Supporting Online Material for

Three periods of regulatory innovation during vertebrate evolution

Craig B. Lowe, Manolis Kellis, Adam Siepel, Brian J. Raney, Michele Clamp, Sofie R. Salama, David M. Kingsley, Kerstin Lindblad-Toh, David Haussler

correspondence to: haussler@soe.ucsc.edu

This PDF file includes:

Materials and Methods
Figs. S1 to S9
Tables S1 to S12

Materials and Methods

Data sets

Data sets used in this study may be visualized as tracks in the UCSC Genome Browser: http://genomewiki.cse.ucsc.edu/index.php/Three_Periods_Of_Regulatory_Innovation

Alignment

Our 40way alignments referenced on the human, mouse, and cow genomes contains: human (hg18), chimp (panTro2), rhesus (rheMac2), tarsier (tarSyr1), mouse lemur (micMur1), bushbaby (otoGar1), tree shrew (tupBel1), mouse (mm9), rat (rn4), kangaroo rat (dipOrd1), guinea pig (cavPor3), squirrel (speTri1), rabbit (oryCun1), pika (ochPri2), alpaca (vicPac1), dolphin (turTru1), cow (bosTau4), horse (equCab2), cat (felCat3), dog (canFam2), microbat (myoLuc1), megabat (pteVam1), hedgehog (eriEur1), shrew (sorAra1), elephant (loxAfr2), rock hyrax (proCap1), tenrec (echTel1), armadillo (dasNov2), sloth (choHof1), opossum (monDom4), platypus (ornAna1), chicken (galGal3), zebrafinch (taeGut1), lizard (anoCar1), frog (xenTro2), tetraodon (tetNig1), fugu (fr2), stickleback (gasAcu1), medaka (oryLat2), and zebrafish (danRer5).

Since alignments containing low-coverage genomes can be enriched for misalignments due to the inability of applying synteny filters to initial pairwise alignments we created an additional alignment, referenced on the human genome, that only contains the well assembled species: human, chimp, rhesus, mouse, rat, guinea pig, cow, horse, dog, opossum, platypus, chicken, zebrafinch, frog, tetraodon, fugu, stickleback, medaka, and zebrafish. Medaka and stickleback alignments contained all five fish genomes as well as well-assembled mammals and chicken to define the most basal branching point.

Initial pairwise alignments were done between the reference genome and all other genomes using Blastz (*29*). We then chained these local alignments together to arrive at whole-genome alignments (*30*). For the well assembled genomes we applied a synteny filter to the alignments (*30*). The synteny filter is effective at distinguishing orthologs from paralogs, it can be too strict for the low-coverage genomes that have not been assembled into large continuous regions, and are therefore unable to produce long stretches of local alignments having the same order and orientation. For the low-coverage assemblies we applied a reciprocal-best filter (*30*). This filter distinguishes between orthologs and paralogs by demanding that each alignment between the target and query genomes is not only the best match in the query for the target, but also the best match in the target for the query. This filter does not require long continuous regions, which do not always exist in low-coverage assemblies. We then used Multiz (*31*) to create the final genome-wide alignments, using all the filtered pair-wise alignments as input.

Identification of conserved non-exonic elements

We defined a neutral model of evolution based on the rate of substitutions in four-fold degenerate sites in protein-coding regions (Fig S3). Our model of nucleotide substitutions in neutrally evolving DNA is reversible and we fit it to the alignment of four-fold degenerate sites using PhyloFit (*32*). We also created a chrX specific model of neutral evolution using only coding regions from that chromosome. We used PhastCons (*16*), a phylogenetic hidden Markov model (HMM), to define sets of conserved elements in our genome-wide alignments based on these models of neutral evolution. PhastCons has a neutral state and a conserved state. We defined the conserved state to be the same as the neutral state, but with all branch lengths scaled by 0.3. We set the transition probabilities of the HMM by expecting a length of 45 base pairs and a coverage of 0.3.

This set of ~4.3 million conserved elements covers ~4.5% of the human genome. To validate that this set of conserved elements is capable of capturing functional sequence, we examined its overlap with perhaps the most well understood set of functional elements, coding exons. We used the high-quality set of coding regions from the consensus coding sequence project (*33*). The set of conserved elements overlaps ~98% of the exons, verifying that this methodology is capable of detecting functional sequence.

We then filtered this set of conserved regions by removing any elements overlapping exons in the reference species' Ensembl gene set (*34*) or ESTs that were mapped to the reference genome using BLAT (*35*). We also removed elements overlapping exons from UCSC Known Genes (*36*), RefSeq Genes (*37*), mRNAs, and spliced ESTs from closely related species that could be mapped to the reference genome using TransMap (*38*). This filtering resulted in a set of conserved bases that do not appear in mature transcripts of either coding or noncoding genes.

Derived allele frequencies

We used SNP allele frequencies for the Yoruban population, as reported by the HapMap Consortium (*20*). We chose the Yoruban population, as opposed to populations outside of Africa, to minimize any effect of population bottlenecks. We only used SNPs that are currently segregating in Yorubans, appear to have only two alleles, and both chimp and rhesus agree on which of the two alleles is ancestral. This is so that we may confidently identify which of the alleles is ancestral and which is derived. We then created three subsets of these SNPs. First, we extracted the set of SNPs that reside within our CNEEs. We then placed this derived allele frequency spectrum in the context of a near-neutral spectrum and a spectrum representing positions under substantial selection (Fig. S2). We used the annotation from dbSNP (*39*) to extract a set of intronic SNPs to represent a near-neutral population and a set of non-synonymous positions to represent positions under substantial selection. We used the Mann-Whitney test to assess the significance of any shift in derived allele frequency between the subsets.

Overlap with experimental data sets to identify regulatory regions

We developed a statistical framework to assess the overlap of two sets comprised of genomic regions of arbitrary size. We used this framework to quantify the enrichment between the set of CNEEs and experimental data sets, such as ChIP-seq.

For the 'i'th regulatory region we computed the number of locations where this region could be placed in the genome that would not overlap an assembly gap (T_i). We then computed the number of locations that do not overlap an assembly gap, but do overlap a CNEE (O_i). These numbers allowed us to compute the probability of the 'i'th regulatory region overlapping a CNEE by chance, assuming the experimental regulatory regions are uniformly distributed ($P_i = O_i/T_i$).

We exploited a recursive dependence in the probability of having ‘n’ out of ‘m’ experimental sites overlap CNEEs, Q(n,m).

The base cases are:
$Q(1,1) = P_1$
and
$Q(0,1) = (1-P_1)$.

The boundary cases are:
$Q(m,m) = Q(m-1,m-1) * P_m$
and
$Q(0,m) = Q(0,m-1) * (1-P_m)$.

The general recurrence relation is:
$Q(n,m) = Q(n-1,m-1) * P_m + Q(n,m-1) * (1 - P_m)$

The p-value of having at least ‘n’ out of ‘m’ experimental sites overlapping CNEEs:
$P(X>=n) = Q(n,m) + Q(n+1,m) + Q(n+2,m) + ... + Q(m-2,m) + Q(m-1,m) + Q(m,m)$
$P(X>=n) = Sum(from:i=n, to:m, equation:Q(i,m))$

Identifying the branch of origin for CNEEs

For each conserved non-exonic element we determined the most recent common ancestor of all species in the alignment that have an orthologous piece of DNA covering at least one-third of the CNEE in the reference genome (Fig. S1). We used this most recent common ancestor as the time of birth unless a subtree, including the reference species, showed an increase in evolutionary constraint greater than the rejection of the neutral model by the alignment as a whole. In this case, the subtree showing the strongest rejection of the null hypothesis was used as the birth of the conserved element (Fig. S1). We used the likelihood ratio test method in the PhyloP program (*19*) to compare CNEE sequences to the model of neutral evolution based on 4-fold degenerate sites. When testing for a significant onset of constraint we used PhyloP to scale the neutral tree based on the distantly related sequences outside the subtree and compared this rate to the rate of substitution seen in the subtree containing the reference.

We do not exclusively use the most recent common ancestor of all species present in the alignment because CNEEs coming under selection in the primates may have neutrally evolving sequences that are still alignable to the more distant primates, or even rodents. For this reason we also test each node in the tree on the way to the reference species to understand if there has been a significant decrease in the rate of substitution, relative to the more distant species appearing in the alignment.

For the human lineage we identified less than one in every 1000 CNEEs as having experienced the onset of constraint after the most recent common ancestor of orthologous DNA in the alignment.

Identifying the branch of origin for genes

For genes we first identified the splice form of the gene with the most exons. This transcript is used to represent the gene as a whole. We then determined the branch of origin by calculating the most divergent species in the alignment that overlaps at least 50% of the bases in the gene. We then inferred that the gene was created on the branch leading to the most recent common ancestor of the genome being analyzed and the divergent species covering at least half of the bases in the gene.

Functional Enrichment

We used the Uniprot database (*40*) to map Gene Ontology (*25*) terms to the Ensembl gene sets (*34*) for human, mouse, cow, chicken, and zebrafish. Using TreeFam's orthology and paralogy mappings (*41*) for Ensembl genes in sequenced vertebrate genomes, we then distributed this annotation among the human, mouse, cow, medaka, and stickleback genomes. If two genes coalesce at the euteleostomi ancestor or more recently, then GO terms are shared between them. This serves to reduce any bias in the GO annotation between different organisms and allows us to analyze the stickleback and medaka lineages that would otherwise not have GO terms associated with their genes. Genes with no functional annotation were removed from the set.

The statistical tests to assess if a set of gene regulatory elements is located near a certain group of genes were implemented based on previously published methods (*42*). Conserved non-exonic elements were assigned to the gene with the closest transcription start site. In the case of genes with multiple transcription start sites, the average of all start sites is used. For each GO term we calculated the number of ungapped bases in the genome that are closest to a transcription start site of a gene with the given GO term. This number of bases associated with each GO term is then divided by the number of ungapped bases in the genome to yield the expected percentage of CNEEs assigned to a gene with the given GO term, provided that CNEEs are uniformly distributed in the genome. The enrichment factor was calculated as the actual percentage of CNEEs near a GO term, divided by the expected number. P-values are calculated using the binomial distribution where the number of trials is the number of CNEEs, the number of successes is the actual number of CNEEs near the GO term, and the null model rate of success is the expected percentage of CNEEs near the GO term. This method compensates for some sets of genes having larger genic or intergenic regions.

The terms featured in the analysis are GO:0003700 (transcription factor activity), GO:0032502 (developmental process), GO:0005102 (receptor binding), GO:0043687 (post-translational protein modification), and MP:0001510 (abnormal coat appearance).

Putatively Lost CNEEs

Ancient CNEEs being enriched near trans-dev genes could be the result of CNEEs near trans-dev genes being less likely to turn-over than CNEEs associated with other genes. To address this we identified CNEEs in the human, mouse, and cow assemblies that were reliably present in the therian ancestor, but putatively lost in other mammalian genomes. For example, we identified human CNEEs that were at least 50% covered in rhesus, rat, mouse, and opossum, but have no orthologous bases in dog or horse. These are CNEEs present in extant humans that were likely present in the therian ancestor, but have since been lost in dog and horse. We also identified human CNEEs 50% covered by rhesus, dog, horse, and opossum, but having no orthologous bases in mouse or rat. We computed similar sets for the mouse and cow genomes looking for losses in human and rhesus, mouse and rat, or dog and horse. By identifying CNEEs missing in two closely related genomes that are well assembled we intend our sets to be further enriched for actual losses and not assembly gaps.

This resulted in 6 related sets of CNEEs that are missing in some mammalian lineages. If some functional categories of genes tended to lose their regulatory regions more or less often, then we should witness an enrichment or depletion in our lost CNEE set relative to the set present in the therian ancestor that we started with. We did not see any consistent enrichments or deletions for the GO terms focused on in this study (Fig. S7). This does not preclude that there might have been an earlier bias on the human lineage for CNEEs near trans-dev genes to resist being lost, but a denser tree of well-assembled vertebrate genomes will be needed to fully address earlier time points.

Robustness of trends

To ensure that the changes we see in enrichments over time are robust against the alignment and species used we have performed our analysis on a separate human-referenced alignment using only deeply-sequenced and well-assembled genomes along with stringent alignment parameters. The results are qualitatively similar (Fig. S9). We performed this additional analysis because our method for dating CNEEs may be sensitive to false-positive alignments in distantly related vertebrate genomes that cause us to believe the origin of a CNEE is more ancient than it actually is. We believe these misalignments are rare as the evolutionary history of 96% of CNEEs from this alignment need three or fewer deletions to explain the absence of species in their section of the alignment.

To ensure that our results are robust against our choice of gene set, CNEE to gene assignment algorithm, and GO term to gene mapping, we performed our human lineage analysis using the GREAT (great.stanford.edu) enrichment software (*26*). GREAT uses the UCSC Known Genes gene set and a more detailed algorithm for assigning CNEEs to genes. The results were qualitatively similar (Fig S9). The few differences are largely due to the different CNEE to gene assignment algorithms. Our method assigns all CNEEs to the gene with the closest transcription start site, while we ran GREAT with a more complex rule that can at times allow a CNEE to be assigned to both of the neighboring genes. Over 200 ancient CNEEs we assigned exclusively to LAMA2 were also assigned to PTPRK by GREAT, which contributed to the enrichment differences for ancient CNEEs being enriched or depleted near genes involved in protein modification.

Approximately 98.8% of the CNEEs defined in the human-referenced 40way alignment are missing in some species where they would be expected based on the branch of origin. This may appear to be a high-percentage, but a number of the mammalian genomes were sequenced at ~2x coverage, where we expect only ~65% of the genome to be present in each assembly (*43*). In our main analysis we include CNEEs that are missing from some species, but we have redone our analysis with the ~1.2% of our CNEEs that are present in every species derived from the common ancestor in which they appear (Fig. S8). All trends are not only present, but possibly exaggerated in this small subset of CNEEs with unambiguous evolutionary origins.

While CNEEs show an overall signature of conservation, it is possible for some subclades to have lost or relaxed this constraint on the element. We used PhyloP (*19*) to identify 679 CNEEs showing a relaxation of constraint on a subtree (p-value threshold of 0.1 after correcting for multiple tests). Removing these CNEEs from our analysis did not affect the trends described (Fig. S8).

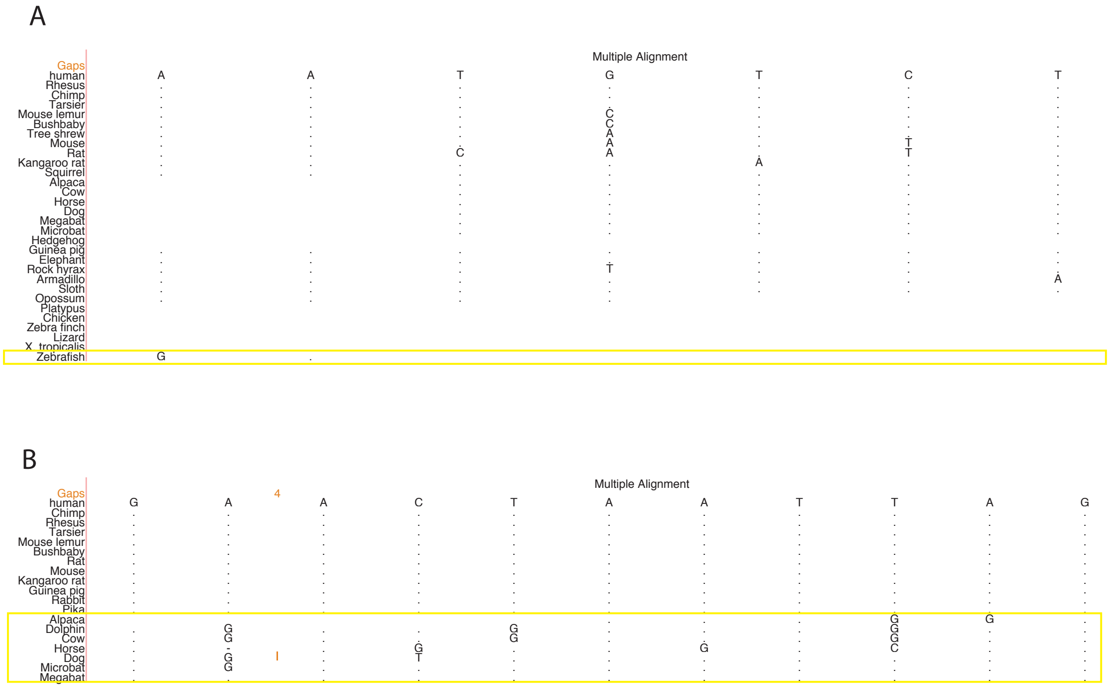


Fig. S1

Dating the origin of CNEEs. Orthologous bases that are identical to human are displayed as a dot while differences are displayed using the one-letter base code. There are two conditions that may cause the CNEE to be placed on a more recent branch than the most recent common ancestor of all species with orthologous DNA in the alignment. (A) If the orthologous DNA does not cover at least one third of the CNEE, then we do not count the species as having an orthologous copy of the CNEE. The zebrafish alignment (yellow rectangle) is not counted and the branch leading to the common ancestor of human and opossum is used as the date for this CNEE. (B) In very rare cases it appears that the majority of constraint originated after the common ancestor of all species in the alignment. In this case it appears that a significant amount of constraint began to act on this CNEE on the branch leading to the human-rodent ancestor. Therefore, the species outside of this subtree (yellow rectangle) are not used to date this CNEE and the branch leading to the human-rodent ancestor is used as the origin.

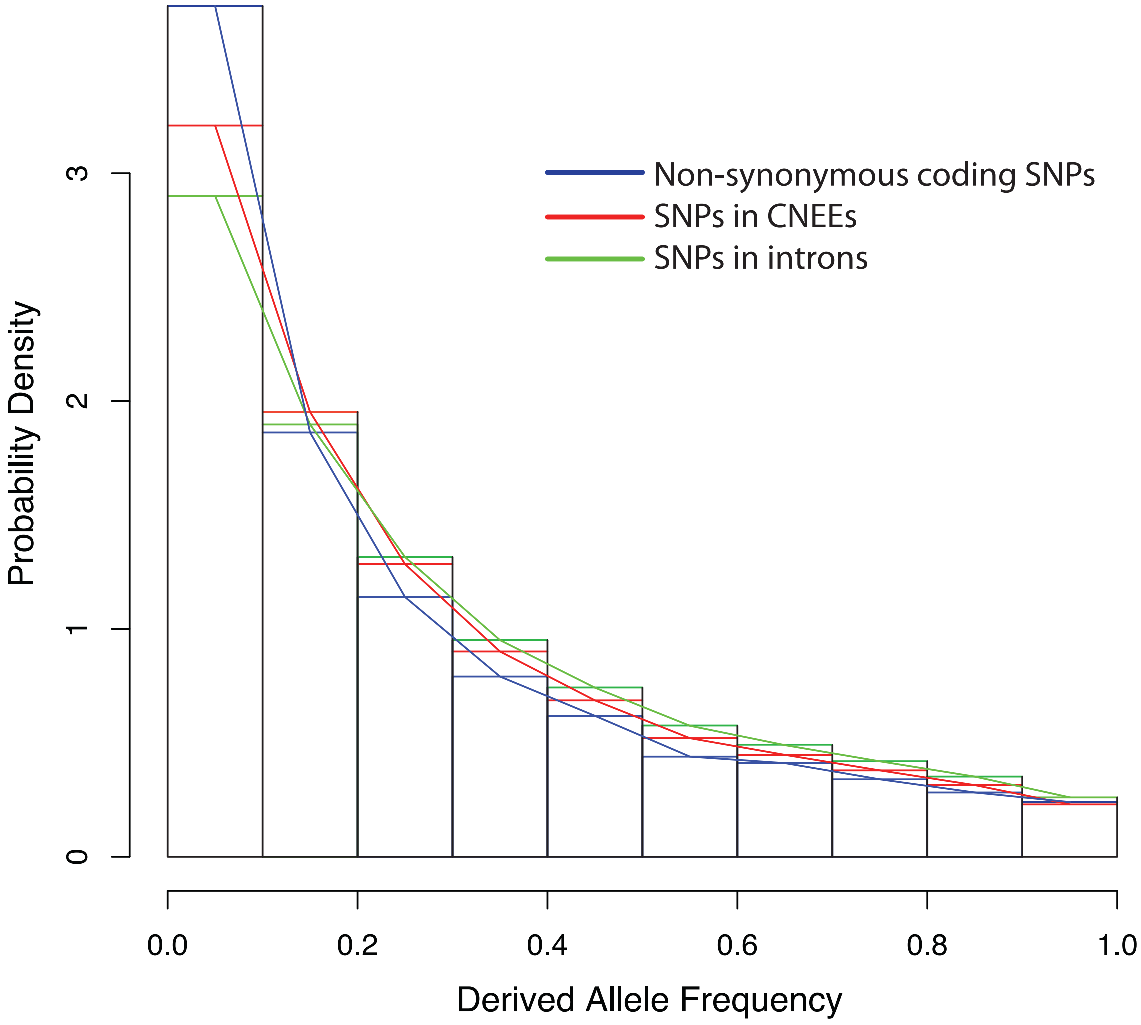


Fig. S2

We compare the derived allele frequency spectra of segregating sites in CNEEs, intronic regions, and coding positions that cause amnio acid changes. The segregating sites in intronic regions should be a conservative representation of neutrally evolving positions since introns are thought to contain some regions under purifying selection along with a majority of positions evolving without constraint. The derived allele frequency spectrum of segregating sites found in CNEEs is shifted towards lower frequencies relative to the set of intronic regions. This is characteristic of regions under significant negative selection where mutations are likely to be deleterious and rarely reach high frequencies in the population. The spectral shift for CNEEs is not as great as that for non-synonymous changes in coding regions.

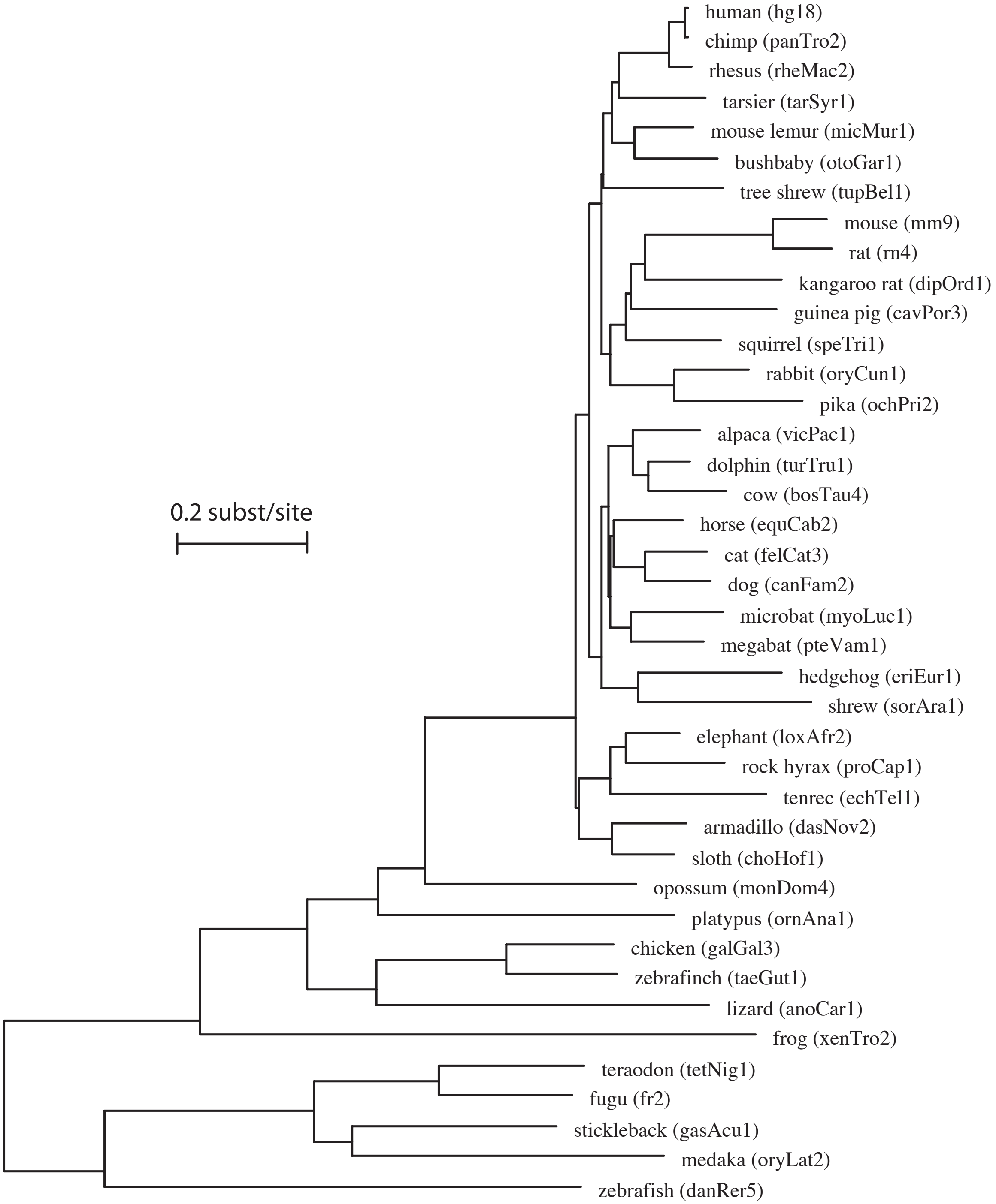


Fig. S3

The topology and branch lengths of the species used in our 40-way alignment. We used 4-fold degenerate sites in codons to construct the neutral tree (see Methods).

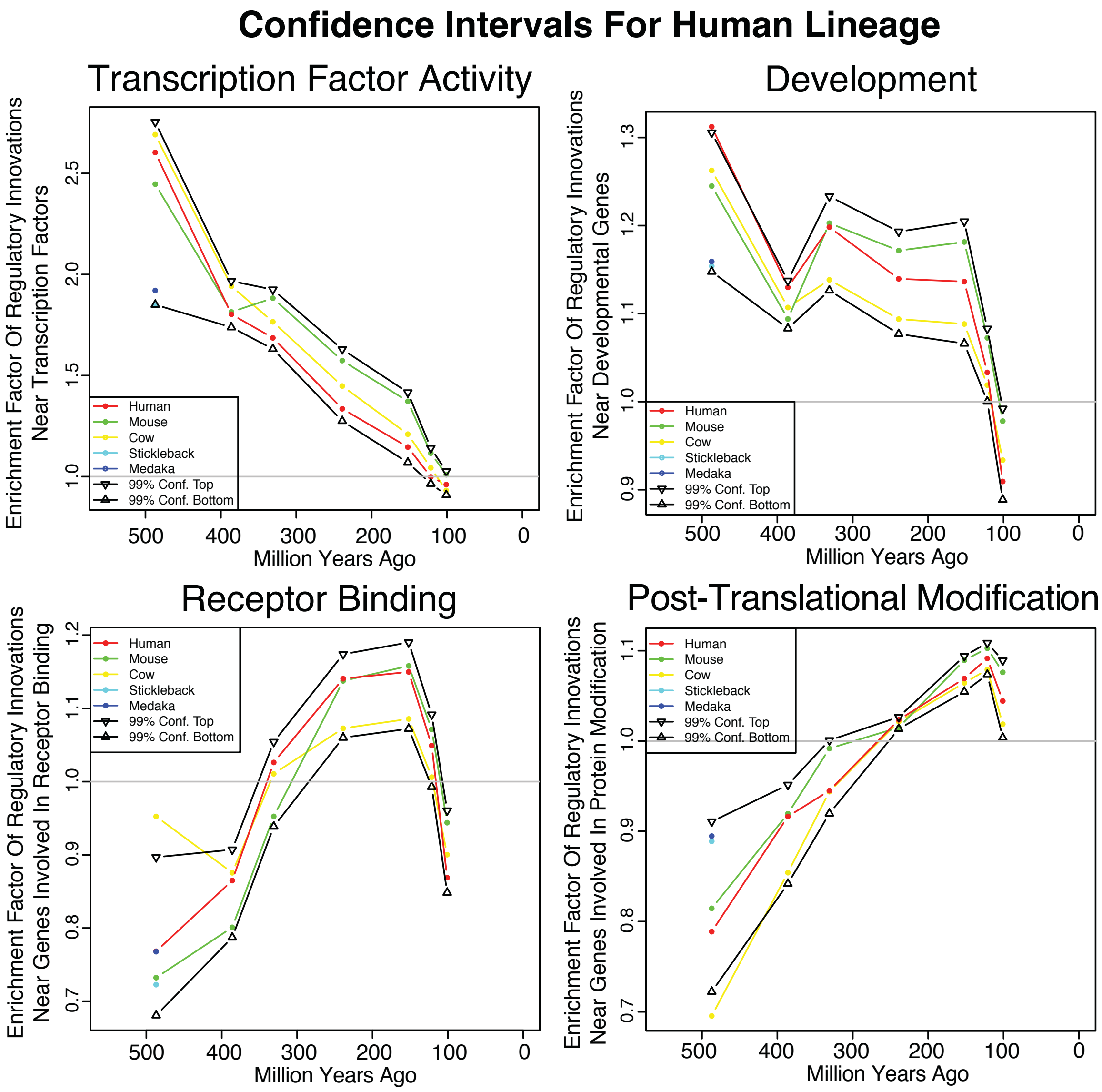


Fig. S4

Confidence intervals from the vertebrate ancestor to placental mammals. The large number of CNEEs assigned to each branch causes there to be little uncertainty in the enrichment value due to under-sampling. However, there is uncertainty due to gene annotation, assembly, genome evolution, and other variables. By assuming that the noise added to the enrichment value is normally distributed, we computed 99% confidence intervals for the enrichment value on each branch from the vertebrate ancestor to placental mammals.

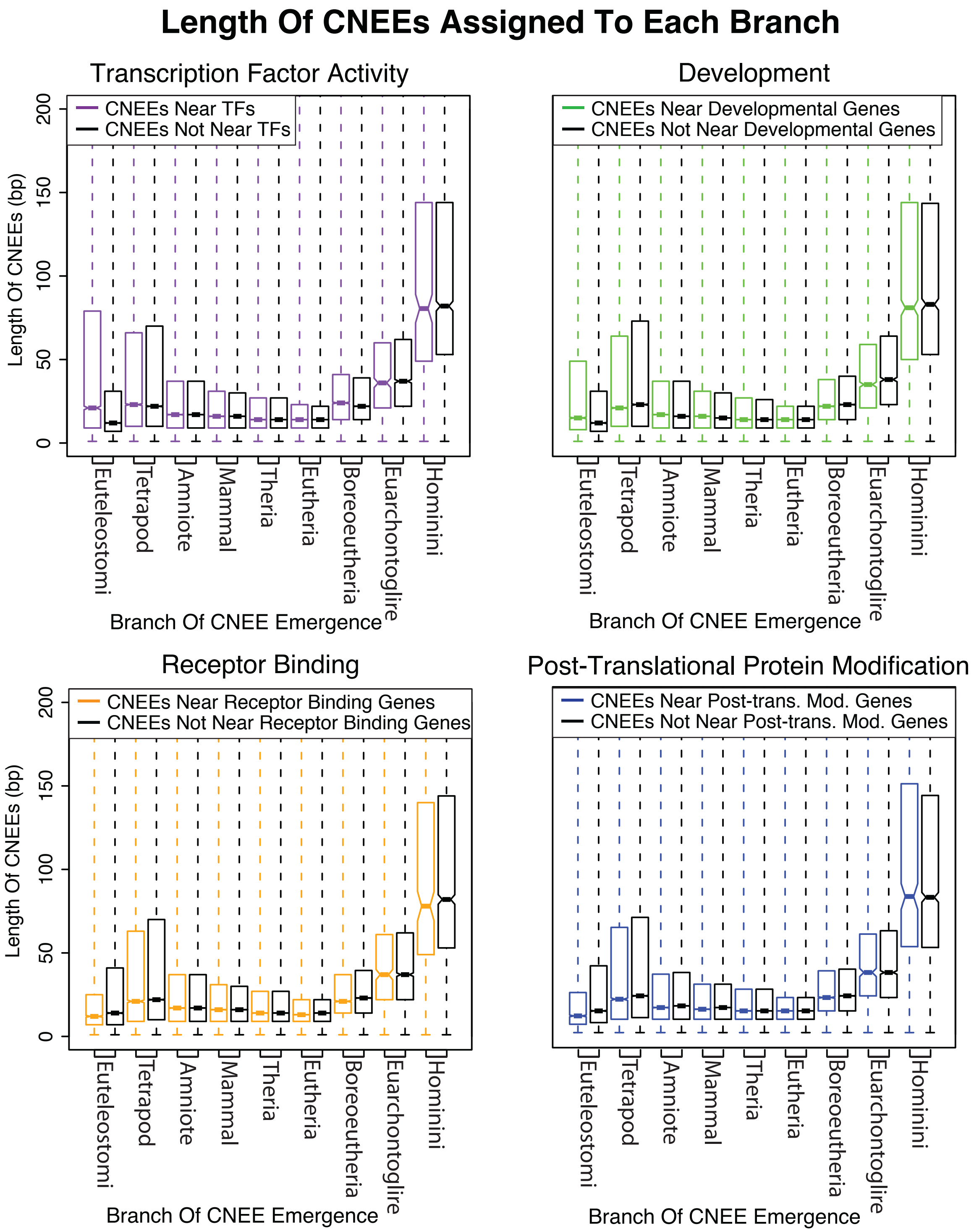


Fig. S5

The length in base pairs of human CNEEs for each of the four GO terms we focus on in the main text. There is a bias for the very recent CNEEs to be longer and to a lesser degree for the very ancient CNEEs to be slightly longer. However, there does not appear to be a consistent trend for the set of CNEEs associated with any of the four GO terms to be biased in their length.

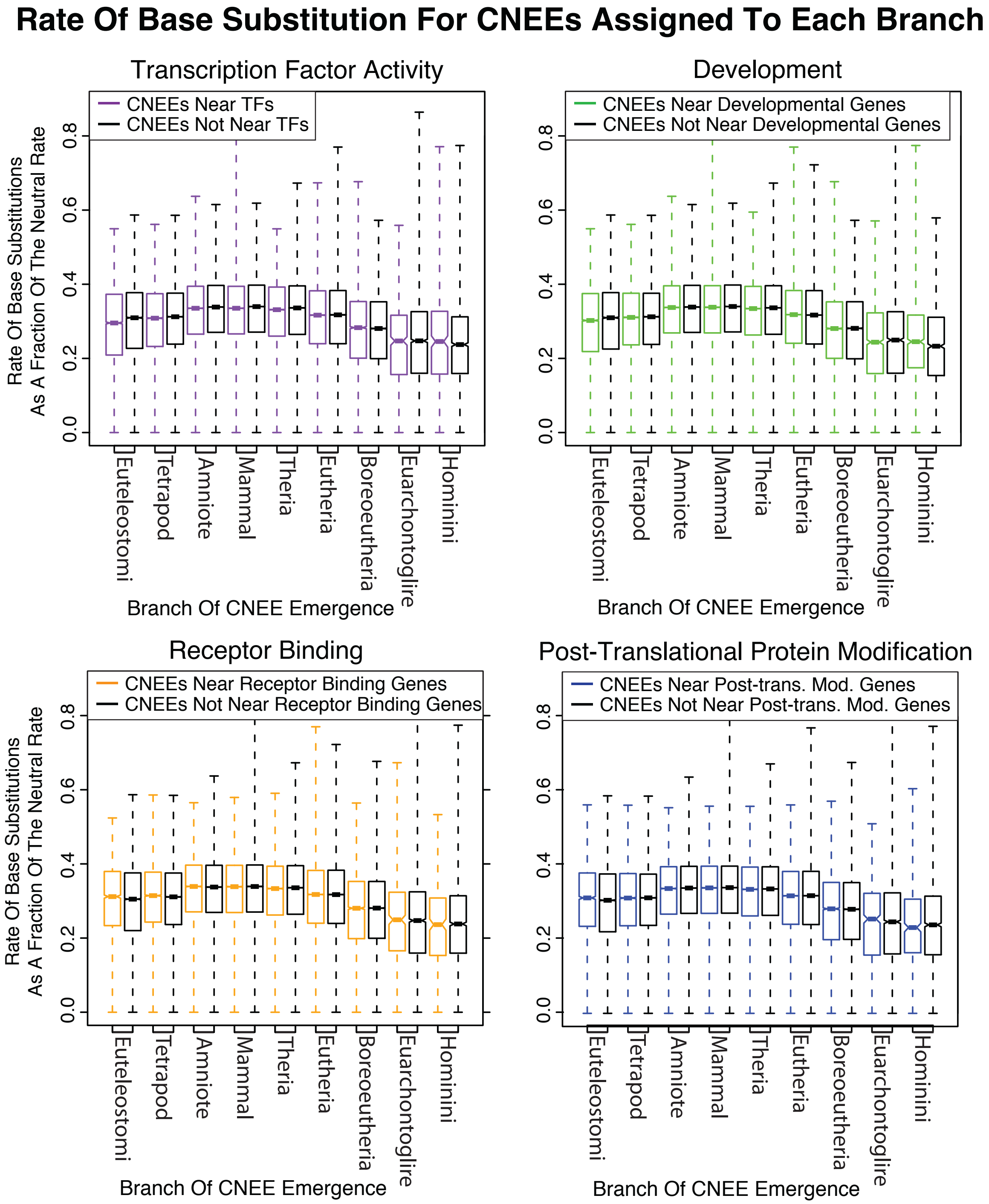


Fig. S6

The rate of base substitution, as a fraction of the neutral rate, for human CNEEs associated with each of the four GO terms we focus on in the main text. There is a bias for the very recent CNEEs to be slower evolving and to a lesser degree for the very ancient CNEEs to be slower evolving. However, there does not appear to be a consistent trend for the set of CNEEs associated with any of the four GO terms to be consistently faster or slower evolving than other CNEEs.

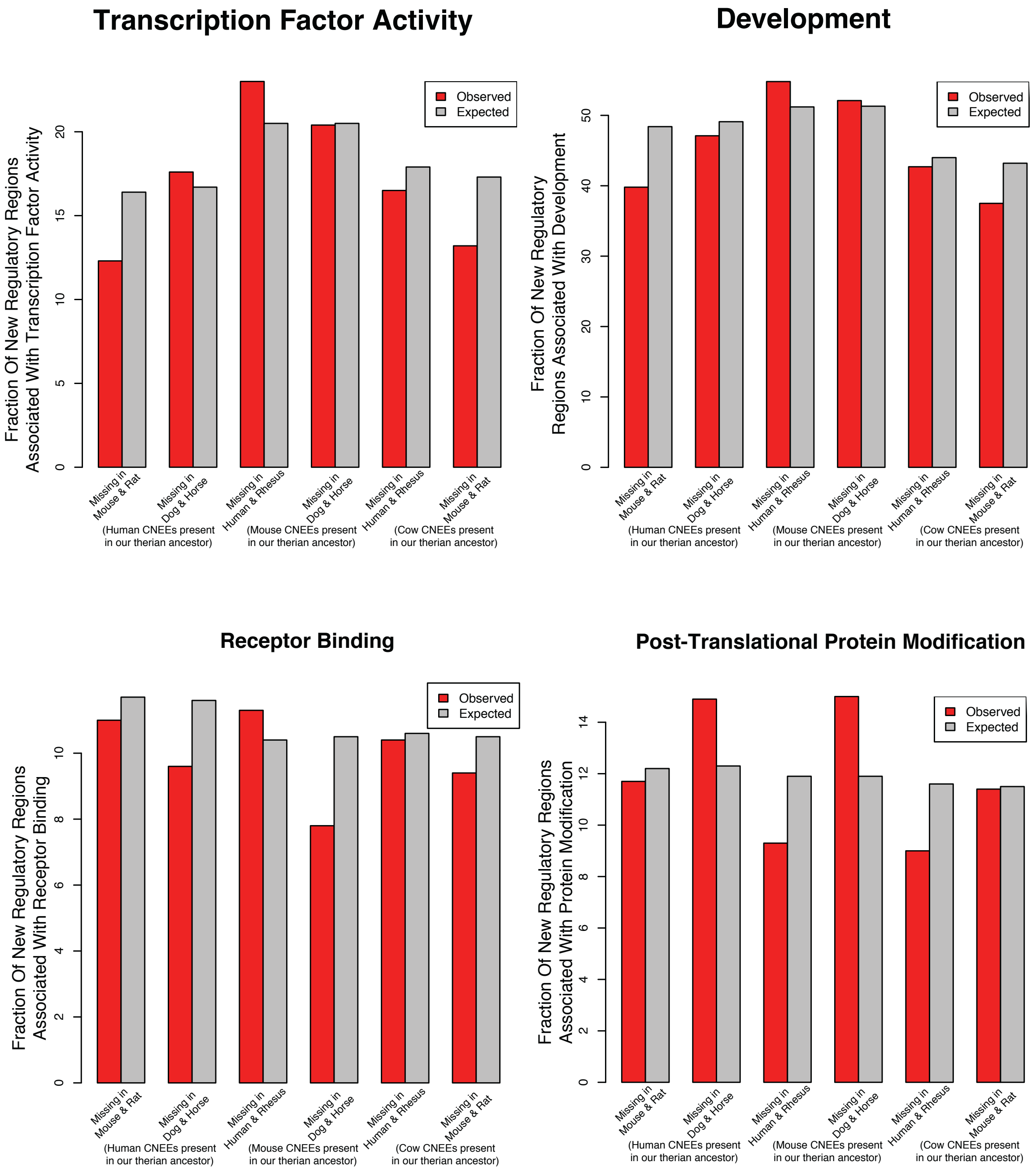


Fig. S7

For each of the human, mouse, and cow CNEE sets we identified a subset that was reliably present in our therian ancestor. We then identified which of these therian CNEEs have been putatively lost in humans and rhesus, mouse and rat, or dog and horse. We compared the set of lost CNEEs to the set of therian CNEEs to understand if the lost CNEEs were enriched or depleted for regulating genes with particular functions. We did not see any enrichments that were consistent across the sets.

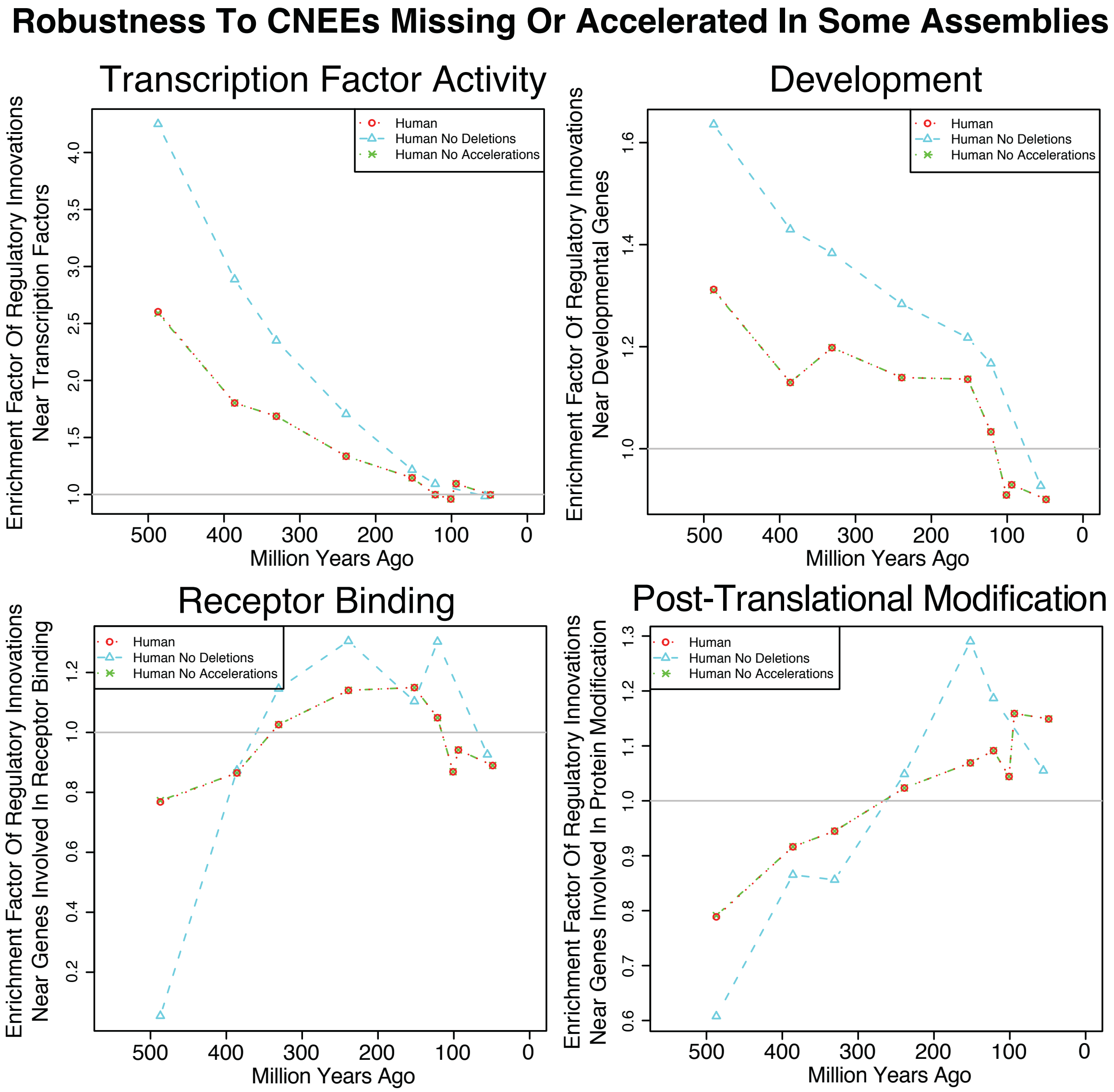


Fig. S8

The analysis of the human lineage when using CNEEs not missing from an assembly or not having an increased rate of evolution in a subtree. We utilize a number of low-coverage mammalian assemblies and therefore many CNEEs are absent in at least one assembly where their presence would be expected, given their branch of origin. The three periods of regulatory innovation are still present when using only the subset of CNEEs whose alignment does not imply any deletions. We also identified 679 CNEEs that have undergone loss of constraint in a subtree. The removal of these CNEEs from our main set of human CNEEs does not remove or diminish the trends we see in the set as a whole.

Robustness To Alignments, Gene Sets, And Association Rules

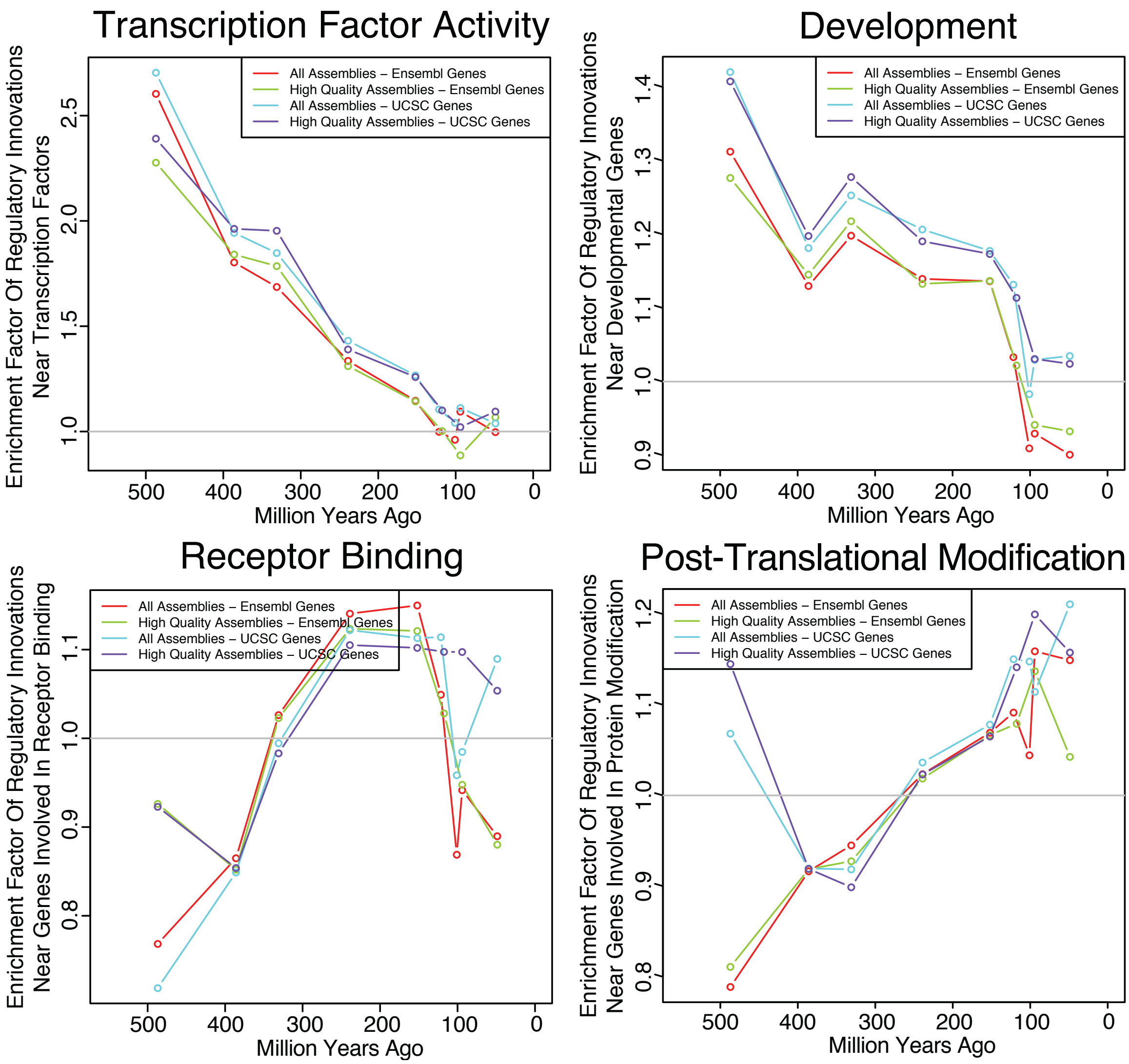


Fig. S9

The analysis for the human lineage when using two different vertebrate-wide alignments and when using two different gene sets and association methods between the CNEEs and the genes they are likely to regulate (see Methods). The analysis for both the transcription factors and developmental genes is highly robust and the analysis for "receptor binding" and "post-translational protein modification" are qualitatively similar. The findings for developmental genes and transcription factors are highly robust to any changes in alignment methods, gene sets, or association rules. The combination of the alignment using all 40 species and UCSC genes detects a resurgence of enrichment for receptor binding genes after the speciation of rodents, but the other three analyses show a decline. When using only the well-assembled species and the Ensembl gene set it appears that the most recent enrichment for protein modification may have already begun a decline, but the other three analyses depict a continued enrichment on the human lineage.

Species A	Species B	Divergence Date	Reference	Midpoint of branch leading to divergence (used as date for CNEEs born on the branch)
Human	Lamprey	552	(*44,45*)	N/A
Human	Stickleback	422	(*46,45*)	487
Human	Frog	350	(*46,45*)	386
Human	Chicken	312	(*46,45*)	331
Human	Platypus	166	(*47,45*)	239
Human	Opossum	138	(*47,45*)	152
Human	Elephant	105	(*46,48,45*)	121.5
Human	Cow	97	(*48,45*)	101
Human	Mouse	91	(*48,45*)	94
Human	Chimp	6.5	(*46,45*)	48.75
Mouse	Rat	12.3	(*48,45*)	51.65
Cow	Dolphin	53.5	(*48,45*)	75.25
Stickleback	Zebrafish	307	(*49*)	364.5
Stickleback	Fugu	165		236
Stickleback	Medaka	138		151.5

Table S1.

Divergence dates used in the analysis. For the mammalian lineages we use speciation dates that incorporate molecular evidence (*44*,*46*,*47*,*48*), yet adjust these dates to fit within a range of dates supported by paleontological data (*45*). We estimated divergence dates for stickleback-fugu and stickleback-medaka based on the molecular divergence since the split with zebrafish 307 Mya.

Reference Genome	Midpoint of Branch (Mya)	CNEEs Assigned to Branch	P-value for TF Enrichment	P-value for Dev Enrichment	P-value for Recep Binding Enrichment	P-value for Post-translational Mod Enrichment
Human	487	36857	~0	~0	1	1
Human	386	136020	~0	~0	1	1
Human	331	474232	~0	~0	4E-10	1
Human	239	574127	~0	~0	1E-280	2E-11
Human	152	526207	1E-290	~0	1E-293	1E-76
Human	121.5	1141626	0.71	1E-190	1E-70	1E-280
Human	101	68920	1	1	1	9E-06
Human	94	4101	0.02	1	0.91	2E-04
Human	48.75	2819	0.53	1	0.98	2E-03
Mouse	487	33564	~0	1E-300	1	1
Mouse	386	110141	~0	1E-140	1	1
Mouse	331	360375	~0	~0	1	0.97
Mouse	239	387706	~0	~0	1E-170	3E-04
Mouse	152	368687	~0	~0	1E-220	1E-80
Mouse	121.5	1159116	~0	~0	1E-140	~0
Mouse	101	116442	0.14	1	1	1E-20
Mouse	94	12410	0.13	1	0.91	0.04
Mouse	51.65	17770	0.01	1	1	4E-06
Cow	487	33281	~0	1E-290	1	1
Cow	386	111895	~0	1E-160	1	1
Cow	331	436363	~0	~0	0.01	1
Cow	239	479774	~0	~0	1E-60	1E-07
Cow	152	457757	~0	~0	1E-80	1E-50
Cow	121.5	1138160	1E-40	1E-50	0.02	1E-200
Cow	101	68898	1	1	1	0.04
Cow	75.25	12426	1	1	1	4E-05
Stickleback	487	20542	~0	1E-70	1	1
Stickleback	364.5	55437	~0	1E-290	0.68	1
Stickleback	236	258749	1E-170	~0	1E-90	0.77
Stickleback	151.5	29029	1	1E-08	0.9	0.4
Medaka	487	19357	~0	1E-70	1	1
Medaka	364.5	43148	~0	1E-290	1	1
Medaka	236	212629	~0	~0	1E-140	7E-05
Medaka	151.5	42234	1E-08	1E-50	1E-13	0.05

Table S2.

Number of CNEEs and enrichment p-values for each branch of the human, mouse, cow, stickleback, and medaka lineages. The visual representation of these enrichments is shown in figure 1 of the main text. More detailed information about the speciation events flanking each branch is available in supplementary table S1.

Gene Ontology Term	Slope Of Model (percent of regulatory innovations near GO term / My)
transcription	-4.26E-04
nervous system development	-4.18E-04
nucleic acid metabolic process	-4.08E-04
DNA binding	-4.04E-04
cellular macromolecule biosynthetic process	-3.94E-04
macromolecule biosynthetic process	-3.91E-04
nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	-3.88E-04
transcription regulator activity	-3.85E-04
gene expression	-3.85E-04
regulation of transcription	-3.78E-04
organ development	-3.75E-04
system development	-3.70E-04
transcription factor activity	-3.69E-04
anatomical structure morphogenesis	-3.68E-04
cellular nitrogen compound metabolic process	-3.68E-04
anatomical structure development	-3.64E-04
regulation of nitrogen compound metabolic process	-3.62E-04
regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	-3.60E-04
regulation of gene expression	-3.60E-04
regulation of macromolecule biosynthetic process	-3.57E-04
regulation of metabolic process	-3.55E-04
multicellular organismal development	-3.55E-04
developmental process	-3.55E-04
nitrogen compound metabolic process	-3.54E-04
regulation of biosynthetic process	-3.54E-04
regulation of cellular biosynthetic process	-3.51E-04
regulation of cellular metabolic process	-3.50E-04
nucleic acid binding	-3.50E-04
positive regulation of cellular process	-3.46E-04
nucleus	-3.45E-04
regulation of primary metabolic process	-3.40E-04
cellular biosynthetic process	-3.37E-04
regulation of macromolecule metabolic process	-3.36E-04
central nervous system development	-3.35E-04
sequence-specific DNA binding	-3.35E-04
biosynthetic process	-3.33E-04
regulation of transcription, DNA-dependent	-3.26E-04
multicellular organismal process	-3.24E-04
regulation of RNA metabolic process	-3.22E-04
neurogenesis	-3.20E-04

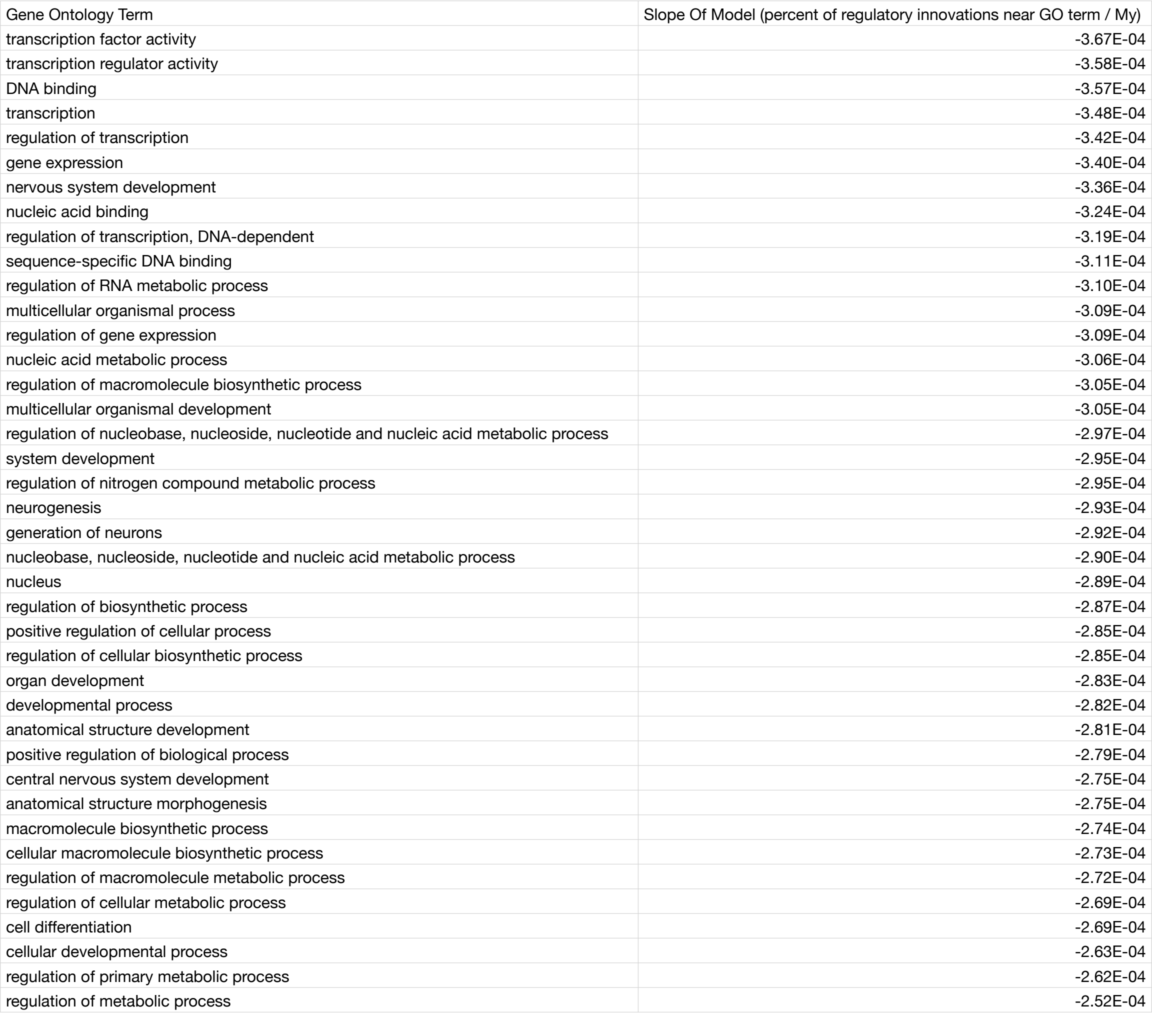
Table S3.

Gene Ontology terms showing the sharpest decrease in rate of regulatory innovations on the human lineage.

Gene Ontology Term	Slope Of Model (percent of regulatory innovations near GO term / My)
DNA binding	-3.77E-04
nucleic acid binding	-3.51E-04
transcription	-3.43E-04
nucleic acid metabolic process	-3.28E-04
gene expression	-3.23E-04
regulation of transcription	-3.20E-04
nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	-3.19E-04
transcription regulator activity	-3.16E-04
transcription factor activity	-3.13E-04
macromolecule biosynthetic process	-3.03E-04
cellular macromolecule biosynthetic process	-3.03E-04
nucleus	-3.00E-04
cellular nitrogen compound metabolic process	-2.99E-04
nervous system development	-2.97E-04
regulation of macromolecule biosynthetic process	-2.95E-04
regulation of transcription, DNA-dependent	-2.95E-04
regulation of gene expression	-2.94E-04
regulation of RNA metabolic process	-2.91E-04
nitrogen compound metabolic process	-2.86E-04
cell differentiation	-2.76E-04
cellular developmental process	-2.76E-04
regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	-2.74E-04
sequence-specific DNA binding	-2.72E-04
regulation of nitrogen compound metabolic process	-2.71E-04
regulation of biosynthetic process	-2.68E-04
regulation of cellular biosynthetic process	-2.67E-04
anatomical structure development	-2.65E-04
cellular biosynthetic process	-2.60E-04
regulation of macromolecule metabolic process	-2.59E-04
generation of neurons	-2.53E-04
biosynthetic process	-2.52E-04
neurogenesis	-2.52E-04
organ development	-2.44E-04
central nervous system development	-2.44E-04
multicellular organismal process	-2.41E-04
regulation of primary metabolic process	-2.39E-04
developmental process	-2.39E-04
system development	-2.38E-04
anatomical structure morphogenesis	-2.36E-04
cellular macromolecule metabolic process	-2.33E-04

Table S4.

Gene Ontology terms showing the sharpest decrease in rate of regulatory innovations on the mouse lineage.



Gene Ontology Term	Slope Of Model (percent of regulatory innovations near GO term / My)
transcription factor activity	-3.67E-04
transcription regulator activity	-3.58E-04
DNA binding	-3.57E-04
transcription	-3.48E-04
regulation of transcription	-3.42E-04
gene expression	-3.40E-04
nervous system development	-3.36E-04
nucleic acid binding	-3.24E-04
regulation of transcription, DNA-dependent	-3.19E-04
sequence-specific DNA binding	-3.11E-04
regulation of RNA metabolic process	-3.10E-04
multicellular organismal process	-3.09E-04
regulation of gene expression	-3.09E-04
nucleic acid metabolic process	-3.06E-04
regulation of macromolecule biosynthetic process	-3.05E-04
multicellular organismal development	-3.05E-04
regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	-2.97E-04
system development	-2.95E-04
regulation of nitrogen compound metabolic process	-2.95E-04
neurogenesis	-2.93E-04
generation of neurons	-2.92E-04
nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	-2.90E-04
nucleus	-2.89E-04
regulation of biosynthetic process	-2.87E-04
positive regulation of cellular process	-2.85E-04
regulation of cellular biosynthetic process	-2.85E-04
organ development	-2.83E-04
developmental process	-2.82E-04
anatomical structure development	-2.81E-04
positive regulation of biological process	-2.79E-04
central nervous system development	-2.75E-04
anatomical structure morphogenesis	-2.75E-04
macromolecule biosynthetic process	-2.74E-04
cellular macromolecule biosynthetic process	-2.73E-04
regulation of macromolecule metabolic process	-2.72E-04
regulation of cellular metabolic process	-2.69E-04
cell differentiation	-2.69E-04
cellular developmental process	-2.63E-04
regulation of primary metabolic process	-2.62E-04
regulation of metabolic process	-2.52E-04

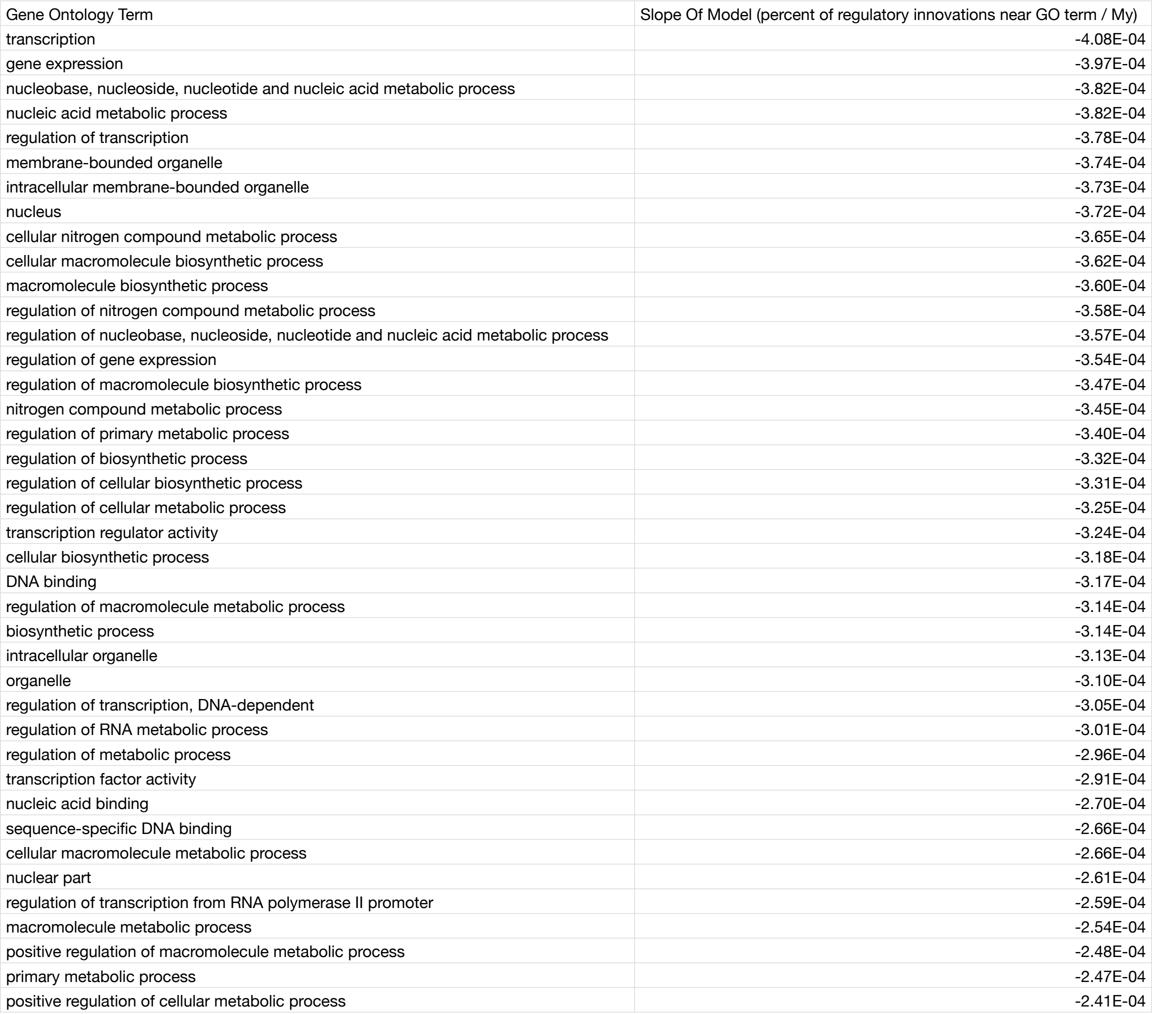
Table S5.

Gene Ontology terms showing the sharpest decrease in rate of regulatory innovations on the cow lineage.

Gene Ontology Term	Slope Of Model (percent of regulatory innovations near GO term / My)
transcription	-4.40E-04
regulation of transcription	-4.30E-04
gene expression	-4.23E-04
nucleus	-4.19E-04
nucleic acid metabolic process	-4.17E-04
nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	-4.09E-04
regulation of gene expression	-3.98E-04
regulation of nitrogen compound metabolic process	-3.98E-04
cellular macromolecule biosynthetic process	-3.98E-04
regulation of macromolecule biosynthetic process	-3.97E-04
macromolecule biosynthetic process	-3.96E-04
regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	-3.96E-04
DNA binding	-3.94E-04
cellular nitrogen compound metabolic process	-3.93E-04
regulation of biosynthetic process	-3.88E-04
regulation of cellular biosynthetic process	-3.87E-04
regulation of primary metabolic process	-3.85E-04
nitrogen compound metabolic process	-3.79E-04
regulation of cellular metabolic process	-3.75E-04
regulation of macromolecule metabolic process	-3.72E-04
membrane-bounded organelle	-3.65E-04
intracellular membrane-bounded organelle	-3.65E-04
nucleic acid binding	-3.63E-04
transcription regulator activity	-3.53E-04
biosynthetic process	-3.53E-04
regulation of transcription, DNA-dependent	-3.48E-04
cellular biosynthetic process	-3.47E-04
regulation of RNA metabolic process	-3.47E-04
regulation of metabolic process	-3.45E-04
transcription factor activity	-3.44E-04
cellular macromolecule metabolic process	-3.38E-04
organelle	-3.36E-04
intracellular organelle	-3.35E-04
sequence-specific DNA binding	-3.34E-04
macromolecule metabolic process	-3.27E-04
metabolic process	-3.16E-04
cellular metabolic process	-2.99E-04
primary metabolic process	-2.90E-04
organ development	-2.71E-04
zinc ion binding	-2.69E-04

Table S6.

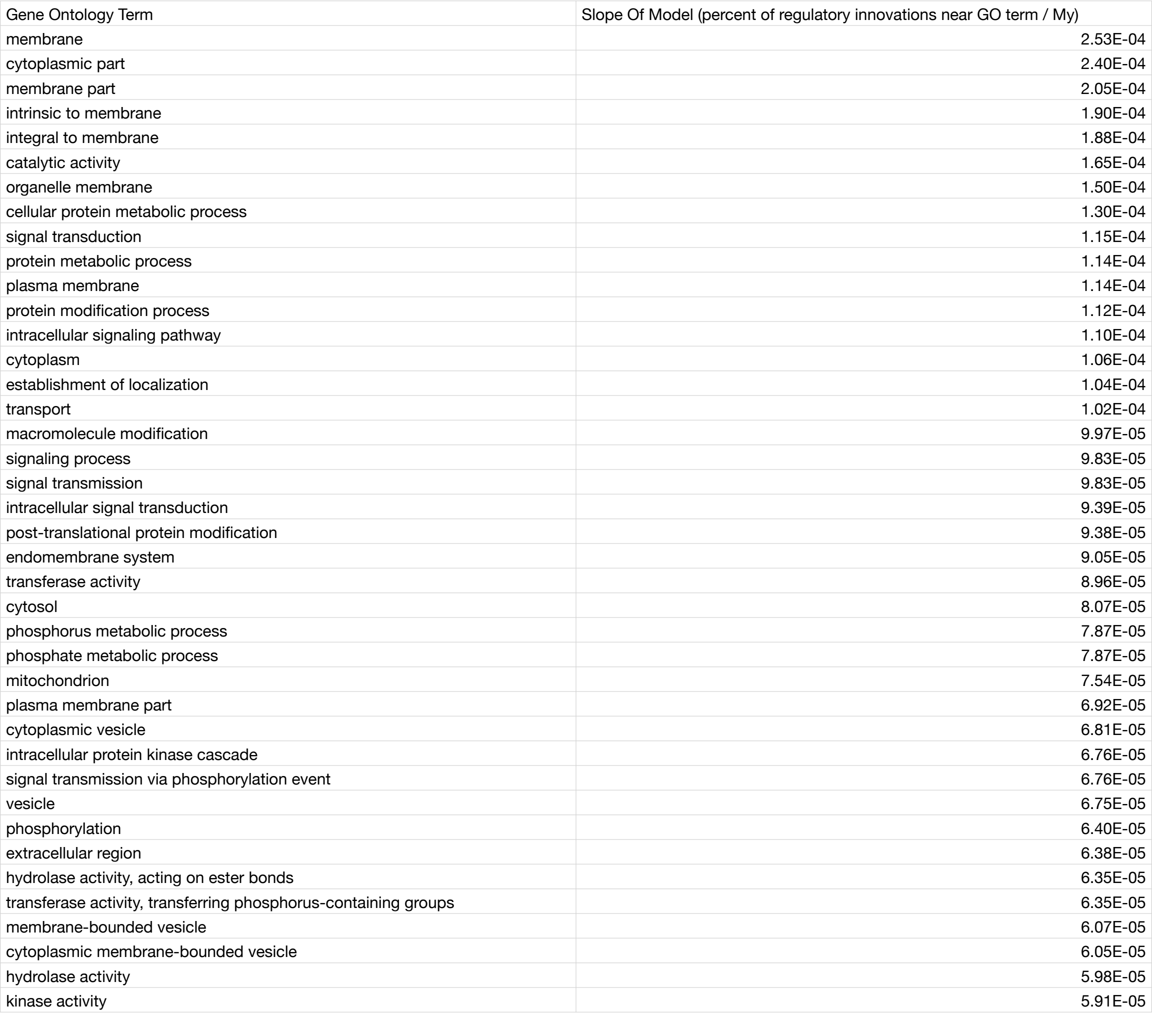
Gene Ontology terms showing the sharpest decrease in rate of regulatory innovations on the stickleback lineage.



Gene Ontology Term	Slope Of Model (percent of regulatory innovations near GO term / My)
transcription	-4.08E-04
gene expression	-3.97E-04
nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	-3.82E-04
nucleic acid metabolic process	-3.82E-04
regulation of transcription	-3.78E-04
membrane-bounded organelle	-3.74E-04
intracellular membrane-bounded organelle	-3.73E-04
nucleus	-3.72E-04
cellular nitrogen compound metabolic process	-3.65E-04
cellular macromolecule biosynthetic process	-3.62E-04
macromolecule biosynthetic process	-3.60E-04
regulation of nitrogen compound metabolic process	-3.58E-04
regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	-3.57E-04
regulation of gene expression	-3.54E-04
regulation of macromolecule biosynthetic process	-3.47E-04
nitrogen compound metabolic process	-3.45E-04
regulation of primary metabolic process	-3.40E-04
regulation of biosynthetic process	-3.32E-04
regulation of cellular biosynthetic process	-3.31E-04
regulation of cellular metabolic process	-3.25E-04
transcription regulator activity	-3.24E-04
cellular biosynthetic process	-3.18E-04
DNA binding	-3.17E-04
regulation of macromolecule metabolic process	-3.14E-04
biosynthetic process	-3.14E-04
intracellular organelle	-3.13E-04
organelle	-3.10E-04
regulation of transcription, DNA-dependent	-3.05E-04
regulation of RNA metabolic process	-3.01E-04
regulation of metabolic process	-2.96E-04
transcription factor activity	-2.91E-04
nucleic acid binding	-2.70E-04
sequence-specific DNA binding	-2.66E-04
cellular macromolecule metabolic process	-2.66E-04
nuclear part	-2.61E-04
regulation of transcription from RNA polymerase II promoter	-2.59E-04
macromolecule metabolic process	-2.54E-04
positive regulation of macromolecule metabolic process	-2.48E-04
primary metabolic process	-2.47E-04
positive regulation of cellular metabolic process	-2.41E-04

Table S7.

Gene Ontology terms showing the sharpest decrease in rate of regulatory innovations on the medaka lineage.



Gene Ontology Term	Slope Of Model (percent of regulatory innovations near GO term / My)
membrane	2.53E-04
cytoplasmic part	2.40E-04
membrane part	2.05E-04
intrinsic to membrane	1.90E-04
integral to membrane	1.88E-04
catalytic activity	1.65E-04
organelle membrane	1.50E-04
cellular protein metabolic process	1.30E-04
signal transduction	1.15E-04
protein metabolic process	1.14E-04
plasma membrane	1.14E-04
protein modification process	1.12E-04
intracellular signaling pathway	1.10E-04
cytoplasm	1.06E-04
establishment of localization	1.04E-04
transport	1.02E-04
macromolecule modification	9.97E-05
signaling process	9.83E-05
signal transmission	9.83E-05
intracellular signal transduction	9.39E-05
post-translational protein modification	9.38E-05
endomembrane system	9.05E-05
transferase activity	8.96E-05
cytosol	8.07E-05
phosphorus metabolic process	7.87E-05
phosphate metabolic process	7.87E-05
mitochondrion	7.54E-05
plasma membrane part	6.92E-05
cytoplasmic vesicle	6.81E-05
intracellular protein kinase cascade	6.76E-05
signal transmission via phosphorylation event	6.76E-05
vesicle	6.75E-05
phosphorylation	6.40E-05
extracellular region	6.38E-05
hydrolase activity, acting on ester bonds	6.35E-05
transferase activity, transferring phosphorus-containing groups	6.35E-05
membrane-bounded vesicle	6.07E-05
cytoplasmic membrane-bounded vesicle	6.05E-05
hydrolase activity	5.98E-05
kinase activity	5.91E-05

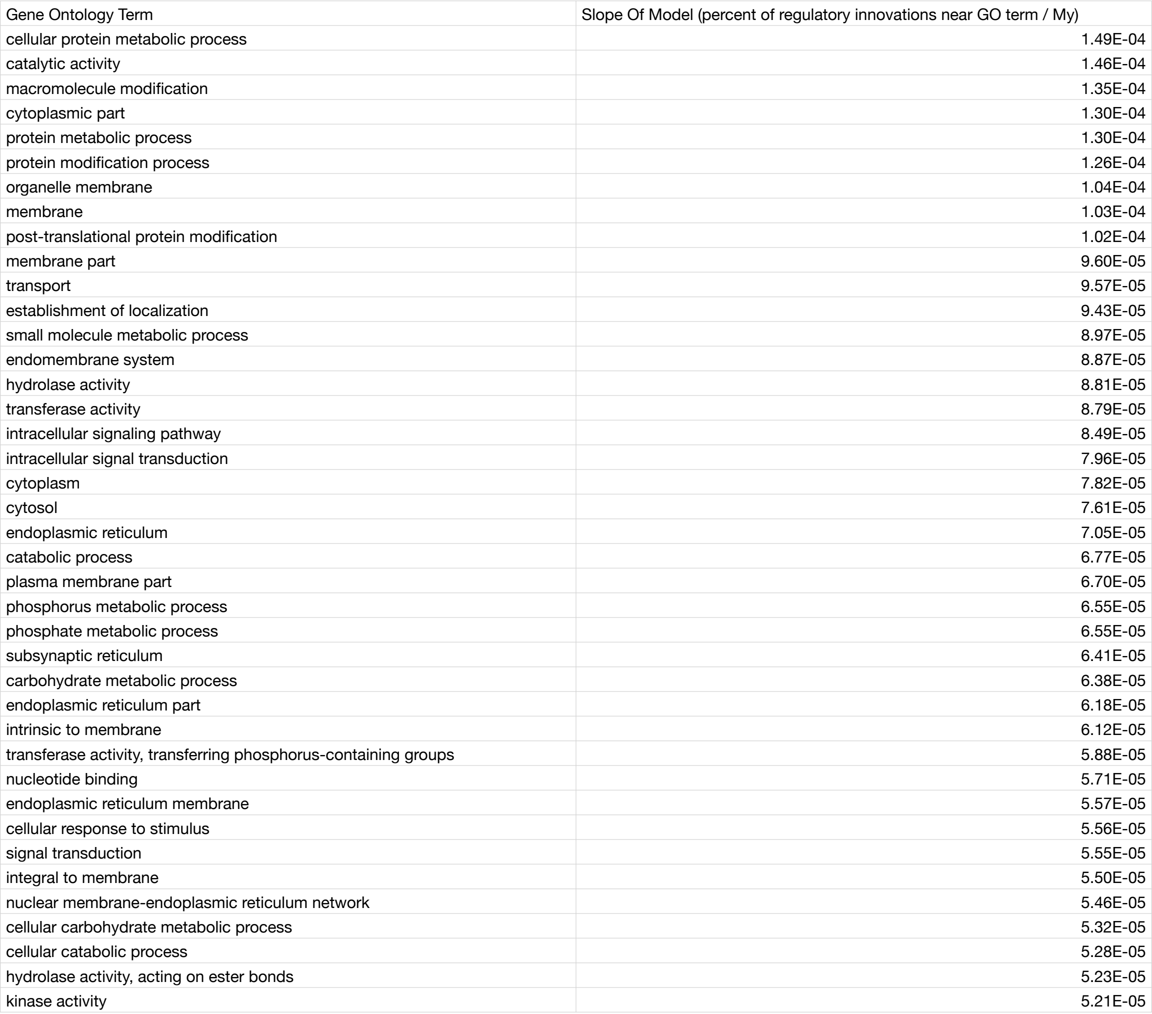
Table S8.

Gene Ontology terms showing the sharpest increase in rate of regulatory innovations on the human lineage.

Gene Ontology Term	Slope Of Model (percent of regulatory innovations near GO term / My)
membrane	2.48E-04
membrane part	2.08E-04
cytoplasmic part	1.91E-04
intrinsic to membrane	1.61E-04
transport	1.56E-04
establishment of localization	1.56E-04
integral to membrane	1.55E-04
catalytic activity	1.43E-04
plasma membrane	1.40E-04
plasma membrane part	1.35E-04
intracellular signaling pathway	1.26E-04
organelle membrane	1.23E-04
signal transduction	1.19E-04
cytoplasm	1.18E-04
signaling process	1.08E-04
signal transmission	1.08E-04
endomembrane system	1.04E-04
intracellular signal transduction	1.04E-04
response to chemical stimulus	8.48E-05
cell fraction	8.33E-05
cellular protein metabolic process	8.29E-05
response to stress	8.20E-05
transporter activity	8.07E-05
localization	8.07E-05
intrinsic to plasma membrane	8.06E-05
integral to plasma membrane	7.97E-05
extracellular region	7.70E-05
endoplasmic reticulum	7.47E-05
protein modification process	7.10E-05
protein metabolic process	7.02E-05
hydrolase activity	6.69E-05
intracellular protein kinase cascade	6.42E-05
signal transmission via phosphorylation event	6.42E-05
membrane fraction	6.37E-05
transferase activity	6.37E-05
vesicle	6.29E-05
transmembrane transporter activity	6.28E-05
Golgi apparatus	6.28E-05
response to organic substance	6.26E-05
cytosol	6.25E-05

Table S9.

Gene Ontology terms showing the sharpest increase in rate of regulatory innovations on the mouse lineage.



Gene Ontology Term	Slope Of Model (percent of regulatory innovations near GO term / My)
cellular protein metabolic process	1.49E-04
catalytic activity	1.46E-04
macromolecule modification	1.35E-04
cytoplasmic part	1.30E-04
protein metabolic process	1.30E-04
protein modification process	1.26E-04
organelle membrane	1.04E-04
membrane	1.03E-04
post-translational protein modification	1.02E-04
membrane part	9.60E-05
transport	9.57E-05
establishment of localization	9.43E-05
small molecule metabolic process	8.97E-05
endomembrane system	8.87E-05
hydrolase activity	8.81E-05
transferase activity	8.79E-05
intracellular signaling pathway	8.49E-05
intracellular signal transduction	7.96E-05
cytoplasm	7.82E-05
cytosol	7.61E-05
endoplasmic reticulum	7.05E-05
catabolic process	6.77E-05
plasma membrane part	6.70E-05
phosphorus metabolic process	6.55E-05
phosphate metabolic process	6.55E-05
subsynaptic reticulum	6.41E-05
carbohydrate metabolic process	6.38E-05
endoplasmic reticulum part	6.18E-05
intrinsic to membrane	6.12E-05
transferase activity, transferring phosphorus-containing groups	5.88E-05
nucleotide binding	5.71E-05
endoplasmic reticulum membrane	5.57E-05
cellular response to stimulus	5.56E-05
signal transduction	5.55E-05
integral to membrane	5.50E-05
nuclear membrane-endoplasmic reticulum network	5.46E-05
cellular carbohydrate metabolic process	5.32E-05
cellular catabolic process	5.28E-05
hydrolase activity, acting on ester bonds	5.23E-05
kinase activity	5.21E-05

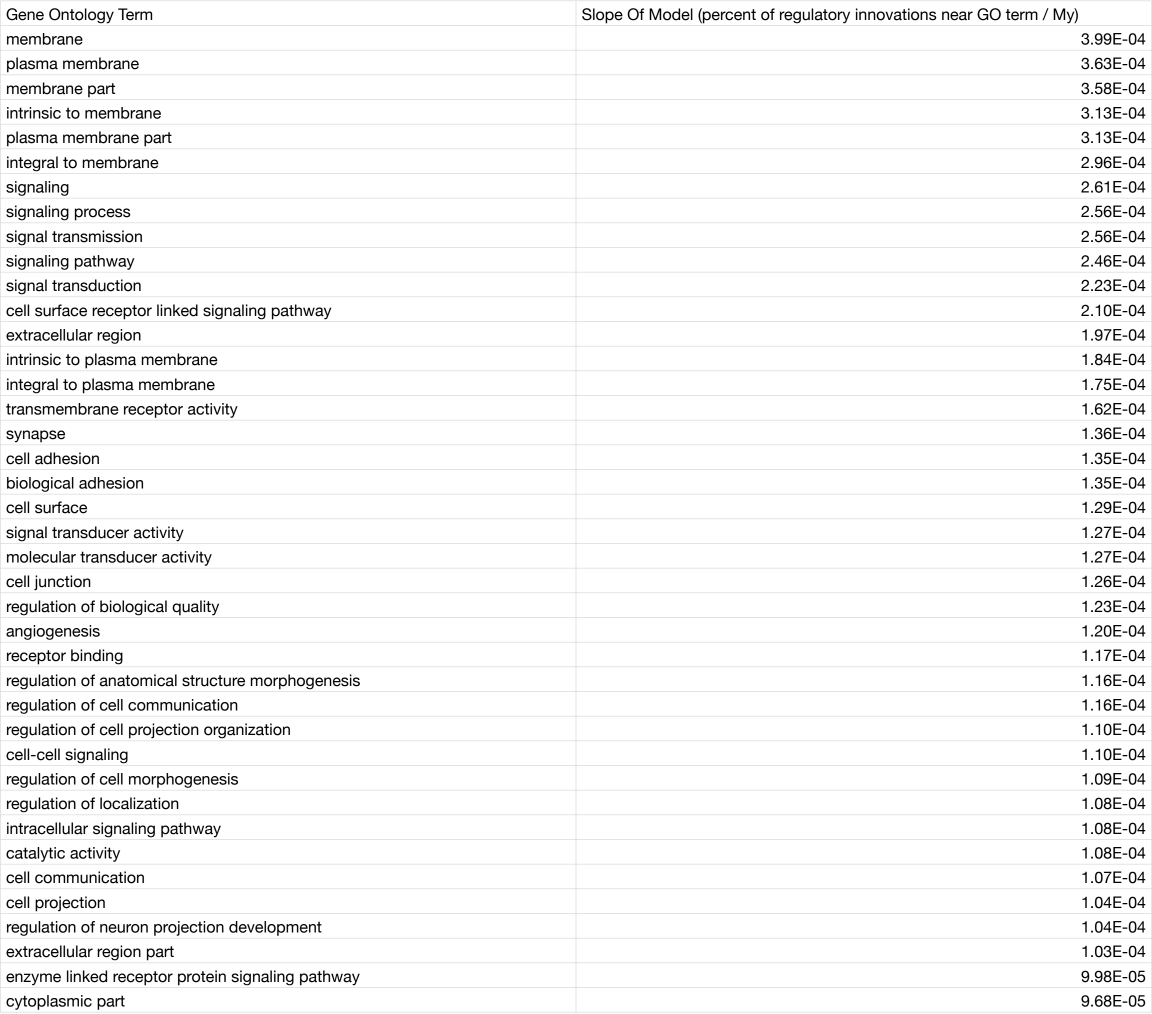
Table S10.

Gene Ontology terms showing the sharpest increase in rate of regulatory innovations on the cow lineage.

Gene Ontology Term	Slope Of Model (percent of regulatory innovations near GO term / My)
membrane	4.40E-04
membrane part	3.93E-04
plasma membrane	3.77E-04
intrinsic to membrane	3.73E-04
integral to membrane	3.56E-04
plasma membrane part	2.46E-04
signaling process	2.19E-04
signal transmission	2.19E-04
signaling pathway	2.17E-04
signaling	2.11E-04
signal transduction	1.95E-04
intrinsic to plasma membrane	1.75E-04
cell surface receptor linked signaling pathway	1.71E-04
integral to plasma membrane	1.68E-04
extracellular region	1.64E-04
signal transducer activity	1.54E-04
molecular transducer activity	1.54E-04
transmembrane receptor activity	1.51E-04
regulation of cell communication	1.25E-04
cell adhesion	1.25E-04
biological adhesion	1.25E-04
receptor activity	1.18E-04
synapse	1.14E-04
cell junction	1.10E-04
intracellular signaling pathway	1.10E-04
regulation of cell projection organization	1.07E-04
regulation of neuron projection development	1.04E-04
regulation of signaling process	1.02E-04
regulation of signal transduction	9.69E-05
cell surface	9.63E-05
intracellular signal transduction	9.39E-05
cell communication	9.12E-05
extracellular region part	8.79E-05
cell-cell signaling	8.64E-05
regulation of axonogenesis	8.63E-05
regulation of cell morphogenesis	8.60E-05
receptor binding	8.57E-05
regulation of cell morphogenesis involved in differentiation	8.49E-05
regulation of anatomical structure morphogenesis	8.46E-05
transferase activity	8.27E-05

Table S11.

Gene Ontology terms showing the sharpest increase in rate of regulatory innovations on the stickleback lineage.



Gene Ontology Term	Slope Of Model (percent of regulatory innovations near GO term / My)
membrane	3.99E-04
plasma membrane	3.63E-04
membrane part	3.58E-04
intrinsic to membrane	3.13E-04
plasma membrane part	3.13E-04
integral to membrane	2.96E-04
signaling	2.61E-04
signaling process	2.56E-04
signal transmission	2.56E-04
signaling pathway	2.46E-04
signal transduction	2.23E-04
cell surface receptor linked signaling pathway	2.10E-04
extracellular region	1.97E-04
intrinsic to plasma membrane	1.84E-04
integral to plasma membrane	1.75E-04
transmembrane receptor activity	1.62E-04
synapse	1.36E-04
cell adhesion	1.35E-04
biological adhesion	1.35E-04
cell surface	1.29E-04
signal transducer activity	1.27E-04
molecular transducer activity	1.27E-04
cell junction	1.26E-04
regulation of biological quality	1.23E-04
angiogenesis	1.20E-04
receptor binding	1.17E-04
regulation of anatomical structure morphogenesis	1.16E-04
regulation of cell communication	1.16E-04
regulation of cell projection organization	1.10E-04
cell-cell signaling	1.10E-04
regulation of cell morphogenesis	1.09E-04
regulation of localization	1.08E-04
intracellular signaling pathway	1.08E-04
catalytic activity	1.08E-04
cell communication	1.07E-04
cell projection	1.04E-04
regulation of neuron projection development	1.04E-04
extracellular region part	1.03E-04
enzyme linked receptor protein signaling pathway	9.98E-05
cytoplasmic part	9.68E-05

Table S12.

Gene Ontology terms showing the sharpest increase in rate of regulatory innovations on the medaka lineage.