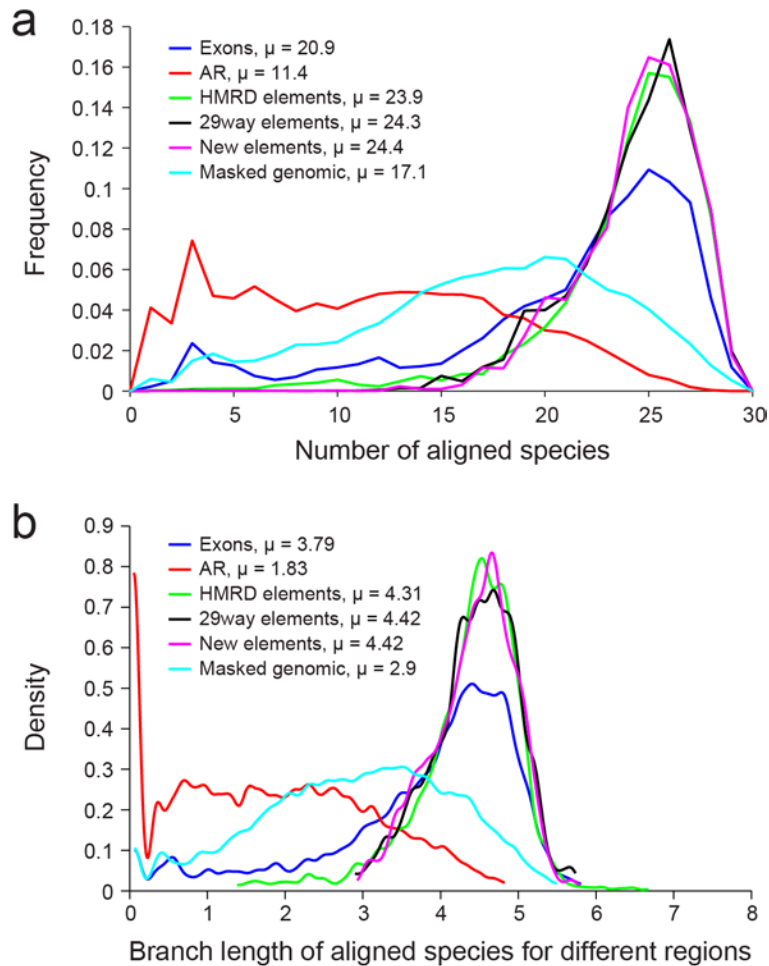


Figure S1 - Evolutionary rate and depth of the 29 mammals multiple alignment in different features in the human genome. a, The number of aligned species for each base in the genome is reported for different features such as neutrally evolving repeats (AR, in red), the whole genome (light blue), exons (dark blue) and non-coding conserved elements (green=top 5% of HMRD elements, black=29mammals elements, purple=newly detected bases). Note that the numbers of aligned species increases with the functional importance of each feature, suggesting that the power is highest over functional elements. **b,** The evolutionary depth for each base in the genome is reported for different features such as neutrally evolving repeats (AR, in red), the whole genome (light blue), exons (dark blue) and non-coding conserved elements (green=top 5% of HMRD elements, black=29mammals elements, purple=newly detected bases). Note that the evolutionary depth increases with the functional importance of each feature, suggesting that the power is highest over functional elements.

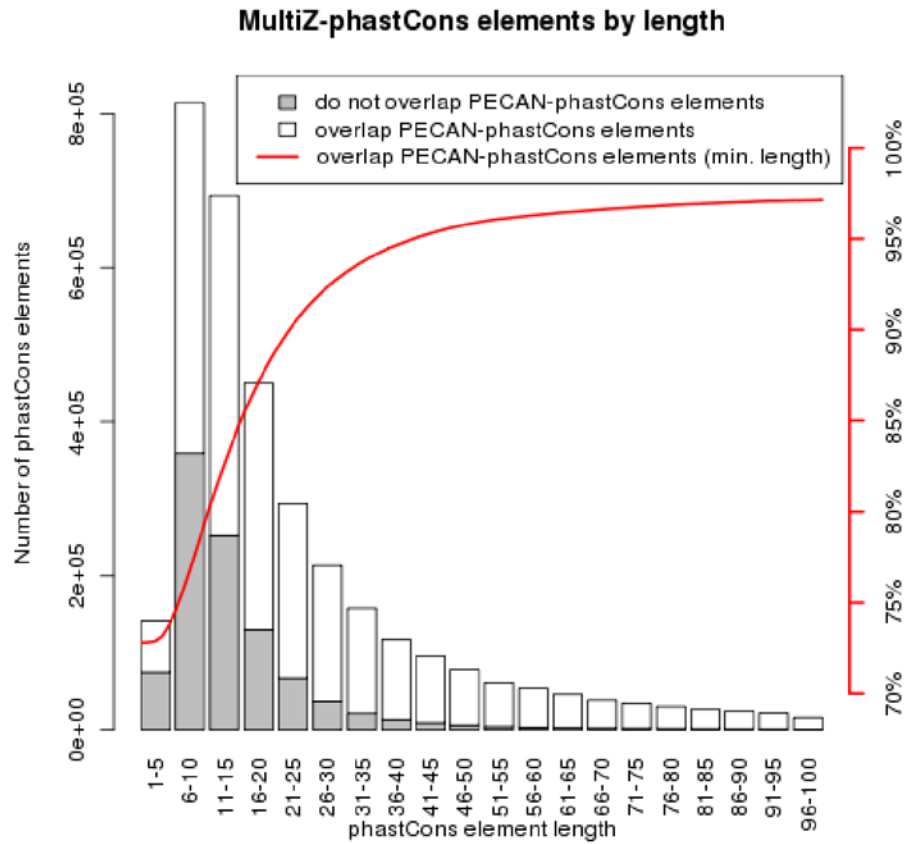
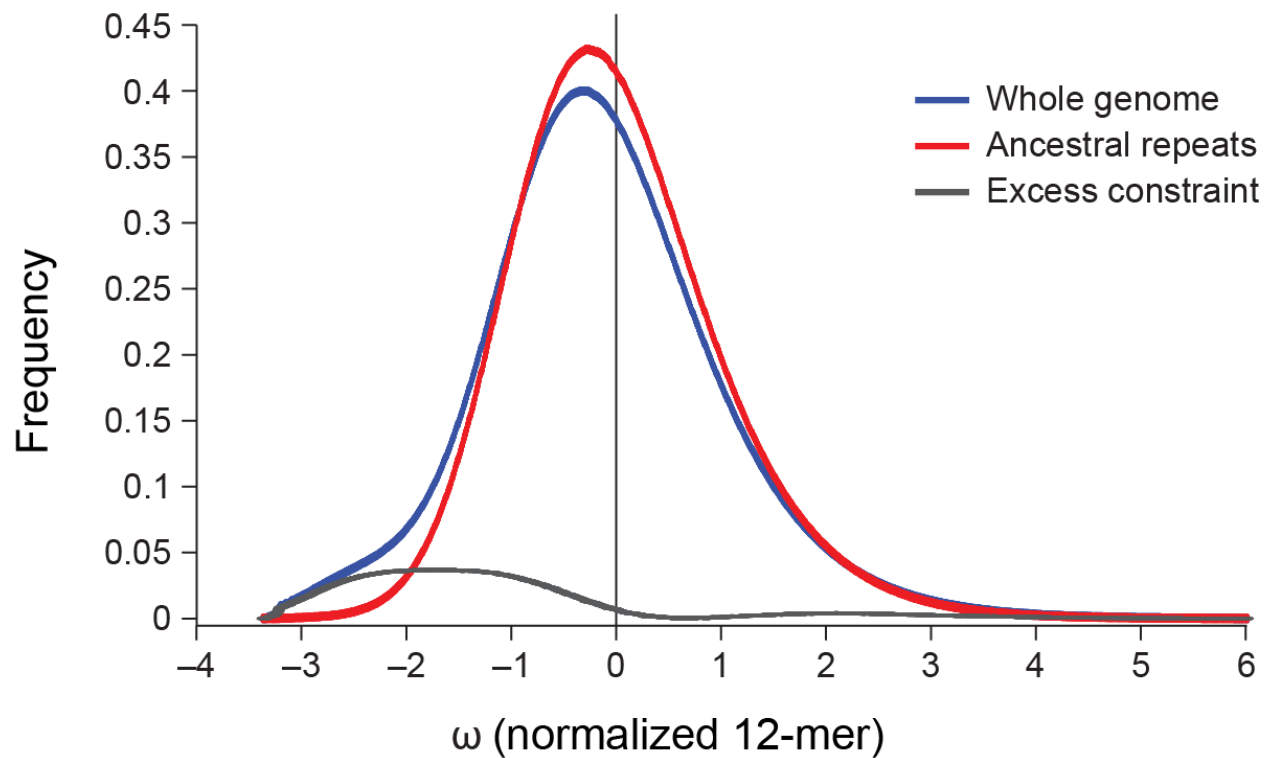
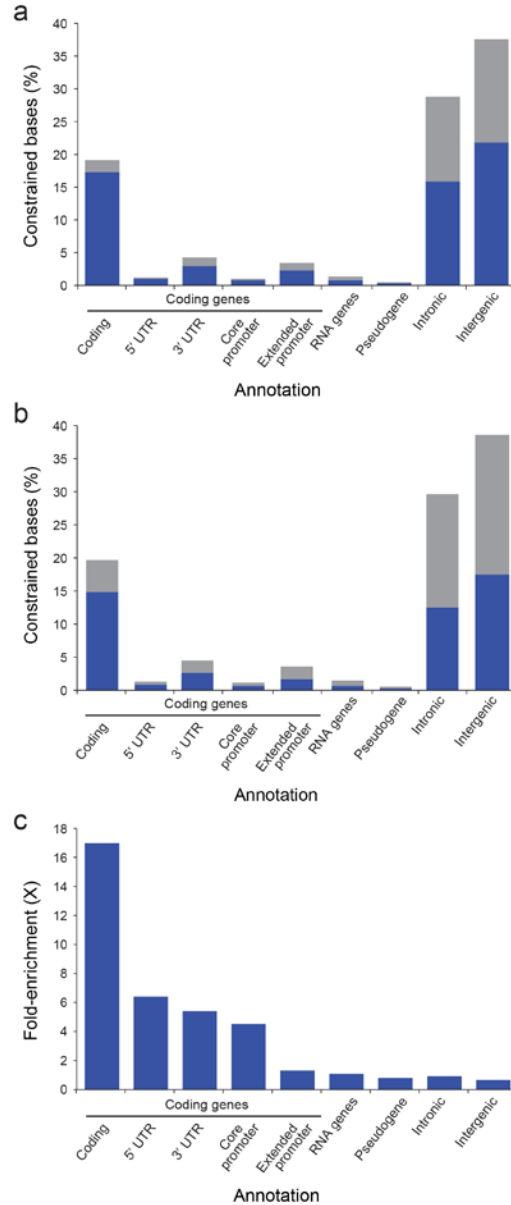


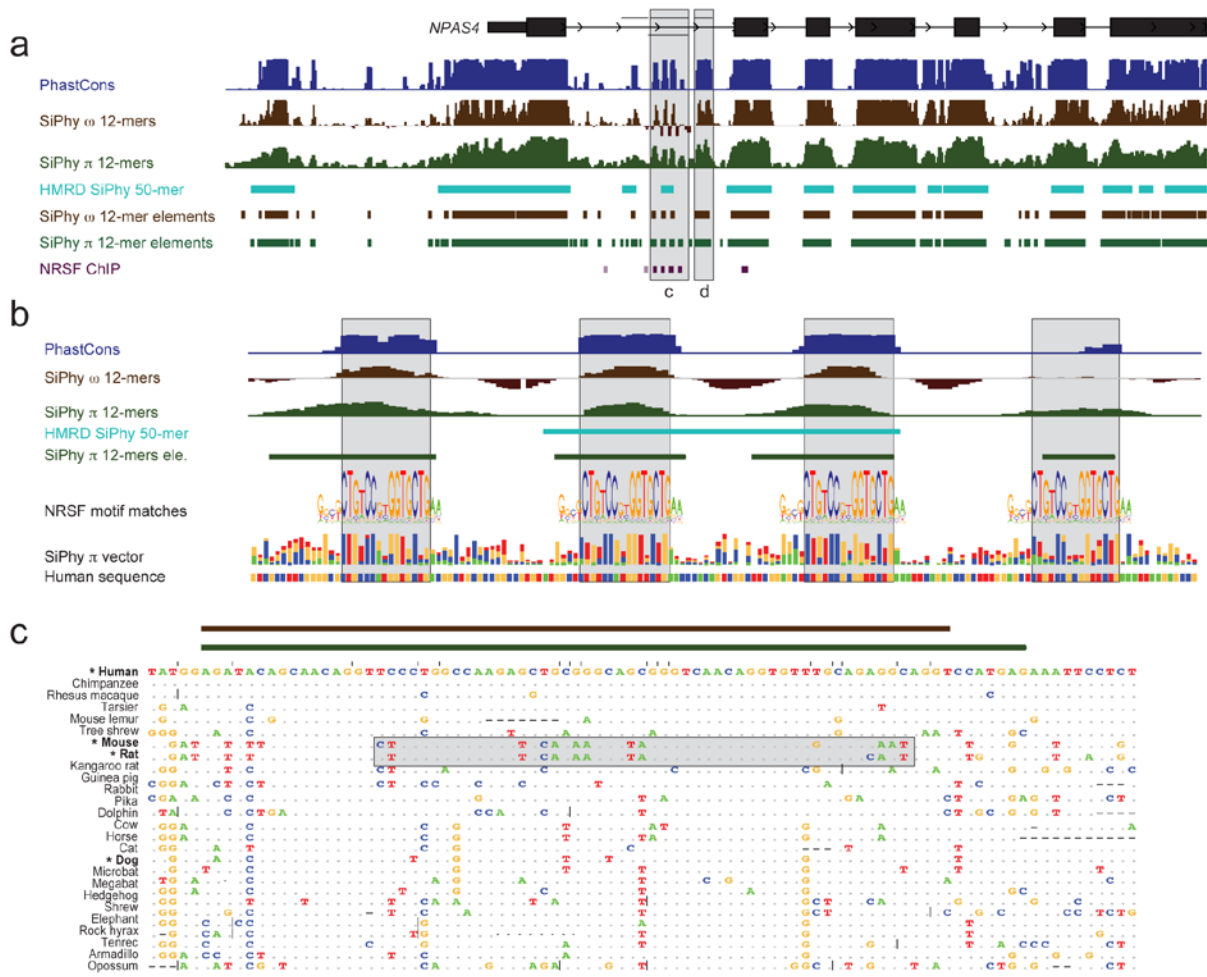
Figure S2 - Comparison between phastCons elements in MultiZ and in PECAN alignments. Histogram of MultiZ-phastCons elements by size, divided into the elements that overlap a PECAN-phastCons element (white bars) and those that do not (grey bars). Also shown in red, the percentage of MultiZ-phastCons elements in agreement with a Pecan-phastCons element as a function of the minimum length of the elements. The agreement between the elements is defined as an overlap of at least 1 nucleotide.



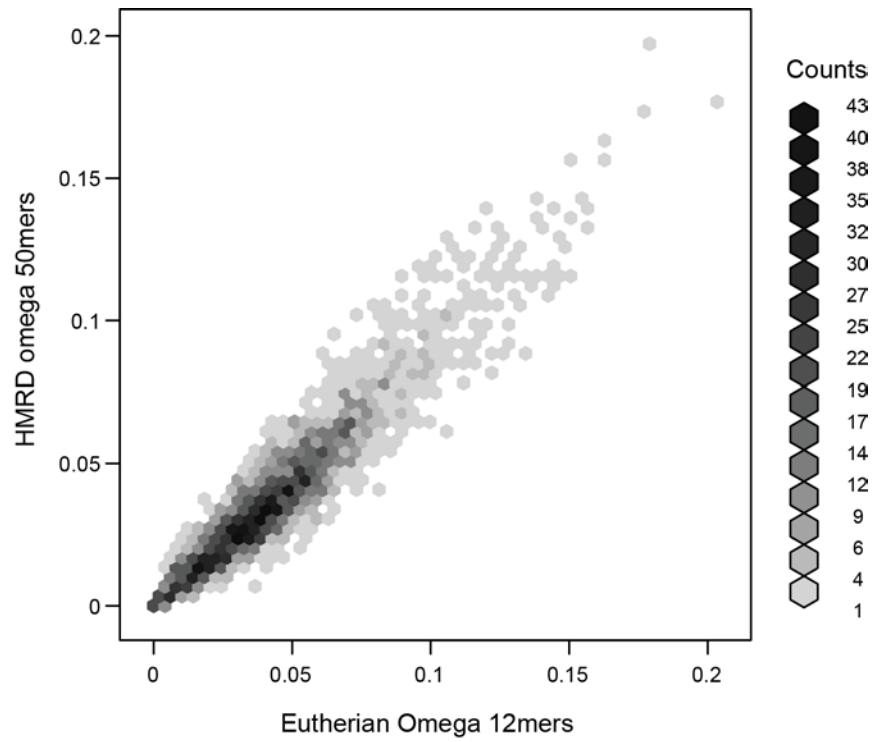
Supplementary Figure S3 - Estimation and detection of constraint. Roughly 5.5% of the genome is estimated to be under constraint using SiPhy- ω with 12-bp windows. The constraint score for ancient repeats is shown in red and the whole genome in blue.



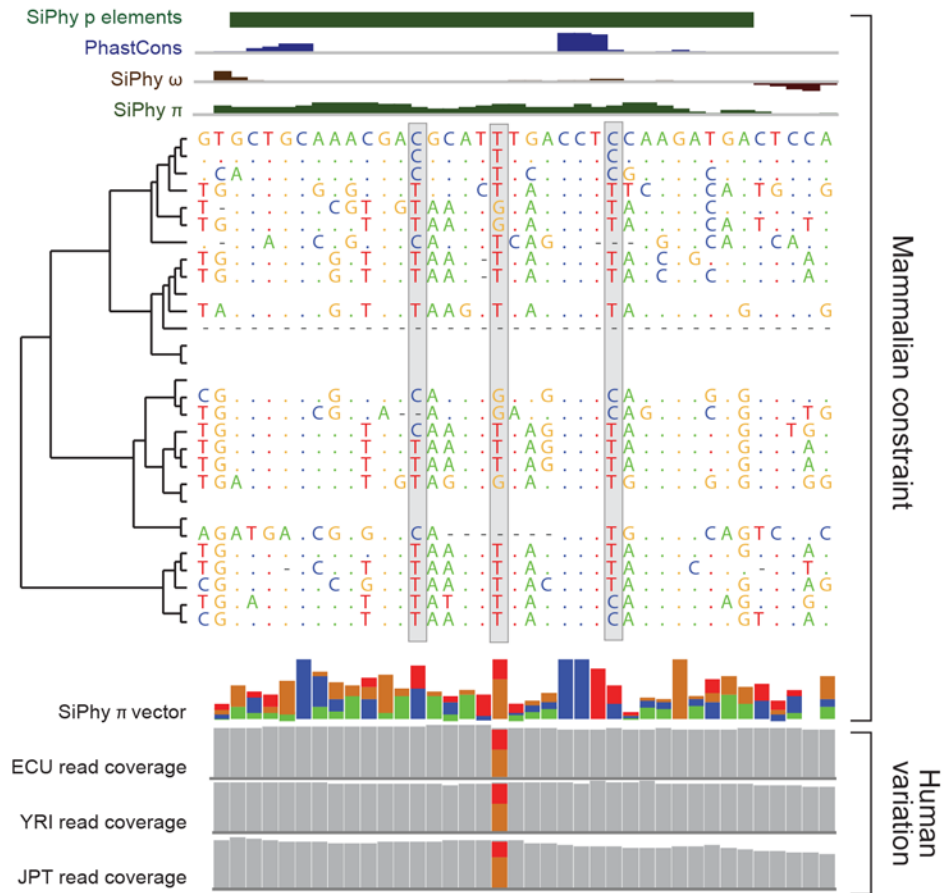
Supplementary Figure S4 - Identification of constrained SiPhy- ω elements. At 10% FDR, 3.6 million constrained SiPhy- ω elements (12-bp windows) can be detected. The largest fraction of constraint can be seen in coding exons, introns and intergenic regions. To account for each constrained base uniquely, the analysis was performed hierarchically as follows: coding exons, 5'-UTRs, 3'-UTRs, promoters, pseudogenes, non-coding RNAs, introns, intergenic. **a**, Distribution of constraint with overlap with HMRD 50 bp + Siepel vertebrate elements shown in blue. **b**, Distribution of constraint with overlap with HMRD elements only shown in blue. Please note the similarity of these data sets. **c**, The 29 mammals constrained bases are particularly enriched in coding transcripts and their promoters. The enrichment was generated by comparing the fraction of constrained bases to the total number of bases in the specific genomic annotation.



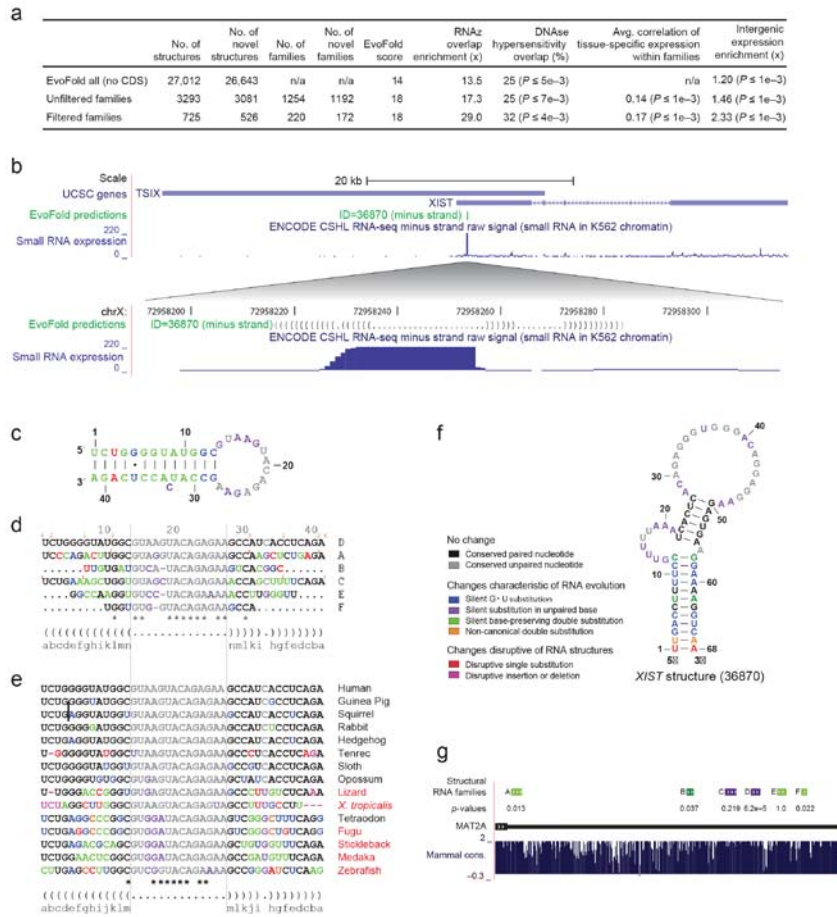
Supplementary Figure S5 - Estimation of fraction of constraint and identification of four NRSF-binding sites in *NPAS4*. The neurological gene *NPAS4* has many constrained elements overlapping introns and the upstream intergenic region. Note that the shaded box c contained only one constrained element using HMRD, while analysis of 29 mammalian sequences reveals four smaller elements. **b**, These four constrained elements in the first intron correspond to binding sites for the NRSF transcription factor, known to regulate neuronal lineages. **c**, Another 70 bp constrained element in the first intron, marked as shaded box d in panel b, was not detected in the HMRD analysis due to unusually high divergence in mouse and rat, but is highly constrained in all other mammals and was therefore detected with sequences from 29 mammals.



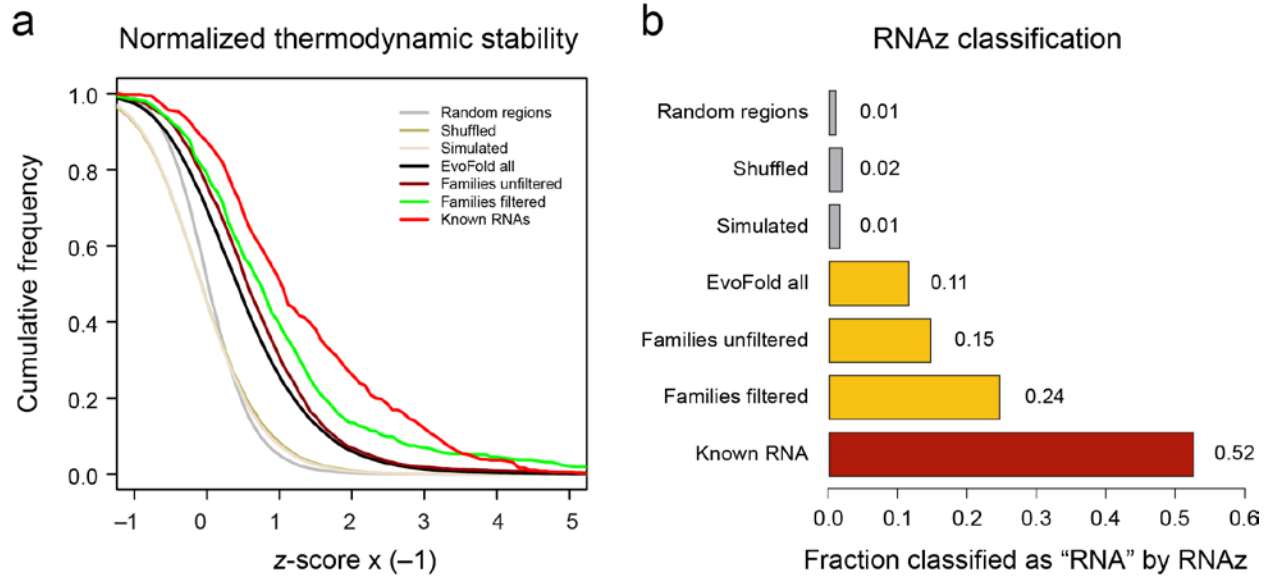
Supplementary Figure S6 - Correlation between element density in 29 mammals and HMRD data sets. The correlation for each megabase in the genome between 29-way eutherian 12mer based element density and HMRD 50mer element density was computed and plotted.



Supplementary Figure S7. Biased nucleotide substitution patterns identifies positions where two bases appear equally constrained and correlating with SNPs in the human population. An example of an intergenic SiPhy- π element (HG18 chr12:1,916,342-1,916,380) detected based on the presence of three 2-fold degenerate constrained bases. Note how these bases (in grey boxes) alternate between bases across the evolutionary tree. One of the degenerate bases matches a SNP present in several human populations, European CEPF (ECU), Yoruban Africans (YRI) and Japanese (JPT).

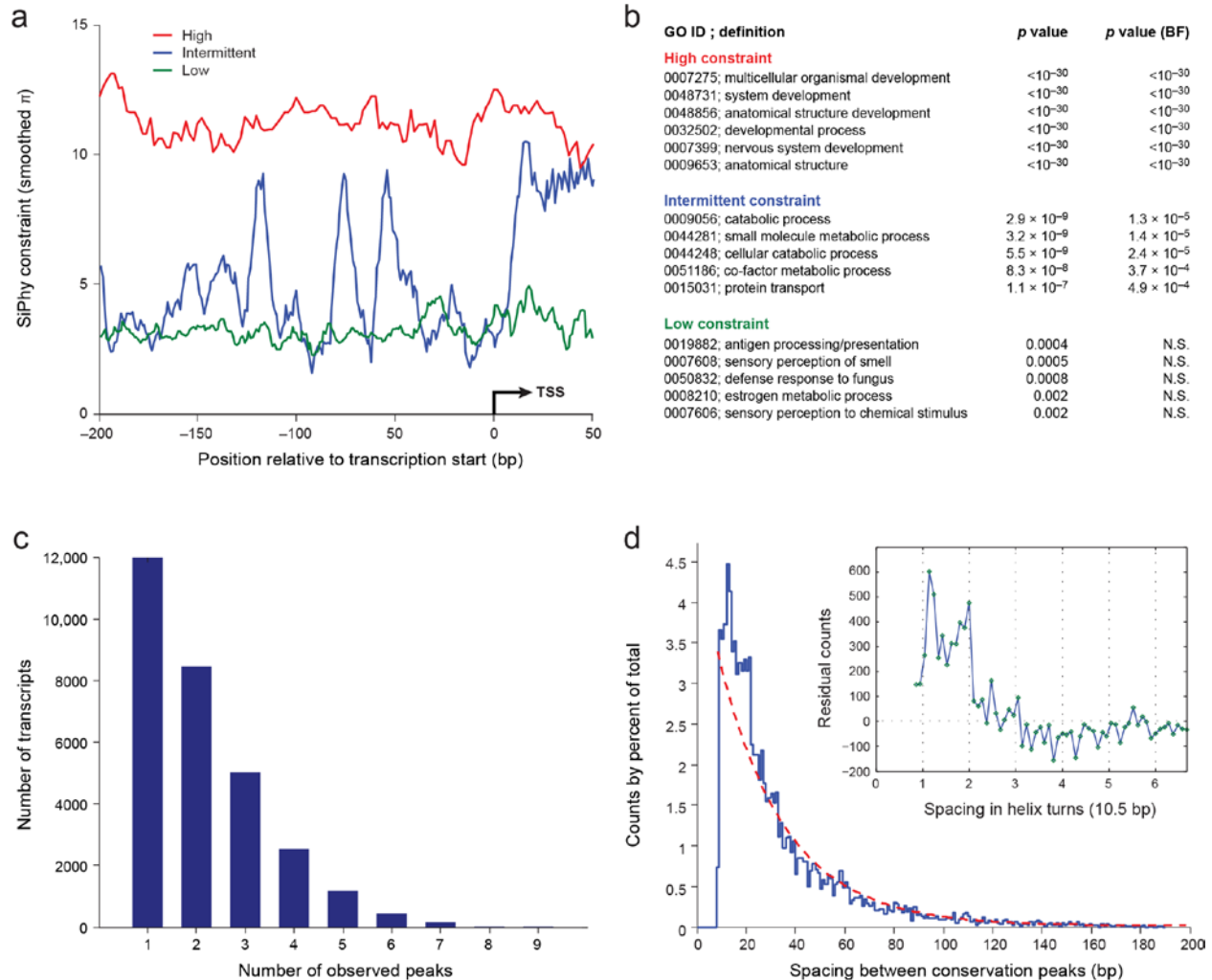


Supplementary Figure S9 - Over 200 families of potential RNA structural elements were identified. **a**, Summary and enrichment statistics for structural RNA predictions sets. **b**, Top: The RNA structure (green) is predicted on the *XIST* strand (purple) and overlaps short RNAs (blue) observed at high abundance in the chromatin cellular compartment. Bottom: base-level resolution showing secondary structure in parenthesis notation (black) and the demarcated expression of short RNAs (blue). **c**, Predicted secondary structure for one of six hairpins (hairpin D) in the 3'UTR of the *MAT2A* gene, responsible for the synthesis of S-adenosylmethionine (SAM), the primary methyl donor in human cells. **d**, The human sequences of all six hairpins were aligned using hairpin D as the reference. Insertions relative to D are shown with orange bars and numbers. Fully conserved positions (*) between the human sequences reveal the same loop region motif. **e** Multiple alignment across vertebrates for hairpin D. **f**, Secondary structure drawing of *XIST* structure with color-coding of substitution evidence (Black= Conserved paired nucleotide, grey= Conserved unpaired nucleotide; blue= Silent G • U substitution; purple= Silent substitution in unpaired base, green = Silent base preserving substitution, Orange= Non-canonical double substitution, Red= Disruptive single substitution, Pink = Disruptive insertion or deletion). **g**, Family of hairpins in 3'UTR of the *MAT2A* gene, responsible for the synthesis of S-adenosylmethionine (SAM), the primary methyl donor in human cells. Purple= initial family members, dark green= family member after paralog search, light green= additional members found by dedicated additional paralog search, with P-values for substitution evidence in species not used for structure inference.

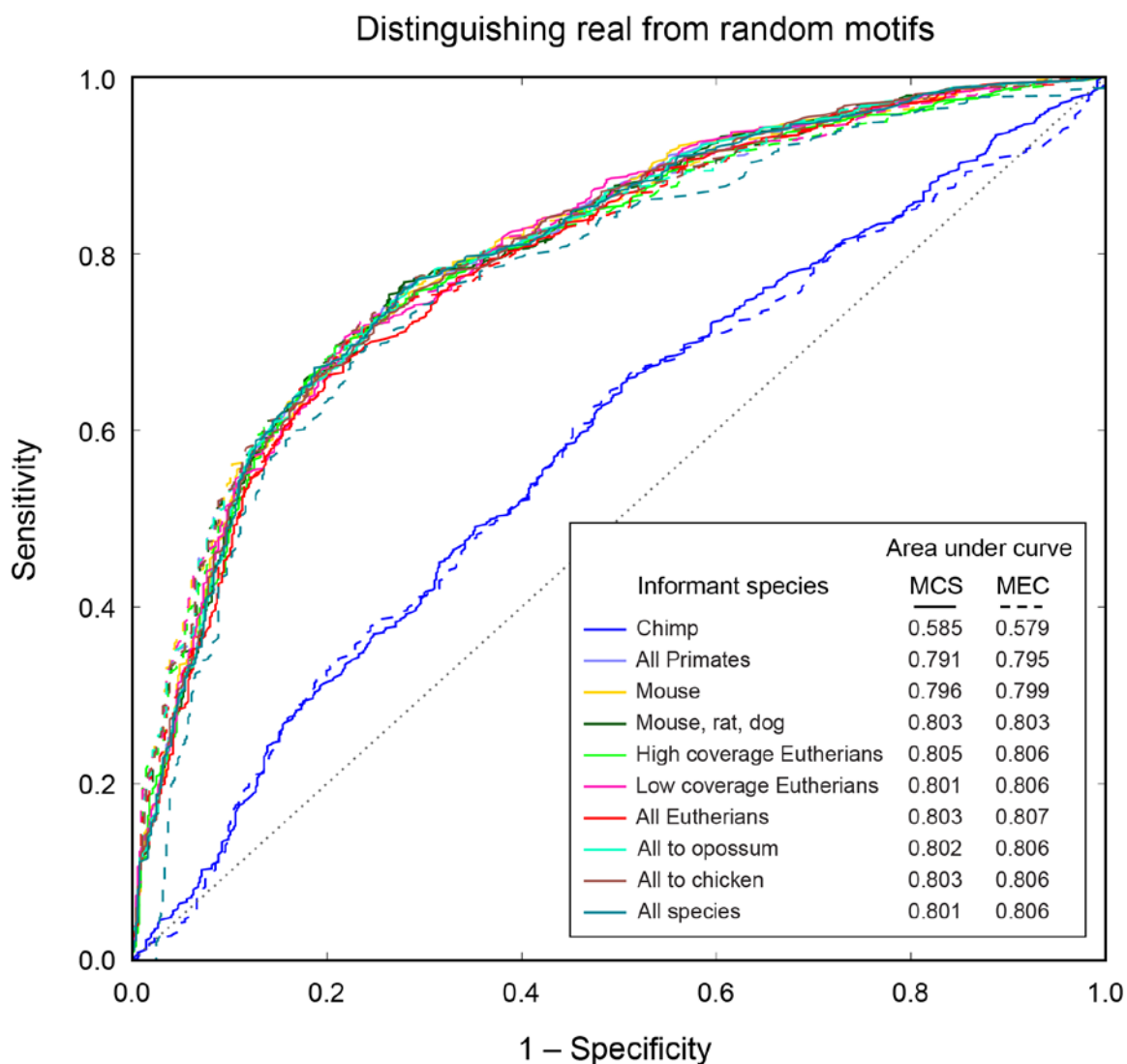


Supplementary Figure S10 - Thermodynamic analysis of EvoFold predictions using RNAz.

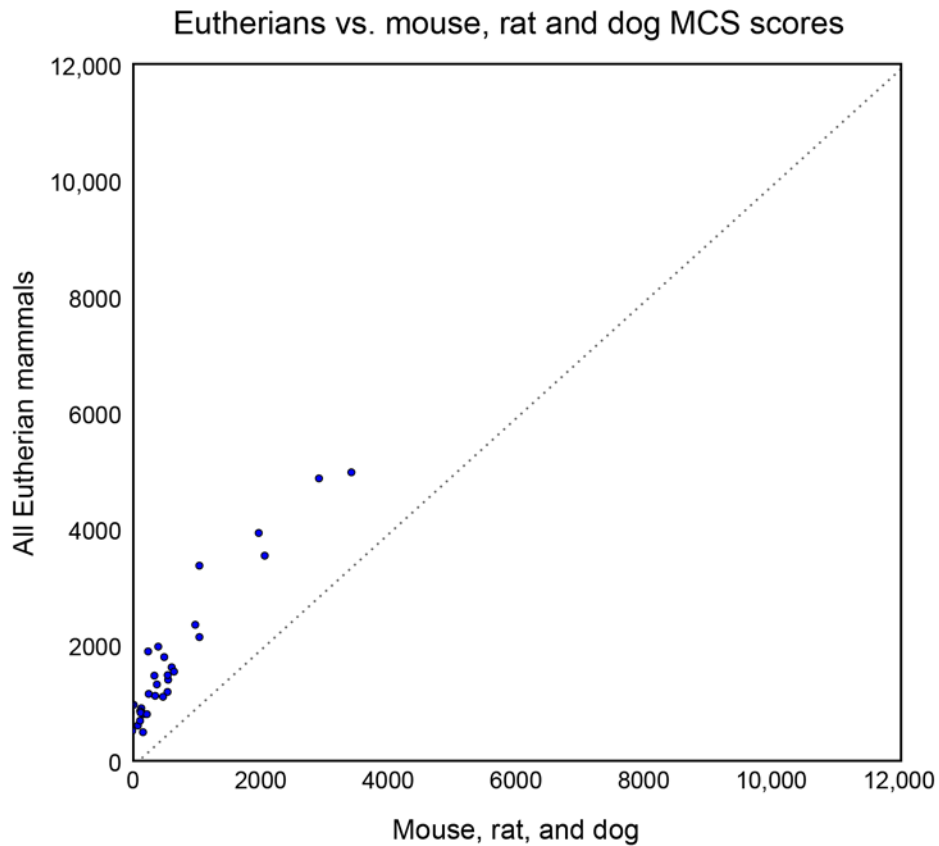
Left: Cumulative stability z-score distribution for all EvoFold structures structures, structures clustered in families (before and after additional filtering), and positive and negative controls. More negative z-score indicate more stable RNA structures. Right: Fraction of structures predicted as "Functional RNA" by RNAz for the different sets.



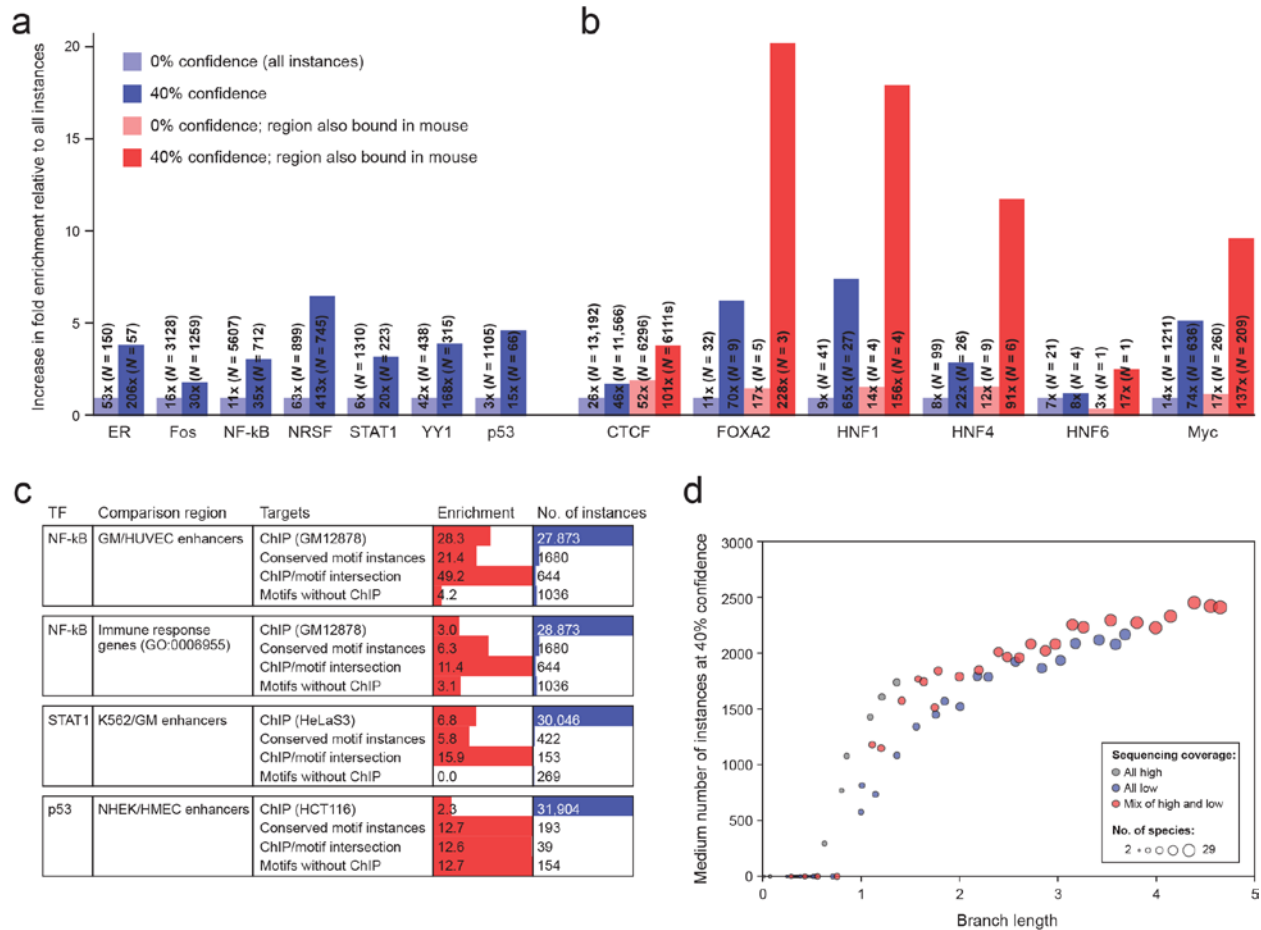
Supplementary Figure S11 - Differing constraint patterns are observed in core promoters. **a**, Analysis of promoters for 47,945 transcripts identified three patterns of high (red), intermittent (blue) and low (green) constraint. **b**, The different promoter types were enriched for different types of genes, with developmental genes significantly associated with high constraint, basic cell function and metabolic processes associated with intermittent constraint, and low constraint promoter overrepresented for sensory and immune response genes. **c**, The genes with intermittent constraint had between 1-9 peaks of constraint within the 200 bp core promoter. **d**, The spacing between constraint peaks varies between 9 and 200 bp, with a strong enrichment compared to an exponential distribution (cyan line), and local maxima at 12 and 21 bp distance between peaks (inset), corresponding to roughly one and two turns of the helix. Residual counts are obtained by subtracting the shown exponential from the spacing distribution curve.



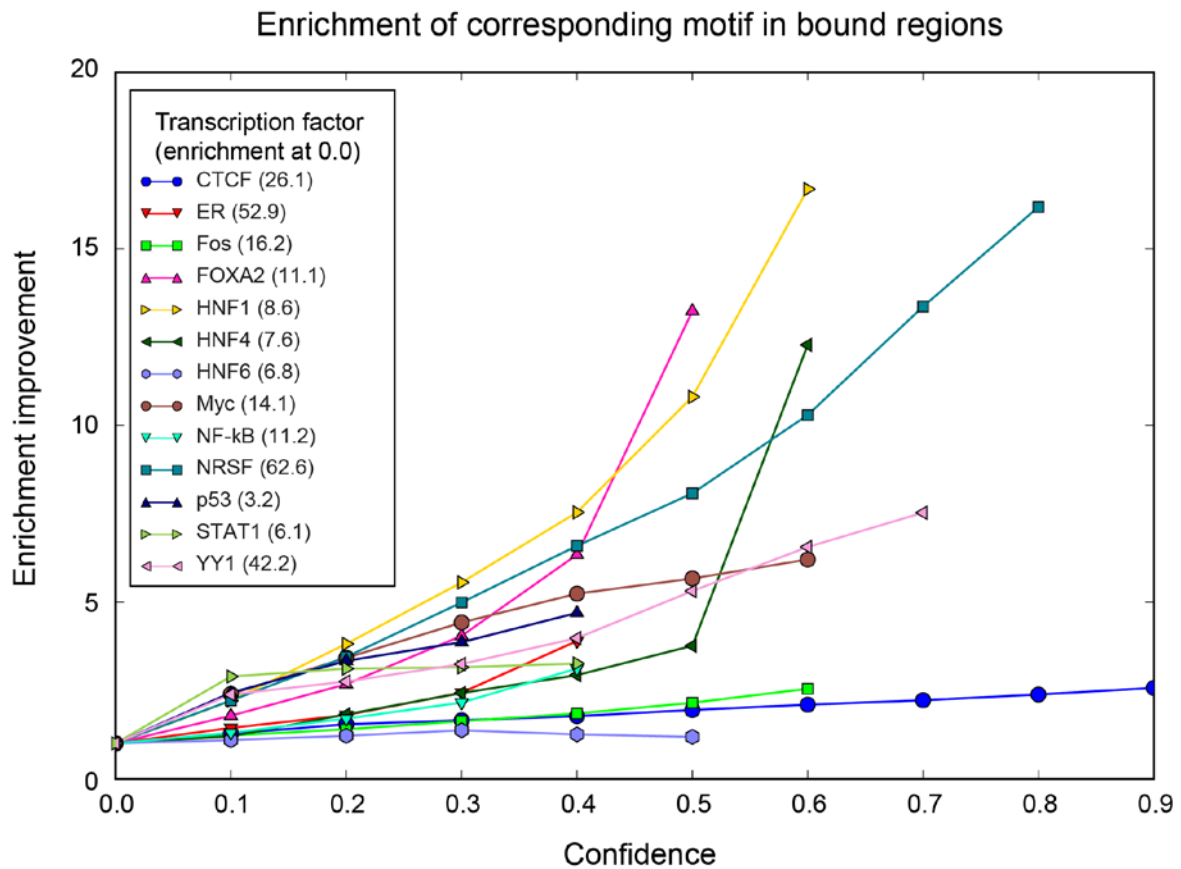
Supplementary Figure S12 - Limited number of species necessary to separate known and random motifs on the basis of conservation. ROC curves comparing different informant species subsets in identifying real motifs (contrasted to motif instances). Two methods are used to score motifs: MCS⁵⁴ in solid lines or MEC⁵⁵ in dashed lines. All motifs with at least two shuffles (N = 577) in the known motif database were scored genome-wide to show a preference for being conserved at the optimal branch length score (in terms of AUC) for each species subset. Additionally, shuffles of these motifs were scored using the same criterion. Using only mouse, rat, and dog as informant species performs essentially identically to using the entire Eutherian tree in separating the known and shuffled motifs. Indeed, even using just a single informant (mouse), has nearly equivalent performance. The two scoring schemes also distinguish between the two motif sets equally well. This demonstrates that at the number of instances and level of conservation seen for motifs in our database, motif discovery will likely not perform better when using motif conservation methods that employ a statistical conservation signal across the instances found genome-wide.



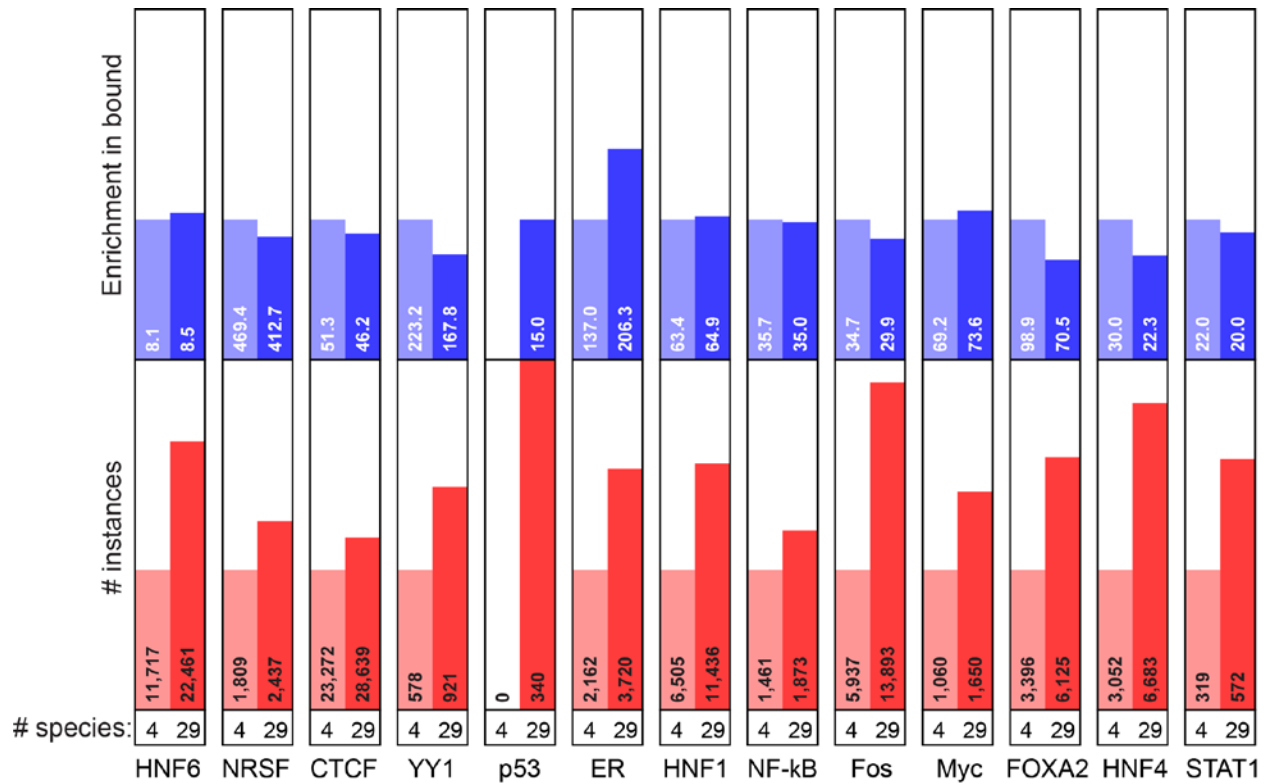
Supplementary Figure S13 - Motif conservation score (MCS) is strongly correlated when using the entire Eutherian tree or only mouse, rat and dog as informant species. A correlation of 0.99 is seen between the MCS scores on known motifs computed using all Eutherian informant species and when only using mouse, rat, and dog as informants. This extreme correlation fails to identify motifs that are better suited to be found with the larger Eutherian clade compared to the three species.



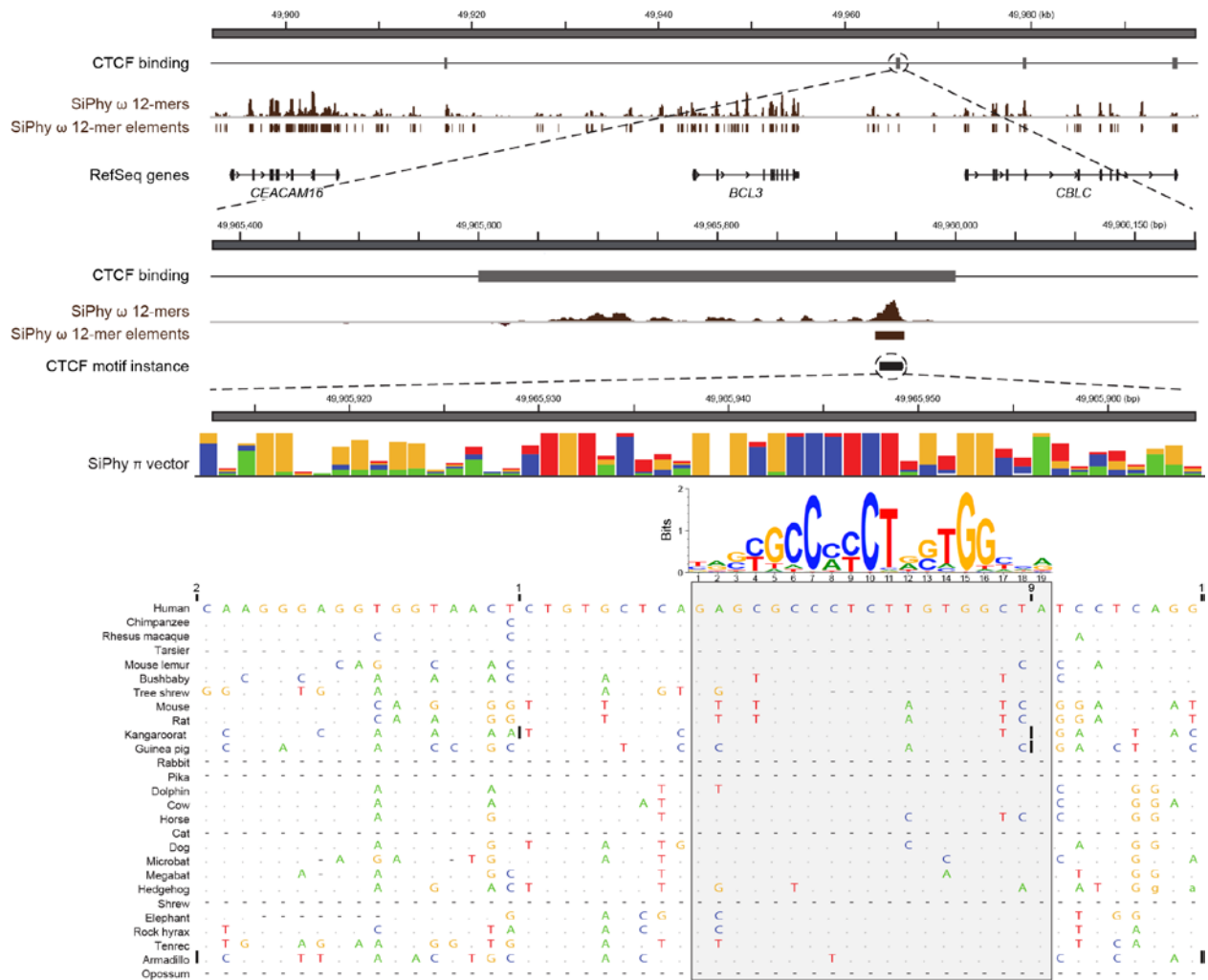
Supplementary Figure S14 - Regulatory motif instances associated with binding and putative functions. **a**, Enrichment of motifs in published experimental data sets. Known motifs for each factor show an enrichment in experimental data sets, which increases with conservation. **b**, Enrichment further increases for regions that are bound both in human and in the orthologous positions in mouse. **c**, Comparison of ChIP and conserved motif instances in identifying regions and genes likely to be bound by a factor. **d**, Scaling of motif instances using different species subsets. Comparison of high and low coverage species demonstrates the value of having low coverage species.



Supplementary Figure S15 - Increase in enrichment of motif instances across several factors. Motif enrichments are divided by enrichment at 0.0 confidence (i.e. all motif instances). Most factors show consistent and substantial increases in enrichment with increasing confidence. Considering only 0.0 and 0.4 confidence values leads to Supplementary Figure S12a.



Supplementary Figure S16 - Comparison of number of motif instances and enrichment of motifs at 40% confidence using either 4 species (human, mouse, rat, and dog) or 29 species (all Eutherian mammals). The number of motif instances (shown in red bars; normalized height to 4 species number) always increases when going from 4 species to 29. Conversely, the enrichment (shown in blue bars) does not show a clear preference for either species. This is consistent with the expectation that motif instances be equally good regardless of the species subset used to identify them as long as their confidence level is the same.



Supplementary Figure S17 - Example of a constrained element with a region of CTCF binding between *BCL3* and *CBL3* in the human genome. SiPhy rate indicates the level of constraint in overlapping 12-mers. A close-up of the constrained element shows overlap with a predicted CTCF site.

		Chromatin mark frequencies (%)																		
		CTCF	H3K27me3	H3K36me3	H4K20me1	H3K4me1	H3K4me2	H3K4me3	H3K27ac	H3K9ac	WCE	Percent of unexplained genome covered (%)	Percent of unexplained constraint covered (%)	Constraint enrichment	Cumulative % of unexplained genome covered	Cumulative enrichment	Chromatin state tentative annotations			
Chromatin states	1	16	2	2	6	17	93	99	96	98	2	0.5	1.5	3.27	1.5	3.27	1	Active promoter		
	3	13	72	0	9	48	78	49	1	10	1	0.6	1.8	2.96	3.3	3.10	3	Inactive/poised promoter		
	6	7	1	1	3	58	75	8	6	5	1	4.9	11.7	2.40	14.9	2.53	6	Weak/poised enhancer		
	4	11	1	15	11	96	99	75	97	86	4	2.0	4.1	2.05	19.0	2.41	4	Strong enhancer		
	2	12	2	6	9	53	94	95	14	44	1	0.5	0.8	1.52	19.7	2.36	2	Weak promoter		
	5	5	0	10	3	88	57	5	84	25	1	3.5	5.0	1.44	24.7	2.09	5	Strong enhancer		
	7	2	1	2	1	56	3	0	6	2	1	7.5	9.8	1.32	34.5	1.79	7	Weak/poised enhancer		
	8	92	2	1	3	6	3	0	0	1	1	1.1	1.4	1.23	35.9	1.76	8	Insulator		
	12	1	27	0	2	0	0	0	0	0	0	13.1	13.4	1.02	49.3	1.47	12	Polycomb repressed		
	9	5	0	43	43	37	11	2	9	4	1	1.6	1.5	0.96	50.8	1.45	9	Transcriptional transition		
	11	0	0	3	2	0	0	0	0	0	0	20.2	17.2	0.85	68.0	1.23	11	Weak transcribed		
	10	1	0	47	3	0	0	0	0	0	1	0.9	1.0	1.03	69.0	1.23	10	Transcriptional elongation		
	13	0	0	0	0	0	0	0	0	0	0	43.6	31.0	0.71	100.0	1.00	13	Heterochromatin; low signal		
	14	22	28	19	41	6	5	26	5	13	37	0.1	0.0	0.06	100.0	1.00	14	Repetitive/copy number variation		
	15	85	85	91	88	76	77	91	73	85	78	0.0	0.0	0.06	100.0	1.00	15	Repetitive/copy number variation		

Supplementary Figure S19 - Overlap of constraint bases with chromatin states. As much as 36% of unexplained constrained bases overlap chromatin states associated with candidate promoter, enhancer, or insulator states in at least one cell type among nine diverse cell lines. Reported enrichments were calculated after first excluding locations assigned to a higher listed state in one or more cell types.

Cumulative % of "unexplained" constrained bases covered

Chromatin State Tentative Annotations	state	9	8	7	6	5	4	3	2	1	0
Active promoter	1	0.05	0.09	0.14	0.19	0.25	0.32	0.45	0.66	1.51	100.00
Inactive/poised promoter	3	0.06	0.12	0.21	0.32	0.46	0.64	0.89	1.35	3.27	100.00
Weak/poised enhancer	6	0.06	0.13	0.24	0.41	0.67	1.15	2.17	4.92	14.92	100.00
Strong enhancer	4	0.12	0.27	0.52	0.91	1.53	2.55	4.32	8.06	18.98	100.00
Weak promoter	2	0.47	0.80	1.13	1.58	2.23	3.22	4.99	8.75	19.73	100.00
Strong enhancer	5	0.55	0.98	1.50	2.30	3.55	5.23	7.83	12.74	24.72	100.00
Weak/poised enhancer	7	0.67	1.28	2.15	3.44	5.40	8.03	11.89	18.82	34.54	100.00
Insulator	8	1.14	1.94	2.94	4.34	6.40	9.14	13.12	20.15	35.92	100.00
Polycomb repressed	12	1.74	3.27	5.12	7.55	10.94	15.48	21.76	31.48	49.29	100.00
Transcriptional transition	9	2.06	3.78	5.81	8.42	11.96	16.65	23.09	32.96	50.82	100.00
Weak transcribed	11	4.19	8.22	12.73	17.99	24.20	31.33	40.01	51.22	68.05	100.00
Transcriptional elongation	10	9.51	14.10	18.38	23.06	28.49	34.89	42.71	53.06	69.02	100.00
Heterochromatin; low signal	13	99.58	99.93	99.96	99.97	99.97	99.98	99.99	99.99	99.99	100.00
Repetitive/copy number variation	14	99.92	99.97	99.98	99.98	99.99	99.99	99.99	100.00	100.00	100.00
Repetitive/copy number variation	15	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00

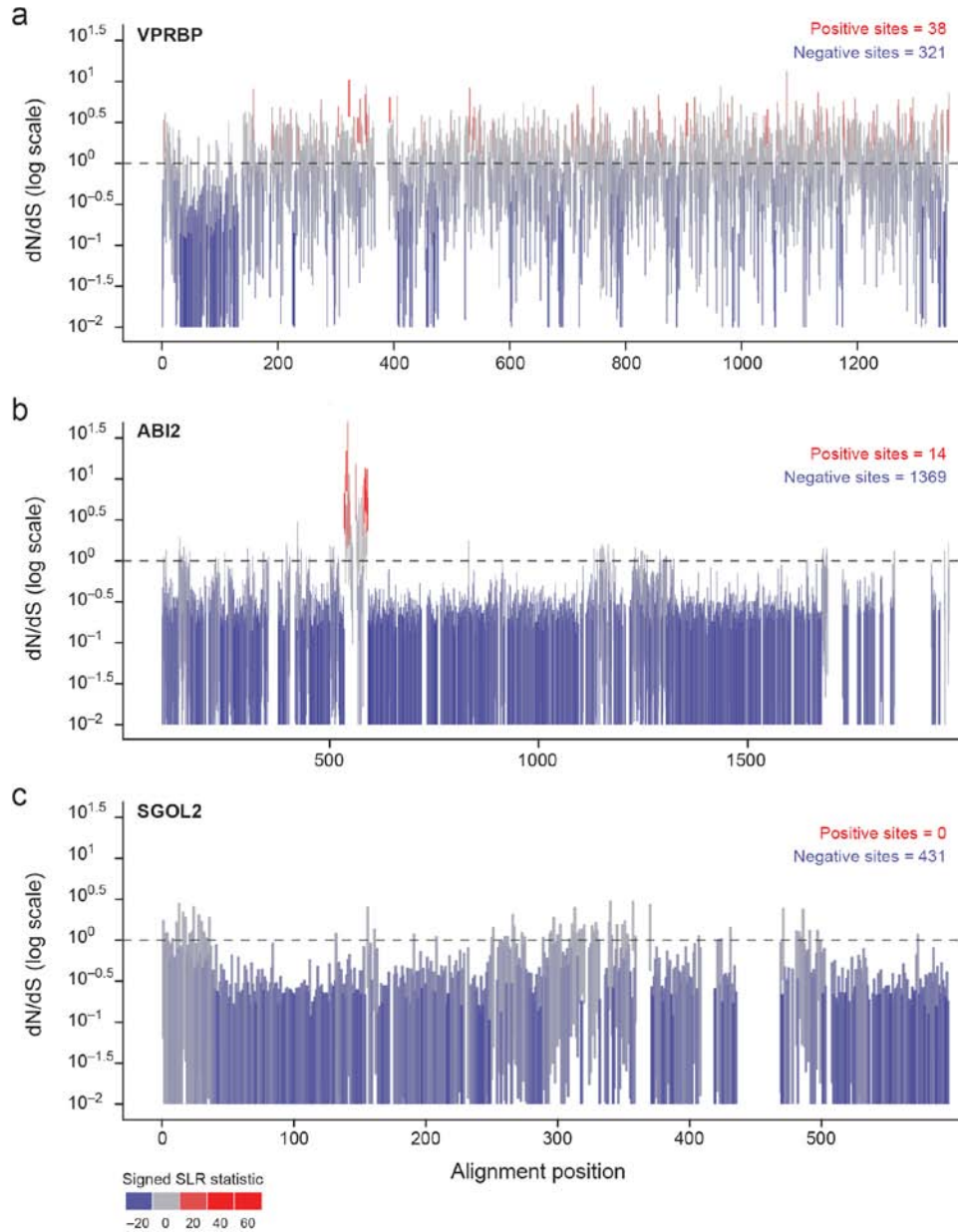
Cumulative % of "unexplained" genome bases covered

Chromatin State Tentative Annotations	state	9	8	7	6	5	4	3	2	1	0
Active promoter	1	0.01	0.02	0.03	0.04	0.05	0.07	0.10	0.15	0.46	100.00
Inactive/poised promoter	3	0.01	0.02	0.04	0.07	0.09	0.14	0.20	0.33	1.06	100.00
Weak/poised enhancer	6	0.01	0.03	0.05	0.09	0.17	0.30	0.61	1.54	5.91	100.00
Strong enhancer	4	0.02	0.06	0.11	0.21	0.37	0.65	1.20	2.57	7.88	100.00
Weak promoter	2	0.13	0.22	0.31	0.43	0.60	0.88	1.46	2.89	8.38	100.00
Strong enhancer	5	0.15	0.26	0.41	0.64	1.03	1.61	2.65	4.93	11.85	100.00
Weak/poised enhancer	7	0.18	0.36	0.63	1.08	1.84	2.97	4.88	8.81	19.31	100.00
Insulator	8	0.38	0.65	1.00	1.52	2.36	3.58	5.60	9.68	20.44	100.00
Polycomb repressed	12	0.55	1.06	1.79	2.94	4.74	7.45	11.68	18.86	33.51	100.00
Transcriptional transition	9	0.68	1.31	2.17	3.45	5.41	8.31	12.76	20.23	35.11	100.00
Weak transcribed	11	2.55	5.46	8.99	13.19	18.16	24.00	31.18	40.54	55.34	100.00
Transcriptional elongation	10	8.18	11.61	14.78	18.27	22.42	27.43	33.76	42.31	56.28	100.00
Heterochromatin; low signal	13	99.31	99.70	99.75	99.78	99.79	99.81	99.83	99.86	99.88	100.00
Repetitive/copy number variation	14	99.80	99.87	99.89	99.90	99.91	99.92	99.93	99.95	99.96	100.00
Repetitive/copy number variation	15	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00

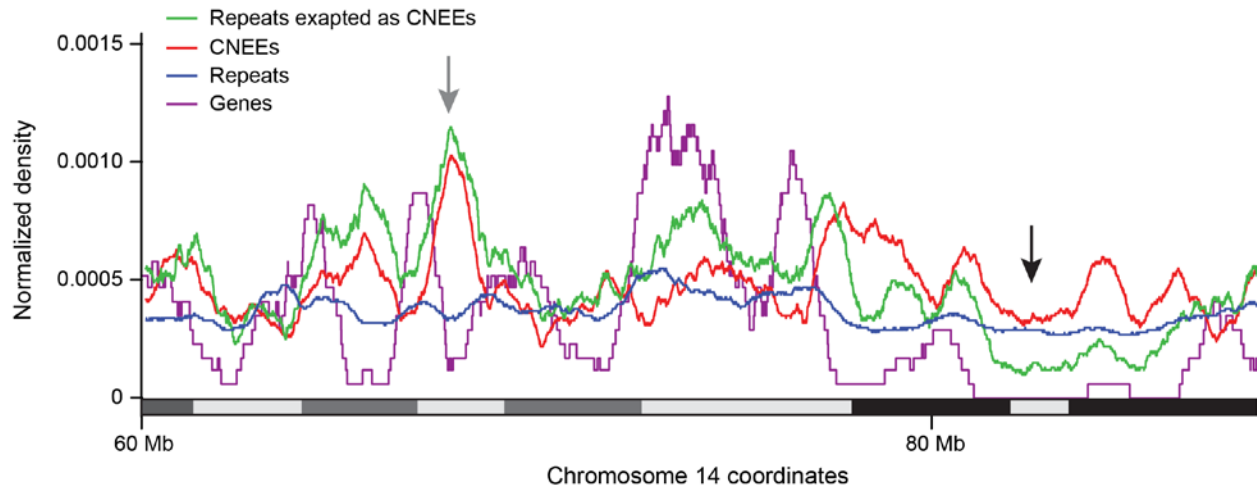
Cumulative enrichments in "unexplained" constrained elements

Chromatin State Tentative Annotations	state	9	8	7	6	5	4	3	2	1	0
Active promoter	1	5.42	5.02	5.00	4.92	4.79	4.62	4.59	4.31	3.27	1.00
Inactive/poised promoter	3	5.59	5.04	5.00	4.93	4.84	4.70	4.49	4.13	3.10	1.00
Weak/poised enhancer	6	5.51	4.92	4.73	4.46	4.07	3.79	3.53	3.20	2.53	1.00
Strong enhancer	4	5.02	4.68	4.53	4.35	4.12	3.89	3.59	3.14	2.41	1.00
Weak promoter	2	3.60	3.63	3.65	3.69	3.71	3.65	3.42	3.03	2.36	1.00
Strong enhancer	5	3.70	3.71	3.68	3.61	3.44	3.24	2.96	2.59	2.09	1.00
Weak/poised enhancer	7	3.73	3.60	3.42	3.18	2.93	2.71	2.44	2.14	1.79	1.00
Insulator	8	2.97	2.98	2.94	2.85	2.71	2.56	2.34	2.08	1.76	1.00
Polycomb repressed	12	3.18	3.09	2.85	2.57	2.31	2.08	1.86	1.67	1.47	1.00
Transcriptional transition	9	3.01	2.89	2.68	2.44	2.21	2.00	1.81	1.63	1.45	1.00
Weak transcribed	11	1.64	1.50	1.42	1.36	1.33	1.31	1.28	1.26	1.23	1.00
Transcriptional elongation	10	1.16	1.21	1.24	1.26	1.27	1.27	1.27	1.25	1.23	1.00
Heterochromatin; low signal	13	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Repetitive/copy number variation	14	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Repetitive/copy number variation	15	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

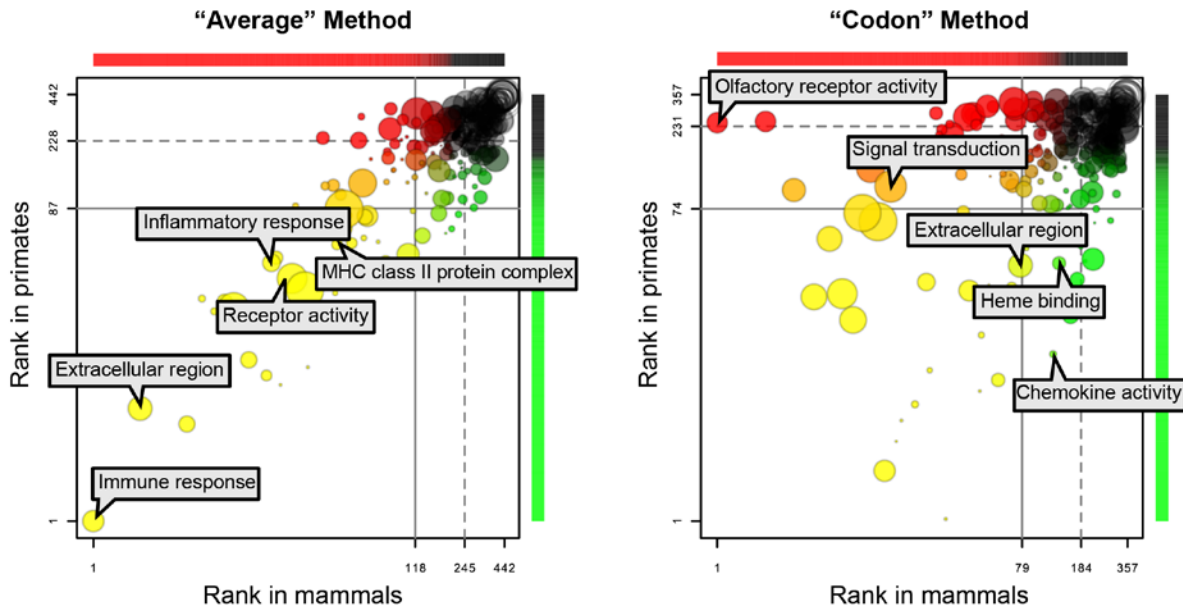
Supplementary Figure S20 - Chromatin state conservation analysis for varying numbers of cell types. The top grid indicates the cumulative total % of constrained bases in the unmasked, that is 'unexplained' portion of the genome, assigned to the state of the row or a higher listed state in that number of cell types of the column or more. The middle table is similar, but for all bases in the 'unexplained' portion of the genome. The last table gives the enrichment for constrained bases, that is the ratio of the values in the first table to the second table.



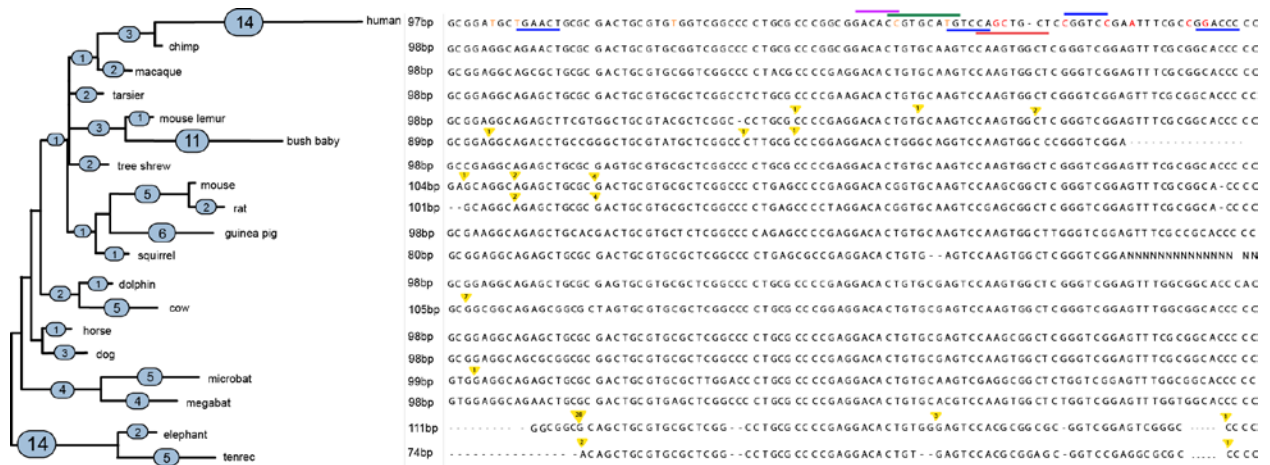
Supplementary Figure S21 - Positive selection acts uniformly for some genes and in a site-specific fashion for other genes. **a**, *VPRBP* shows a mixture of negative (purifying) and positive selection along its length, **b**, *ABI2* shows strong purifying selection with negatively-selected sites over most of the sequence and a localized region of positively-selected sites, **c**, *SGOL2* shows only purifying selection with negatively-selected sites over the entire gene. Each vertical bar covers the estimated 95% confidence interval for dN/dS at that site (with values of 0 truncated to 0.01 to accommodate the log scaling), and bars are colored according to a signed version of the SLR statistic for non-neutral evolution: blue for sites under negative selection, gray for sites under neutral evolution, and red for sites under positive selection. A dotted line is shown at neutral evolution ($dN/dS=1$) for reference, and the number of statistically significant sites under positive and negative selection ($FDR < 5\%$) is shown for each gene.



Supplementary Figure S22 - Exaptation of non-coding constrained elements is overrepresented near developmental genes and depleted in the middle of large gene deserts. Repeats exapted as CNEEs follow the density of all CNEEs, and not that of the repeat elements (grey arrow). The normalized density of CNEEs exapted from repeat elements (green) is shown compared to protein coding genes (purple), CNEEs (red), and repeats (blue). Densities are calculated as the number of elements within a 1 Mb window divided by the total number of elements in the genome. The trends in this section of chromosome 14 are representative of the genome as a whole. CNEEs exapted from repeats tend to be depleted in gene deserts larger than 1 Mb (black arrow).



Supplementary Figure S23 - GO enrichments for positively selected genes in primates versus mammals. GO term enrichments were calculated using the distributed positive selection (left) and localized positive selection (right) methods (see Methods). Each bubble represents one GO term, and its size is proportional to the log of the number of times the term was annotated in the gene “universe” (17,709 genes). Colors represent unadjusted p-values for enrichment in primates (red), mammals (green), both (yellow), or neither (black). The horizontal and vertical positions of the bubble correspond to the term’s rank in mammals or primates, respectively, plotted on a log scale. Lines on each axis indicate ranks corresponding to $p=0.05$ (solid) and $p=0.5$ (dashed). Using the distributed positive selection method, statistical significance of enrichment is highly correlated in primates and mammals, suggesting that similar genes and pathways are under overall selection. In contrast, the localized positive selection method shows greater discordance between primates and mammals, indicating that different sets of genes show signals of selection at the single codon level in the two clades.



Supplemental Figure S24 - Rapid human evolution in the 5' UTR of FGF13. 2xHAR1 is a conserved non-coding element in the 5' UTR of the shortest FGF13 isoform that has evolved rapidly in humans since divergence from chimpanzee. Human-specific substitutions in 2xHAR1 could potentially have altered isoform-specific transcription. *Left:* Unrooted mammalian phylogeny for 2xHAR1 with branch lengths proportional to the expected number of substitutions on each lineage, annotated in a bubble on each branch. *Center:* Length of the 2xHAR1 sequence for each species. *Right:* Multiple sequence alignment. Species with very few aligned bases are omitted. Insertions in non-human species are shown as yellow triangles with the number of inserted bases annotated. Human-specific substitutions* are marked in red (fixed differences) or orange (polymorphic substitutions). The five polymorphic sites suggest that the function of 2xHAR1 may still be in the process of diverging from its ancestral state and are potential candidates for BFLS (Borjeson-Forssman-Lehmann syndrome) and other disease associations in humans. Predicted transcription factor binding site turnover events are marked with lines above (losses) or below (gains) the human sequence, with line color denoting the JASPAR family: blue= nuclear receptor, green= bZIP cEBP-like, red= bHLH(zip), purple= HMG. Note that the number of red and orange bases (12) does not match the expected number of substitutions (14), because of insertions and deletions in the alignment that are not displayed.