SUPPLEMENTARY INFORMATION

Supporting Information Corrected July 17, 2014

1. SUPPLEMENTARY METHODS

1.1 Python Genome Sequencing

A single *Python molurus bivittatus* female was obtained from a commercial breeder, euthanized and tissues preserved following protocols outlined in the Guidelines for Use of Live Amphibians and Reptiles in Field Research, established jointly by the American Society for Ichthyologists and Herpetologists, Herpetologist's League, and the Society for the Study of Amphibians and Reptiles (http://www.aaalac.org/accreditation/Guidelines_for_Use_of_Live_Amphibians_and_Reptiles.pdf). The specimen was deposited in the University of Texas at Arlington – Amphibian and Reptile Diversity Research Center's collection. This individual was used for all complete genome sequencing. Multiple whole genome shotgun libraries were prepared and sequenced, including the following: 454 FLX and 454 FLX+ shotgun libraries, paired end Illumina 300bp insert, 500bp insert, and 3kb mate pair.

1.2 Genome Assembly

The genome was assembled using two different approaches, and the results were later merged. First, all Illumina data, including data from shotgun and mate pair libraries, were assembled using *SoapDeNovo* v1.0.5 (1). This assembly (representing ~50x sequence coverage) resulted in 1,323,545 contigs, with a contig N50 size of 4,097 bp, and a total contig length of 1.4227 Gbp. The scaffold N50 for this assembly was 183,903 bp.

A second independent assembly was created using the 454 *Newbler* assembler based on all 454 reads plus 22.4 Gbp of Illumina shotgun data from the 500bp insert shotgun paired-end library; the complete raw data set was sub-sampled for this assembly because the *Newbler* assembler

suffered fatal errors when attempting to use all available short read Illumina data. This Newbler assembly resulted in a total of 375,259 contigs with a total length of 1.3041 Gbp. The contig N50 was 3,771 bp and a scaffold N50 of 20,227 bp.

The *SOAPdenovo* and *Newbler* python assemblies were merged into a single assembly using the Graph Accordance Assembler (GAA (2)). The GAA algorithm constructs an accordance graph to capture the mapping information between the target assembly, *SOAPdenovo* in this case, and the query assembly, *Newbler*. The merged assembly was further improved by iterative mapping and local assembly of short reads to eliminate as many gaps as possible. After gap closing efforts, the resulting 1.5090 Gbp assembly (including gaps) with 448,617 scaffolds was labeled as 5.0.1 (Supplementary Table S2). The final assembly resulted in 759,403 contigs, with a contig N50 size of 10,203 bp, and a total contig length of 1.4440 Gbp. The scaffold N50 for this assembly was 201,400 bp.

1.3 Genome Annotation

Annotations for the Python genome assembly were generated using the automated genome annotation pipeline MAKER (3-5), which aligns and filters EST and protein homology evidence, identifies repeats, produces *ab initio* gene predictions, infers 5' and 3' UTR, and integrates these data to produce final downstream gene models along with quality control statistics. Inputs for MAKER included the *Python molurus bivittatus* genome assembly, a snake-specific repeat library constructed using the complete python genome assembly, the complete king cobra genome assembly, and the sample sequencing of other snakes (see section 1.4-1.5 below) with repeats identified using RepeatModeler (http://www.repeatmasker.org/RepeatModeler.html) and classified further using Repclass (6). Gene annotations were made using a protein database combining the Uniprot/Swiss-Prot (7, 8) protein database and all sequences for *Python molurus* and *Anolis carolensis* from the NCBI protein database (9). *Ab initio* gene predictions were created by MAKER using the programs SNAP (10) and Augustus (11). Gene models were further improved by providing MAKER with all mRNAseq data generated in this study and others (12-14) for *Python molurus bivittatus*, which

were combined to generate a joint assembly of transcripts using Trinity (15). A total of three iterative runs of MAKER were used to produce the final gene set.

Following genome annotation, final gene models were analyzed using the program InterProScan (16) to identify putative protein domains. The final annotation set contained a total of 25,385 genes, 68% of which contain a protein domain as detected by IPRscan, and 85.5% of which have an annotation edit distance less than 0.5, consistent with a well annotated genome (4, 5, 17). The average gene length is 18,441 bp with median exon and intron lengths of 130 bp and 1,116 bp respectively (Supplementary Table S3).

1.4 Additional Genome Resources

To increase the research value of the python genome, and facilitate future investigation into specific regions of interest, we constructed a large insert genomic DNA Bacterial Artificial Chromosome (BAC) library. This BAC library was constructed from DNA from the same individual python that was used for whole genome sequencing. The library is estimated to represent approximately 5-fold coverage, and is comprised of 55,296 clones with an average insert size of 135 kb. This resource is publically available on a cost recovery basis through Amplicon Express (Pullman WA). The BAC library can be screened for specific genes of interest and individual alleles can be separated out from the mosaic haploid genome assembly presented herein.

1.5 Physiological, metabolic, and triglyceride analyses of fasted and postfeeding pythons

To identify postprandial changes in organ masses, Burmese pythons (833±15 g) were studied after a 30-day fast and at 1, 4, and 10 days (n=4 for each time period) following the consumption of a rodent meal equal in mass to 25% of the snake's body mass. All python feeding experiments, and subsequent sampling of tissues from these specimens, were conducted under approved IACUC protocols at the University of Alabama. Snakes were humanely sacrificed by severing the spinal cord immediately behind the head and from a mid-

ventral incision we extracted and weighed each organ (18). For fed snakes, the small intestine was emptied of its contents before weighing. Postprandial metabolic response was studied using closed-system respirometry of six Burmese python (659±45 g) prior to (following a 30-day fast) and following the consumption of rodent meals equaling 25% of snake body mass (19). Plasma triglycerides were measured from seven Burmese pythons (4720±430 g) prior to (following a 30-day fast) and following the consumption of rodent meals equaling 25% of snake body mass. Blood was drawn by cardiac puncture, centrifuged at 4000 rpm at 5°C, and triglycerides quantified from the plasma using Sigma reagents. We report true triglyceride concentration (mg/dL) following the subtraction of endogenous free glycerol.

1.6 Sample sequencing of additional snake genomes

Whole genome random shotgun libraries were constructed from DNA extracted from muscle or liver tissue from 10 additional snake species (see Supplementary Table S4). For each, DNA was extracted using phenol-chloroform-isoamyl alcohol methods. Shotgun genome libraries were prepared using the Roche 454 FLX shotgun genome rapid library kit and protocol except libraries were size-selected (at between 500-700bp) using a pippen prep. Each library was barcoded and run in a 1/8 plate of a 454 FLX sequencer using the FLX-XL Titanium kit, plates (70x75cm) and reagents. Raw reads were quality filtered and trimmed. New data was combined with existing data (13) for the Burmese Python and Copperhead (*Agkistrodon contortrix*). The Burmese Python sample is the same individual that was used for genome sequencing, thus allowing a direct comparison of repeat estimates between the complete assembled genome and the sample sequencing approach.

Mitochondrial reads were filtered based on blast searching against available snake mitochondrial genomes (following approach in (13)). In brief, reads with a score > 100 and a length > 75 were mapped using the 454 gsMapper software to the reference mitochondrial genomes, and resulting contigs were used as a reference for a second round of blastn. Reads with a score > 50 and length > 50 were then mapped back to the original reference sequence, and reads that successfully mapped to the reference sequence from both steps were assembled

using the 454 gsAssembler to create new contigs. These new contigs became the reference sequence for another round of blastn using all other reads. Any read with a blast score > 50 and a match length > 50 were iteratively added to the assembly to generate mitochondrial contigs. Because mitochondrial genomes were not always available for closely-related species, this step was repeated until no further improvement in the mitochondrial assembly was detected from successive rounds. All reads that did not assemble into the final mitochondrial contigs were considered to be nuclear genome reads in subsequent analyses. Finally, once the mitochondrial reads were filtered out, exact duplicate nuclear reads were discarded. Reads were considered duplicates if the first 100bp matched exactly. The number of reads and total bases collected after mitochondrial and duplicateread filtering is given in Supplementary Table S4. Genome size data for sampled species was approximated based on the most recent values for these or related species (20).

1.7 Repeat Element Analysis

There was little information on repeat elements of snakes prior to this study, and to increase this knowledge to better annotate snake repeat elements, repeat consensus sequences were combined from multiple species into a large joint snake repeat library for analyses of complete and sampled snake genomes. The current release of the Tetrapoda RepBase (RepBase Update 20090604 (21)) was used as the repeat library with RepeatMasker (22) to identify known repeat elements in the snake genomes, and the RepeatModeler pipeline (http://www.repeatmasker.org/RepeatModeler.html) to identify *de novo* repeat sequences in our snake datasets, based on the run parameters suggested as defaults by the program. The sequences were pre-annotated using our Bov-B/CR1 library to avoid misannotation due to the BOVB_VA compound element that exists in RepBase (13). To recover as many elements as possible in RepeatModeler analyses, the new *Anilius scytale, Boa constrictor, Casarea dussermieri, Crotalus atrox, Leptotyphlops dulcis, Loxocemus bicolor, Micrurus fulvius, Sibon nebulata, Thamnophis sirtalis*, and *Typhlops reticulatus* libraries were combined with the previously identified *Python molurus bivittatus* (same individual used in genome sequencing), and *Agkistrodon contortrix* libraries (13) and *de novo* libraries estimated from the complete

python and cobra genomes into a single joint snake library. Thus, each complete or sample-sequenced genome was annotated with exactly the same repeat library, thereby controlling for differences in sequencing depth for sample-sequenced genomes.

Repclass was used to better classify *de novo* identified elements (6). From the *RepeatModeler* consensus sequences, redundancy was removed by counting as one the repeats that hit the same top hit in the Repbase family through the HOM search in Repclass (6).

For the complete cobra and python genomes, the amount of repetitive sequence was also estimated using P-clouds (23, 24), via a pipeline designed to automate P-clouds analysis, the generation of related statistics, and determine ideal conditions for analysis. Jellyfish (25) counting algorithms were used to count oligonucleotides, and then P-clouds software (24) was used to build the repeat probability clouds and annotate the genome using the C10 parameter set. Dinucleotide simulator software was used to randomly generate a genome lacking repeat elements but with the same frequency of dinucleotides as the original genome (24). This simulated genome was also annotated using P-clouds and the resultant information used to calculate the false positive rate. The P-clouds results were compared to previous RepeatMasker results using BED tools (26) to determine the percent recovery of known elements and estimate the total recovery of all repeat elements (23, 24).

1.8 Gene Family Analyses

Genome-wide gene family analysis **-** Protein sequences of *Anolis carolinensis* were obtained from NCBI protein database and pooled with *Python molurus bivittatus*, and *Ophiophagus hannah* (the King Cobra) MAKER annotated protein sequences. A BLASTP all-vs-all comparison was performed with the WU-BLAST (http://blast.wustl.edu) package v2.0 with the following parameters: T=9 wordmask=seg hitdist=60 matrix=BLOSUM50 Q=13 R=1 E=1e-4, on the combined FASTA file. Cluster analysis was used to condense similar or related protein coding genes to help simplify results. The cluster analysis was based on the algorithm of Single-linkage clustering (27), and the pairwise Jaccard index (28) was calculated for evaluation of connections between protein coding genes built up by edges defined by BLASTP hits. The Jaccard index for

any 2 genes *i* and *j* in the dataset is the number of BLASTP hits shared between *i* and *j*, divided by the union of all BLASTP hits including *i* and *j*. A threshold value for the Jaccard index was then chosen. Next, any *i-j* edge with a value less than the threshold was broken, and the remaining single-linkage clusters were collected. For these analyses, a Jaccard threshold of 0.6 was used, as this value recovered the 2 existing curated gene families in *Anolis carolinensis*-cytochrome c oxidase subunit II and NADH dehydrogenase subunit 2 (http://www.ncbi.nlm.nih.gov/proteinclusters/?term=anolis%20carolinensis%20) in terms of the highest weighted Jaccard index.

Olfactory receptor gene family analyses **-** To compare the olfactory receptor (OR) gene model predictions from cobra and python against green anole, we collected green anole transcriptomes constructed from testes [UniGene: Lib.23344], dewlap skin [UniGene: Lib.23339], regenerated tail [UniGene: Lib.23343], embryo [UniGene: Lib.23340], brain [UniGene: Lib.23338], ovary [UniGene: Lib.23342], and mixed tissues (kidney, lungs, tongue, liver, and heart) [UniGene: Lib.23341].

Queries were conducted on 3,863 intact OR amino acid sequences from *Homo* (29), *Mus* (30), *Pelodiscus* (31), *Xenopus*, *Gallus*, *Anolis*, and *Danio* (32) against the green anole transcripts, and cobra and python gene predictions using TBLASTN. The best BLAST hit from each predicted OR transcript was extracted and the resulting OR translated amino acid sequences aligned from green anole, cobra, and python with human non-OR GPCRs using LINSI (33). A neighbor-joining tree was constructed from the resulting alignment with FastTree (34). Annotations were derived from (32) classifications. Tests of nodal support were conducted in FastTree using the S-H test option (35).

Other receptor gene family analyses – In addition to analyses of olfactory receptor genes, the Vomeronasal Receptor (V1R and V2R) and Ephrin-like Receptor gene families were also analyzed. For both of these families, all genes in the python and cobra genomes that were annotated as being members of these families were extracted. All annotated genes were also extracted for both of these families from the *Anolis* lizard and from the human (for Ephrin-like receptors only). Sequences from each receptor gene family were used for analysis only if they

were specifically identified as being within the gene family of interest. The exception to this was in the search for vomeronasal genes in *Anolis*. There were 11 hits in ENSEMBl in relation to vomeronasal genes, but no annotations were explicitly annotated as being "similar to vomeronasal": they were instead all described as "novel genes". Due to this discrepancy, all 11 hits were used. Nucleotides were aligned based on translated amino acid, and converted back to nucleotides after alignment. Phylogenetic trees were estimated based on amino acid alignments using neighbor joining in FastTree2. Support values of the three OR groups (alpha, beta, and gamma) were estimated using the Shimodaira-Hasegawa test as implemented in FastTree2.

Opsin gene family analyses – Visual and non-visual opsin coding sequences from *Python molurus bivittatus* and *Ophiophagus hannah* (King Cobra) genomes were obtained from BLAST searches (blastn, discontinuous megablast, megablast, tblastn, and tblastx) of the genomes, annotated gene CDS libraries, and de novo-assembled cDNA transcriptome libraries using *Anolis* and *Gallus* mRNA, exon, and protein queries. CDS and BLAST results were manually edited to ensure proper exon boundaries and corrected with direct parsing of the genomes, as necessary. The identities of the opsins were confirmed by phylogenetic analysis as follows. A complete set of vertebrate opsin sequences was obtained from GenBank and ENSEMBL for representative species across major groups (Bony Fish–*Danio*, *Oreochromis*; Amphibians–*Xenopus*, *Cynops*; Reptiles–*Anolis*, *Gallus*; Mammals–*Ornithorhynchus*, *Monodelphis*, *Bos*, *Homo*). Sequences from additional taxa (*Takifugu, Bufo, Uta, Gekko, Xenopeltis, Python regius, Mus*) were used to supplement low sampling or unavailable sequences from the chosen representative species. Four human non-opsin GPCRs (ADRA1DA, NPY1R, P2RY14A, and SST) were used as outgroups. Sequences were aligned with the results from the python and cobra genomes for each major gene class individually using codon alignment in MEGA5 (36). Sequences were manually adjusted and trimmed to exclude areas of poor alignment (generally the ends of the sequences and lineage-specific insertions). Trimmed sequences from each gene alignment were combined and aligned preserving established gaps. The complete alignment was analyzed phylogenetically by maximum likelihood (ML) using PhyML 3 (37) under the GTR+G model with

a BioNJ starting tree, the best of NNI and SPR tree improvement, and aLRT SH-like branch support (38) rooted using the four human non-opsin GPCRs. Instances where we inferred the absence of opsin genes from the snake genomes were verified by conducting multiple types of blast searches (Megablast, tBlastx, blastn) against the entire cobra and python genome assemblies, annotated gene CDSs, and de novo-assembled cDNA sets.

1.9 Transcriptomic Analyses of Gene Expression

Generation of RNAseq data - Total RNA was extracted using Trizol Reagent (Invitrogen), following the manufacturer's protocol. Illumina mRNAseq barcoded libraries were constructed with the Illumina TruSeq RNAseq kit and protocol. Total RNA and mRNA was quality-checked using a BioAnalyzer RNA 6000 pico chip (Agilent). Completed libraries were quantified and checked for appropriate size distribution using the DNA 7500 nono chip on a BioAnalyzer (Agilent).

Analysis of gene expression – Data generated for heart, liver, kidney, and small intestine samples across time points before and after feeding were analyzed to quantify and compare gene expression. Raw Illumina RNAseq reads were trimmed based on quality scores (limit = 0.05, maximum of 2 ambiguities). RNAseq reads from all samples (individual snake organs at a particular timepoint) were mapped to the python annotated transcript set from the gene model predictions from the MAKER annotation pipeline, which incorporated all available RNAseq data for the Burmese Python (Supplementary Table S3) for digital gene expression analysis using the CLC Genomics Workbench. In instances where replicates of a sample were available, replicated samples were combined and mapped together per individual. Mapping parameters were as follows: maximum number of mismatches = 2, minimum length fraction = 0.5, minimum similarity fraction = 0.8, maximum number of hits for a read = 10. Expression was determined by counting the number of unique gene reads that mapped only to a particular annotated transcript. To allow for meaningful statistical inference, the raw expression data (counts) were normalized using the quantile normalization method in the CLC Genomics Workbench. All further statistical analyses of gene expression used the normalized data. For tissues in which

multiple replicates (multiple individuals) per time point were available, a Baggerley's T-test, with FDR-based p-value correction, was used to test for significant differential expression. In the case of the liver, no replicates were available so a Kal's Z-test was used, with FDR-based p-value correction.

Tissue expression cross-referencing – an R-script was used to filter and count the number of genes that were significantly differentially expressed in individual tissues or in all tissues in multi-tissue comparisons. For a given tissue comparison and a particular pairwise timepoint comparison, genes were excluded if they were not significantly differentially expressed (FDR <= 0.05) in one or more of the tissues being compared. Therefore, only genes that were significant in all of the tissues being compared were kept. The output values from the significance filtering were further filtered in some cases to report numbers of genes that were both significant and up/downregulated > 2-fold between time point comparisons.

Heatmap generation – For heatmap generation, normalized expression counts were scaled in R using default parameters to improve visualization and plotted using R's heatmap function. Genes were clustered based on Euclidean distance, and in some instances, samples were clustered based on the same measure.

Short time series expression miner (STEM) analyses – Normalized gene expression data were analyzed in the program STEM (39). STEM is designed to search expression data to identify clusters of expression profiles that are statistically overrepresented in the data. Analyses were conducted on the four tissues for the three main time points (fasted, 1DPF, 4DPF). Normalized gene expression counts were averaged per time point (across individual replicates per time point). These counts were log-normalized in STEM, and the default analysis parameters were kept except for a maximum number of profiles = 50, and max unit change =3.

Gene ontology analyses - We used python transcript CDS set as queries in BLASTx searches against the NCBI non-redundant protein database (E-value 1e-10-5) and gene ontology (GO) annotations were identified with the Blast2GO bioinformatics suite by utilizing the homology search feature based on this BLASTx output. We created a custom reference GO database that comprised 16,858 python CDS sequences associated with at least one GO annotation. Custom

GO enrichment analysis was done using the Singular Enrichment Analysis tool (40); http://bioinfo.cau.edu.cn/agriGO/analysis.php). Enrichment was determined using Fisher's exact test with false discovery controlled by the Yekutieli FDR statistical method. Enriched GO terms were called using a 0.05 significance threshold and a minimum of 5 mapping entries/GO terms.

1.10 Protein Coding Gene Molecular Evolutionary Analyses

Ortholog predictions and annotations - Transcript trio sets were first assembled for python, cobra, and the most closely-related previously sequenced species, *Anolis carolinensis*. Coding regions for all MAKER-annotated transcripts were extracted from both the cobra and python genomes, while coding regions for all annotated transcripts in EnsEMBL Compara v70 were extracted for *Anolis*. *Anolis* and cobra transcripts were queried against python using LAST v274 (41). Hits were filtered so that the best cobra and *Anolis* transcripts were retained for each python coding-region. All transcript trio sets were aligned using PRANK (42) under an empirical model of codon substitution. PRANK-aligned gene sets were then filtered by several criteria. First, any cobra or *Anolis* transcript that matched multiple python transcripts were filtered so that only the transcript set containing the highest scoring alignment was retained. Lineage-specific synonymous substitution rates were estimated for each filtered alignment using a free-ratio model of codon substitution by maximum likelihood. Empirical cutoffs were used to eliminate alignments in which dS > 0.5 for python or cobra, and > 1.5 for *Anolis*. These liberal thresholds were chosen to filter out extreme cases where the alignment was poor or where paralogs or different transcript isoforms may have been incorrectly grouped together.

For all transcript sets passing these criteria, we extracted the coding regions for all 1:1 vertebrate orthologs from a local installation of EnsEMBL Compara v70 using the *Anolis* gene as an index. The resulting transcript sets were then realigned using PRANK and were subjected to additional quality control steps. Although initially all vertebrate 1:1 orthologs were included, the inclusion of fishes substantially reduced the apparent quality of the alignments. Fishes were therefore removed and the resulting alignments were subjected to a final quality control

pipeline. Any sequence that had >20% gaps was removed entirely from the alignment, except for python, *Anolis*, and cobra. In addition, when an alignment column contained gaps in 8 or more species, the entire codon column was excluded. For downstream analyses that threw out gapped columns, this step had no effect. For downstream analyses that integrated over gaps as missing data, this step limited the computational burden of missing data imputation. The tree relating the 1:1 orthologs for all species that passed quality control filtering was output for each transcript set separately by pruning the full species tree. A total of ~7,400 ortholog sets passed these criteria and had data for at least 10 species. 87% of these alignments had 30 or more tetrapod species (including python, cobra, and *Anolis*). On average alignments had 35 species, and some had as many as 47.

Extraction of four-fold sites and other codon positions - For each alignment, custom Perl scripts were used to extract synonymous codon positions by identifying alignment columns that were 4-fold degenerate (or gapped) in all available species. All such 4-fold degenerate sites were later concatenated to create a super matrix across all species. An additional concatenated alignment was made for only those genes with 4-fold sites present from a core set of ten species: human, opossum, mouse, finch, chicken, turtle, lizard, frog, python, cobra. An additional, more conservative "core set" alignment was produced that included only those genes where <33% of the 4-fold positions had gaps (we refer to this dataset as *Ensembl10_4-fold*). Non-gapped codon columns were extracted to produce a dataset we refer to as Ensembl10_all, and this was randomly sampled to produce a manageable subset having 10,000 codon positions (*Ensembl10_10k*). These datasets were later used to estimate rates of molecular evolution (see Section 1.13 below).

Analyses of positive selection - Each PRANK-aligned transcript set that passed all above quality control steps was analyzed for the presence of positively selected positions along the ancestral snake, python, and cobra lineages. Thus, we focused on the subset of ~7,400 alignments with sequences present from python, cobra, and *Anolis*. Analyses were conducted with the "consensus" species tree from EnsEMBL.

Analyses were conducted using a maximum likelihood program (de Koning, available upon request). Likelihood maximizations were run in replicate when the null and alternative models in the branch-site likelihood ratio test had significantly different likelihoods (discussed below). In these cases, the null hypothesis model was rerun using the alternative hypothesis MLEs as a starting point (subject to the parameter constraints of the null model). Starting the null hypothesis likelihood maximization at the MLEs under the alternative hypothesis is a conservative strategy because it insures that the parameter space 'closest' to the alternative hypothesis MLEs (in the KLD sense) will be explored during maximization, and thus that if local optima are present that there will be an increased chance of converging on the most conservative optimum with respect to the LRT. Poor convergence during likelihood maximization is therefore expected to not have been a substantial source of false positives in our analyses.

Additional steps were taken to reduce the impact of potential false positives. The alternative and null hypotheses used in our branch-site analyses were similar to those described previously (43), 'ZNY', but with some modifications devised to improve performance for the distantly-related taxa analyzed here. First, the parameter constraint on the mixing proportions in the ZNY test was relaxed resulting in an additional free parameter. This was done because the data sets analyzed here span deep evolutionary distances and thus are likely informative enough to justify the additional parameter. This choice is also expected to lead to increased power when the assumptions of the ZNY parameter constraint are violated by the data. Second, an additional class of sites (imposing two additional free parameters) was added to the branch-site mixture model that accounts for 'persistent' positive selection along all branches of the phylogeny at a subset of sites. This category of sites was implemented identically in both the alternative and null hypothesis models in order to focus the branch-site LRT more on sites that *uniquely* experienced episodic positive selection on the foreground lineage. Although this modified branch-site test is expected to have increased power and decreased false positives under a proscribed set of conditions, its specification is otherwise identical to the ZNY test and thus is expected to have largely similar properties. As in the ZNY test, the alternative and null hypotheses differ by the addition of a single free parameter in the alternative hypothesis, which

is on its boundary in the null hypothesis. The LRT was therefore performed using a chi-squared mixture distribution that was a 50:50 mixture of d.f.=1 and d.f.=0.

Despite these efforts to improve the specificity and conservativeness of our analyses, it should be noted that recent studies have found that PRANK does an excellent job of minimizing the impact of alignment errors on tests for positive selection (Jordan and Goldman, 2012) and that even standard branch-site tests generally have low false positives (Yang and dos Reis, 2011), even in the presence of various types of errors in model specification (44).

Tests for phenotype and gene ontology category enrichment - For all human genes in the final transcript sets, gene ontology category assignments were extracted from EnsEMBL. In addition, mouse knockout phenotypes were extracted from the Mammalian Phenotype Ontology, and were cross-referenced to the transcript alignments using the mouse 1:1 ortholog IDs. Out of the 7442 transcript alignments, 7214 had GO category assignments, while 3363 had mouse knockout phenotypes. Enrichment of phenotype categories was then assessed using a Fisher's exact test under a variety of stringencies. For each analysis, only the set of genes with relevant phenotype assignments available (GO or MP) were included.

1.11 GC Isochore Structure and Patterns of GC

To examine whether the python genome exhibited regional variation in nucleotide composition (e.g "isochores"), GC content was quantified at several spatial scales. To do this, the GC content standard deviation was calculated for 3-, 5-, 10-, 20-, 80-, 160-, and 320-kb windows. The standard deviation of GC content of a compositionally homogeneous genome will halve as window size quadruples (45). Thus, examining how GC variation declines at different window sizes can quantify the heterogeneity of a genome, e.g. a genome that has large GC content standard deviation of 320-kb windows has significant nucleotide composition heterogeneity at a large spatial scale, indicative of strong isochore structure. Multiple mammal, bird, and reptile genomes were used to compare the compositional structure of genomes among tetrapods and to see how the snakes compare to genomes with strong isochore structure (mammals) and those with no GC-rich isochores (*Anolis* lizard (46)).

The GC contents of third-codon positions (GC3) from alignments of 1:1 orthologous protein-coding genes were used to examine the trajectories of GC content evolution among tetrapods. Orthology was determined using the Ensembl pipeline. For each gene, the program nhPhyml (47) was used to calculate ancestral GC3 as well as equilibrium GC3 (GC3*) under a nonhomogeneous model of molecular evolution (four rate categories and estimated transition/transversion ratio and shape parameter), using the following tree: (((((cobra,python),anolis),(Pelodiscus,((Meleagris,Gallus),Taeniopygia))),((((((Gorilla,(Human,Pan)),Pongo),Macaca),(Mus,Oryctolagus)),(((Felis,(Canis,(Ailuropoda,Mustela))),(Tursiops,Bos)),Pteropus)),Loxodonta)),Xenopus). The divergence in GC3 between nodes, D_{ij}, as branch lengths to visually assess the magnitude of GC3 divergence among vertebrates (Supplementary Figure S23 (48)):

$$D_{ij} = \sqrt{\sum_{k=1}^{n} (GC3_k^i - GC3_k^j)^2}$$

where i and j are ancestor and descendant nodes, and n is the number of genes. GC3* can be considered the GC3 content toward which a lineage is evolving; differences between GC3 and GC3* are indicative of non-equilibrium models of molecular evolution.

1.12 Estimation of Evolutionary Rates Across Amniotes

We used the Ensembl-plus-snake gene set constructed for analysis of positive selection (see section 1.11 above). To make analysis of this dataset computationally tractable, we reduced the taxa represented to: human, mouse, opossum, zebra finch, chicken, softshell turtle, *Anolis* lizard, python and cobra. Additional methods for obtaining gene sets are given above (Section 1.11). The largest dataset (*Ensembl10_all*; see Section 1.11) containing all codon position in the 10-species alignment was analyzed using RaxML (49) with a fixed tree (based on (50)), to estimate branch lengths – this figure is presented in the main text as Fig. 3a.

We conducted additional more detailed analyses of rates of evolution on two subsets of this dataset: 1) *Ensembl10_10k* (see Section 1.11), which contains 10,000 randomly sampled aligned codons, and 2) *Ensembl10_4-fold* (see Section 1.11), which contains a set of over 62,000 four-fold degenerate 3rd codon positions for these ten taxa. Each dataset was run independently in BEAST v1.7.5 (51) with 4 independent trials per dataset, each for 40 million generations, sampling trees and parameters every 1000 generations. We estimated the relative rate of substitution on each branch of the tree using a lognormal relaxed clock model. A birth-death process was assumed for the tree model and a log normal distribution prior was used to describe among-branch substitution rate variation. The tree root height was assumed to follow a normal distribution with mean 324 and SD of 12. This node age represents the split between mammals and reptiles (52-55). For each analysis, the topology was constrained to: frog, (((human,mouse), opossum), (((zebra finch, chicken) softshell turtle),(anolis lizard, (python,cobra)))). Convergence and proper mixing was confirmed across multiple runs of the same analysis by comparison of posterior estimates of likelihood values, and sample sizes >100 for parameter estimates as estimated in Tracer v 1.5 (56). Results of these analyses are shown in Supplementary Figs. S27-28.

In addition to analysis of the Ensembl alignment-based data, we analyzed 4-fold degenerate 3rd codon position sites from existing phylogenetic dataset for >150 squamate reptiles for 44 nuclear encoded genes (57) that is available from the Dryad Data Repository (doi:10.5061/dryad.g1gd8). We inferred times and rates simultaneously under a relaxed clock model using a Bayesian approach with the program Beast 1.7.5 (51).

We constrained the main nodes within squamates and reptiles based on the original publication. We constrained the monophyly of mammals, archosaurs (birds and crocodiles) and squamata. We used the dataset as a concatenated matrix that was assigned a GTRGI model of nucleotide evolution. We assumed a relaxed clock with uncorrelated lognormal distribution and a birth-death model of speciation. The divergence time at the root (split between mammals and reptiles) was assumed to follow a normal distribution with a mean of 324 My and SD=10 (58). We initiated 2 independent runs with random starting trees, and ran each for 50 million generations. Chains were sampled every 1000 generations, and convergence and stationarity

were verified by examining the ESS values for parameter estimates using the program Tracer 1.4. We discarded the first 5 million generations as burn-in period. The posterior probabilities for nodal support were obtained after combining the post burn-in samples from the two independent run. Results are shown in Supplementary Fig. S29.

2. SUPPLEMENTARY TABLES

Supplementary Table S1. Data used for Burmese Python genome assembly.

Library Type	Sequencing platform	Read type	Reads (millions)	Gigabases
Illumina shotgun 300 bp insert	Illumina GAIIx	36 bp paired-end	81.88	2.95
Illumina shotgun 300 bp insert	Illumina GAIIx	76 bp paired-end	74.87	5.69
Illumina shotgun 300 bp insert	Illumina GAIIx	114 bp paired-end	132.82	15.14
Illumina shotgun 500 bp insert	Illumina GAIIx	120 bp paired-end	162.75	19.53
Illumina shotgun 500 bp insert	Illumina GAIIx	150 bp paired-end	137.93	20.69
454 FLX shotgun	Roche 454 FLX	200 cycle	0.119	0.30
454 FLX+ shotgun	Roche 454 FLX+	400 cycle	6.64	3.20
Illumina mate pair 3 kb insert	Illumina HiSeq2000	50 bp paired-end	125	6.25

Supplementary Table S2. Genome assembly statistics for final assembly (Pmo2.0).

Assembly Statistic	Value
Total size (scaffold length, including gaps)	1,435,035,089bp
Scaffold number	39,115
Scaffold N50	207,524 bp
Scaffold N50 number	1,924
Total contig length	1,384,533,364 bp
Contig number	274,247
Contig N50	10,658 bp
Contig N50 number	38,693

Supplementary Table S3. RNAseq libraries used in this study to assemble the python transcriptome for gene annotation.

Tissue	Feeding Status	Number of Individual libraries	Sequencing Platform	Data Source
Heart	Fasted	3	Illumina GAIIx	This study
Heart	Fasted	1	454 FLX	Castoe et al. (2011a)
Heart	Fasted	1	Illumina GAIIx	Wall et al., 2011
Heart	1 DPF	3	Illumina GAIIx	This Study
Heart	1 DPF	1	Illumina GAIIx	Wall et al., 2011
Heart	1DPF	1	Illumina GAIIx	Wall et al., 2011
Heart	4 DPF	3	Illumina GAIIx	This Study
Liver	Fasted	1	Illumina GAIIx	This Study
Liver	Fasted	1	454 FLX	Castoe et al. (2011a)
Liver	1 DPF	1	Illumina GAIIx	This Study
Liver	4 DPF	1	Illumina GAIIx	This Study
Kidney	Fasted	3	Illumina GAIIx	This Study
Kidney	1 DPF	3	Illumina GAIIx	This Study
Kidney	4 DPF	3	Illumina GAIIx	This Study
Small Intestine	Fasted	3	Illumina GAIIx	This Study
Small Intestine	1 DPF	3	Illumina GAIIx	This Study
Small Intestine	4 DPF	3	Illumina GAIIx	This Study
Blood	na	1	Illumina HiSeq 2000	This Study
Ovary	na	1	Illumina HiSeq 2000	This Study
Testes	na	1	Illumina HiSeq 2000	This Study
Stomach	na	1	Illumina HiSeq 2000	This Study
Pancreas	na	1	Illumina HiSeq 2000	This Study
Brain	na	1	Illumina HiSeq 2000	This Study
Rictal gland	na	1	Illumina HiSeq 2000	This Study
Skeletal muscle	na	1	Illumina HiSeq 2000	This Study
Spleen	na	1	Illumina HiSeq 2000	This study

Supplementary Table S4. Summary of gene annotations in the Python genome.

Feature	
Total genes annotated	25,385
Average gene length	18,441 bp
Average exon length	130 bp
Average intron length	1,116 bp

Supplementary Table S5. Details of samples included in analysis of repeat element landscapes across snake species.

Species	Common name	Estimated Haploid Genome Size	Nucleotides sequenced (bp)	Number of reads	Percent of nuclear genome sampled	GC content
Agkistrodon contortrix	Copperhead	1.35 Gbp	60,344,580	280,303	4.50%	42.53%
Python molurus	Burmese python	1.42 Gbp	28,496,896	118,973	2.00%	39.78%
Typhlops reticulatus	Reticulate worm snake	1.92 Gbp	6,741,155	50,087	0.35%	46.13%
Rena dulcis	Texas blind snake	Unknown	11,828,885	71,058	Unknown	43.18%
Anilius scytale	Pipe snake	Unknown	7,542,192	50,319	Unknown	43.52%
Boa constrictor	Boa constrictor	1.71 Gbp	11,575,550	38,037	0.67%	39.83%
Casarea dussermieri	Round Island boa	Unknown	76,243,119	470,682	Unknown	43.43%
Loxocemus bicolor	Mexican python	Unknown	6,172,347	40,583	Unknown	42.87%
Crotalus atrox	Western diamondback rattlesnake	1.71 Gbp	19,098,306	63,094	1.11%	38.77%
Micrurus fulvius	Eastern Coral snake	1.42 Gbp	7,735,311	26,831	0.54%	39.35%
Sibon nebulatus	Snail-eating snake	Unknown	12,772,185	43,542	Unknown	41.01%
Thamnophis sirtalis	Garter snake	1.87 Gbp	49,533,818	176,307	2.64%	42.59%

Supplementary Table S6. Repeat element content for the Burmese python and King Cobra genomes estimated by *RepeatMasker*.

Species	Burmese Python	King Cobra
Total masked	31.82%	35.22%
Total interspersed repeats	27.60%	31.28%
Retroelements	14.37%	16.50%
SINEs	1.60%	2.09%
Squam1/Sauria	0.05%	0.40%
LINEs	8.57%	10.55%
L2/CR1/Rex	4.49%	7.17%
L2	2.36%	2.63%
L3	0.00%	0.01%
R1/LOA/Jockey	0.01%	0.06%
R2/R4/NeSL	0.88%	0.42%
RTE/Bov-B	2.30%	1.05%
L1/CIN4	0.64%	1.41%
PLEs	0.52%	0.54%
LTR elements	0.85%	1.75%
BEL/Pao	0.00%	0.01%
Ty1/Copia	0.01%	0.25%
Gypsy	0.15%	0.69%
DIRS1	0.02%	0.45%
Retroviral	0.09%	0.06%
DNA transposons	3.45%	3.49%
hobo-Activator	0.35%	0.89%
Tc1-IS630-Pogo	1.81%	2.26%
En-Spm	0.00%	0.01%
MuDR-IS905	0.00%	0.00%
PiggyBac	0.09%	0.02%
Other (Mirage,	0.00%	0.00%
Unclassified	12.61%	12.87%
Small RNA	0.31%	0.14%
Satellites	0.04%	0.06%
Simple repeats	1.23%	1.51%
Low complexity	0.82%	1.23%

Supplementary Table S7. Repetitive sequence estimates for cobra and python complete genomes based on P-Clouds. P-clouds analyses were run using the C10 parameter setting.

	Python Genome	**Cobra Genome**
Repeat Masker - Total Repetitive	0.3182	0.3522
Repeat Masker percent recovery	0.675219426	0.798259369
Estimated false negative rate (1-RM recovery)	0.324780574	0.201740631
Estimated False Positive	0.001744	0.003434
Estimated True Positivies (1-FP)	0.998256064	0.996566153
P-clouds estimated bp	580614344	800629541
P-clouds estimated false positive bp	1012554.09	2749239.36
P-clouds Estimate (-FP)	579601789.9	797880301.7
Total genome bp (contig length)	1444020805	1655084175
Final P-clouds estimate (-FP)	0.401380498	0.482078382
Final P-clouds estimate (wFP)	0.402081703	0.48373947
Final P-clouds estimate (RM recovery, -FP)	0.59447137	0.603919615

Supplementary Table S8. Repeat element content for the sample sequenced snake genomes estimated by *RepeatMasker*.

Species	*Rena*	*Typhlops*	*Anilius*	*Loxocemus*	*Casarea*	*Boa*
SINEs	1.74%	0.96%	1.67%	1.41%	2.15%	2.28%
LINEs	15.33%	10.77%	8.19%	9.23%	7.98%	8.41%
PLEs	0.69%	0.30%	1.22%	1.88%	0.41%	0.63%
LTR elements	2.78%	3.00%	1.76%	0.85%	0.94%	1.18%
DNA transposons	3.12%	4.53%	4.33%	2.45%	3.22%	2.11%
Unclassified	19.89%	20.78%	15.73%	17.34%	20.73%	10.88%
Small RNA	0.06%	0.24%	0.19%	0.23%	0.22%	0.30%
Satellites	0.04%	0.01%	0.04%	0.01%	0.03%	0.01%
Simple repeats	0.82%	1.26%	0.83%	0.58%	0.87%	0.91%
Low complexity	0.60%	0.40%	0.31%	0.27%	0.26%	0.70%
Total interspersed elements	43.56%	40.35%	32.90%	33.16%	35.42%	25.49%
Total Repetitive Content	48.60%	44.79%	36.83%	37.11%	38.54%	30.10%

Species	*Python* (sampled)	*Crotalus*	*Agkistrodon*	*Micrurus*	*Sibon*	*Thamnophis*
SINEs	1.38%	1.66%	1.57%	2.37%	3.50%	4.30%
LINEs	7.63%	11.44%	12.36%	11.29%	15.81%	9.78%
PLEs	0.68%	1.02%	1.18%	0.84%	0.69%	0.67%
LTR elements	0.84%	3.58%	4.04%	2.83%	2.93%	2.51%
DNA transposons	2.99%	4.57%	5.54%	3.13%	5.72%	4.42%
Unclassified	11.36%	13.33%	17.71%	16.23%	14.25%	17.58%
Small RNA	0.27%	0.12%	0.11%	0.27%	0.24%	0.29%
Satellites	0.02%	0.03%	0.05%	0.01%	0.08%	0.06%
Simple repeats	0.77%	1.82%	3.77%	1.55%	1.83%	1.89%
Low complexity	0.56%	0.82%	1.20%	1.12%	1.24%	0.64%
Total interspersed elements	24.89%	35.59%	42.40%	36.70%	42.90%	39.26%
Total Repetitive Content	28.77%	41.42%	48.58%	43.65%	51.65%	46.42%

Supplementary Table S9. Results of gene expression analysis for physiological remodeling after feeding in the Burmese Python.

Heart	Fasted vs. 24h	24h vs. 96h	0h vs. 96h	Overall
Total differentially expressed genes (p < 0.05)	731	658	246	1334
Total upregulated genes (p < 0.05)	418	371	141	
Upregulated > 2 fold (p < 0.05)	301	275	89	
Upregulated > 5 fold (p < 0.05)	72	85	27	
Total downregulated genes (p < 0.05)	313	287	105	
Downregulated > 2 fold (p < 0.05)	243	171	74	
Downegulated > 5 fold (p < 0.05)	81	54	24	

Kidney	Fasted vs. 24h	24h vs. 96h	0h vs. 96h	Overall
Total differentially expressed genes (p <0.05)	927	3669	3244	5445
Total upregulated genes (p < 0.05)	520	2032	2026	
Upregulated > 2 fold (p < 0.05)	421	1283	1716	
Upregulated > 5 fold (p < 0.05)	121	513	689	
Total downregulated genes (p < 0.05)	407	1637	1218	
Downregulated > 2 fold (p < 0.05)	292	1049	998	
Downegulated > 5 fold (p < 0.05)	83	525	472	

Liver	Fasted vs. 24h	24h vs. 96h	0h vs. 96h	Overall
Total differentially expressed genes (p <0.05)	7011	5367	7178	10093
Total upregulated genes (p < 0.05)	3593	3048	3617	
Upregulated > 2 fold (p < 0.05)	2754	1672	2843	
Upregulated > 5 fold (p < 0.05)	1388	807	1486	
Total downregulated genes (p < 0.05)	3418	2319	3561	
Downregulated > 2 fold (p < 0.05)	2783	1339	2891	
Downegulated > 5 fold (p < 0.05)	1490	700	1570	

Small Intestine	Fasted vs. 24h	24h vs. 96h	0h vs. 96h	Overall
Total differentially expressed genes (p <0.05)	1220	638	1189	2351
Total upregulated genes (p < 0.05)	616	346	622	
Upregulated > 2 fold (p < 0.05)	539	285	531	
Upregulated > 5 fold (p < 0.05)	203	97	192	
Total downregulated genes (p < 0.05)	604	292	567	
Downregulated > 2 fold (p < 0.05)	527	230	489	
Downegulated > 5 fold (p < 0.05)	159	65	140	

Footnote: For heart, kidney, and small intestine, p-values are based on the FDR-corrected p-value from Baggerley's T-test. For the liver, because there are no replicates per time point, p-values are based on FDR-corrected values for Kai's Z-test. Fold changes for all samples are calculated using the weighted proportions fold change measure.

Supplementary Table S10. Numbers of significantly differentially expressed genes shared between tissues.

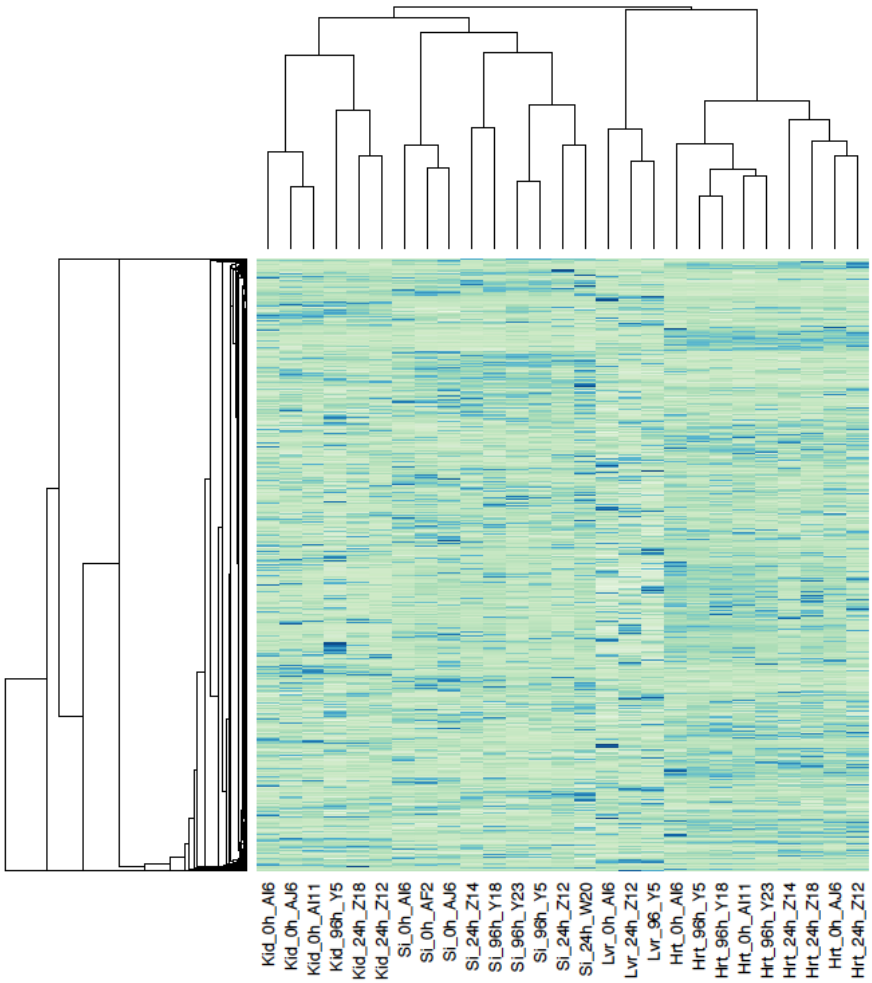
Tissue Comparison	0h vs. 24h	24h vs. 96h	0h vs. 96h
All Tissues	20	13	4
Heart-Kidney-Liver	59	105	35
Heart-Kidney	67	191	110
Heart-Liver	413	275	69
Heart-Small intestine	70	34	17
Kidney-Liver	578	1338	1632
Kidney-Small intestine	179	191	356
Liver-Small intestine	712	261	682
Heart-Kidney-Small intestine	21	18	6
Heart-Liver-Small intestine	52	19	52
Kidney-Liver-Small intestine	143	86	227

	0h versus 24h (>2-fold change)			24h versus 96h (>2-fold change)		
Tissue Comparison	All	upregulated	downregulated	All	upregulated	downregulated
All Tissues	8	3	1	3	2	0
Heart-Kidney-Liver	28	13	6	28	8	6
Heart-Kidney	39	22	8	83	34	20
Heart-Liver	235	86	95	110	46	29
Heart-Small intestine	37	11	18	18	11	5
Kidney-Liver	357	190	122	710	233	198
Kidney-Small intestine	134	73	55	124	56	46
Liver-Small intestine	479	223	187	145	68	41
Heart-Kidney-Small intestine	9	5	1	6	5	1
Heart-Liver-Small intestine	23	4	11	7	4	1
Kidney-Liver-Small intestine	88	45	38	49	21	12

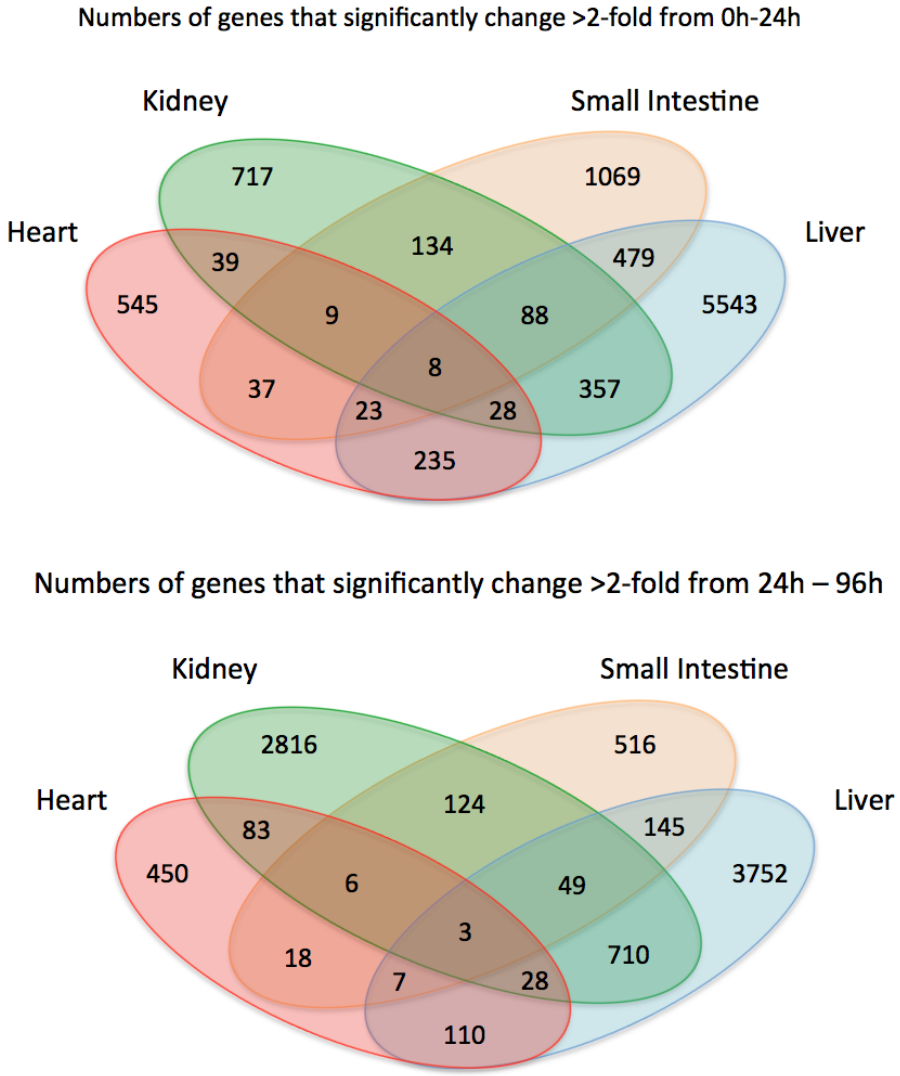
Supplementary Table S11. Complete list of vertebrate opsin genes. Genes present in snakes are bolded, while those lost from snakes (but otherwise present in squamates) are greyed.

Code	Gene Name	Synonymy	Lineage-specific Duplicates
LWS	Long-wavelength Sensitive Opsin	OPN1LW, red sensitive opsin, MWS, OPN1MW	LWS1 and LWS2: Zebrafish MWS: Primates
RH1	Rhodopsin	RHO, RHO1	RH1-1,-2: Zerbrafish
exoRHO	Exorhodopsin	n/a	Bony Fish specific
RH2	Medium-wavelength Sensitive Opsin	RHO2, OPN1MW, green sensitive opsin	RH2-1,-2,-3,-4: Zebrafish RH2a1,a2,B: Tilapia
SWS1	Short-wavelength Sensitive Opsin 1	OPN1SW1, violet/UV sensitive opsin	n/a
SWS2	Short-wavelength Sensitive Opsin 2	OPN1SW2, blue sensitive opsin	SWS2a,b: Tilapia
PIN	Pinopsin	n/a	n/a
PPIN	Parapinsopin	n/a	PPINa,b: Bony Fish
PARIE	Parietopsin	n/a	n/a
VAOP	Vertebrate Ancient-long Opsin	Val Opsin	VAOPa,b: Zerbrafish
ENC	Encephalopsin	OPN3	n/a
MEL1	Melanopsin 1	OPN4m, OPN4a, OPN4-2 Melanopsin mammalian-like	MEL1a,b,c: Bony Fish
MEL2	Melanopsin 2	OPN4x, OPN4b, OPN4-1 Melanopsin Xenopus-like	MEL2a,b: Bony Fish
NEUR1	Neuropsin 1	OPN5, Neuropsin	n/a
NEUR2	Neuropsin 2	OPN5L1	n/a
NEUR3	Neuropsin 3	OPN5L2	NEUR3a,b: Bony Fish
NEUR4	Neuropsin 4	n/a	n/a
NEUR5	Neuropsin 5	n/a	n/a
TMT2	Teleost Multiple Tissue Opsin 2	n/a	TMT2a,b: Bony Fish
TMT3	Teleost Multiple Tissue Opsin 3	n/a	Lost in amniotes
TMTa	Teleost Multiple Tissue Opsin a	n/a	TMTa1,2: Bony Fish
RRH	Retinal pigment epithelium-derived rhodopsin homolog	Peropsin	n/a
RGR	Retinal G Protein Coupled Receptor	n/a	RGRa,b: Bony Fish

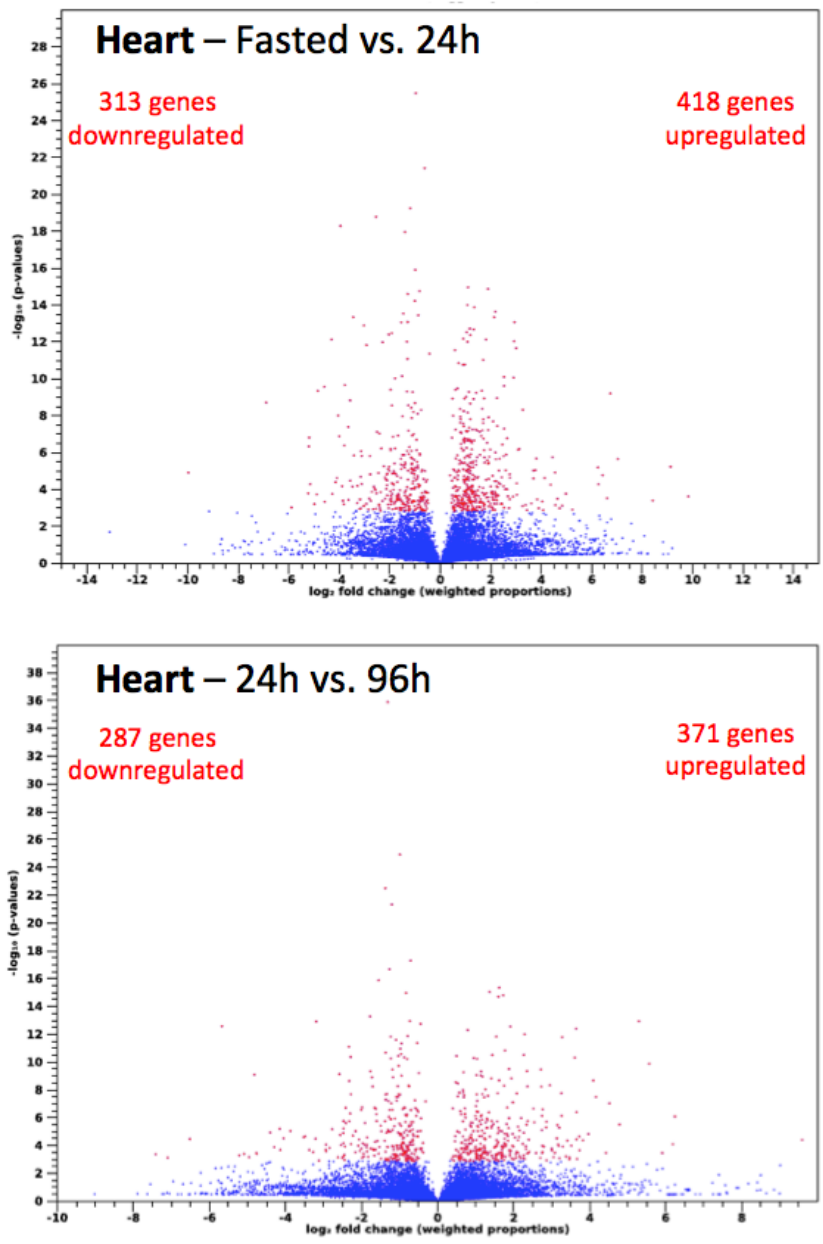
SUPPLEMENTARY FIGURES



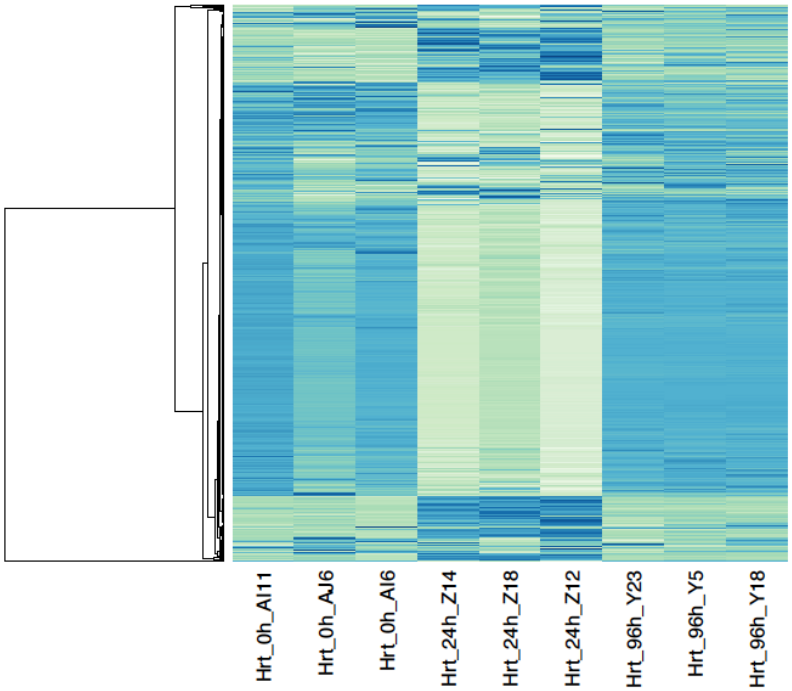
Supplementary Figure S1. Cluster analysis of all transcriptomic samples. Each individual sample for each organ sample is shown, and gene expression levels are indicated for all genes that significantly change in at least one organ between time points. Samples are indicated by abbreviations for organs (kidney, small intestine, liver, and heart) for the three time points samples (0 hour, 24 hour, and 96 hour post-feeding). Individual animal identifiers are indicated by the last acronym in each sample name. Relative levels of gene expression are represented by colors, with light green being low expression, and darker blue being high expression.



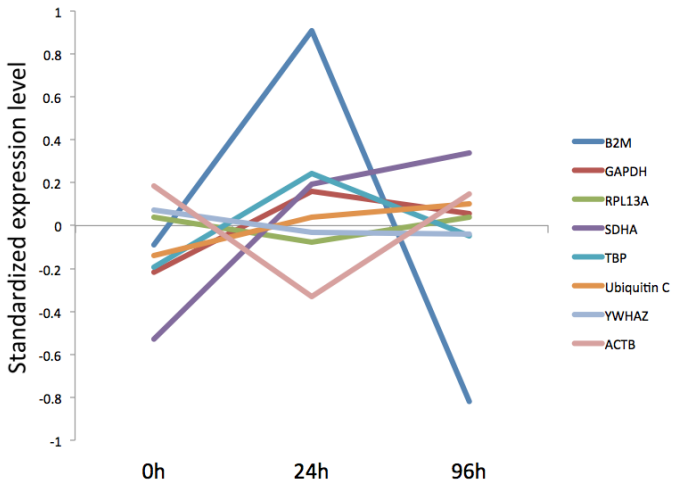
Supplementary Figure S2. Venn diagrams of numbers of genes that significantly change expression level (> 2-fold) between time points across tissues.



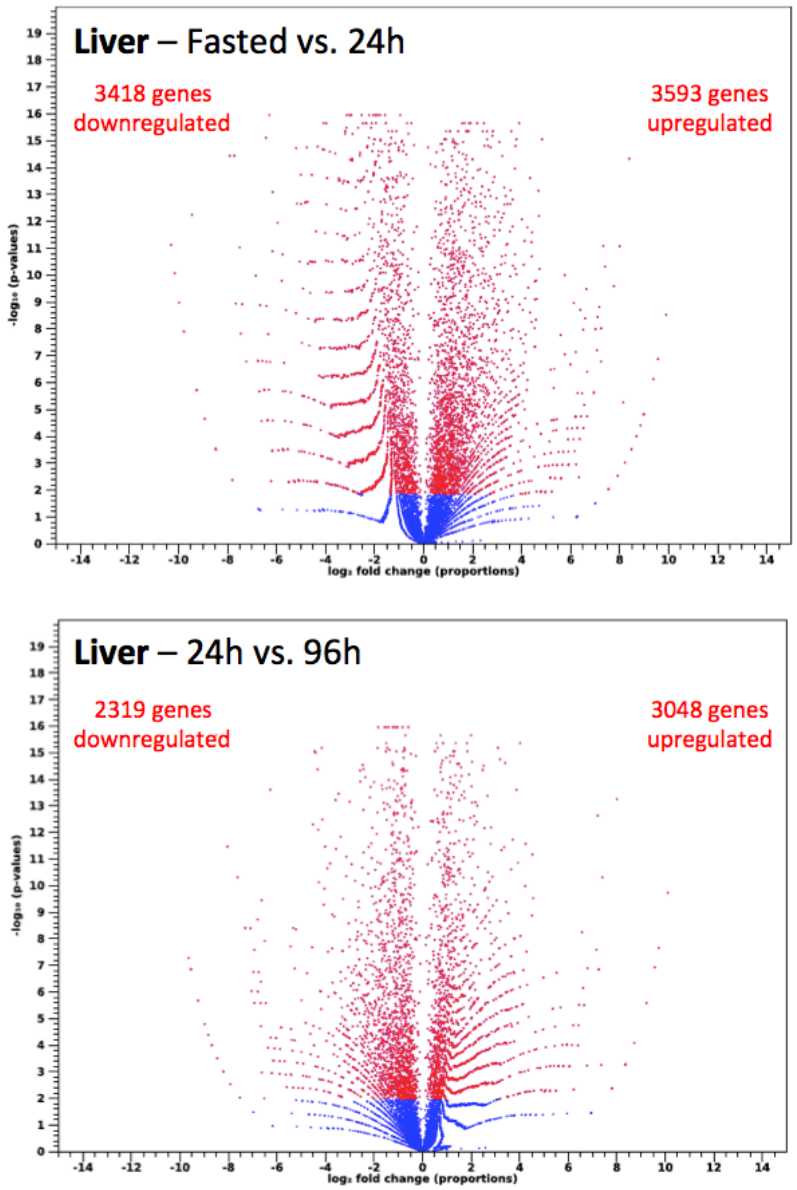
Supplementary Figure S3. Volcano plot comparison of expression changes between timepoints in the heart. Genes that are significantly differentially expressed between time points are shown in red. Significance is based on Baggerleys test statistic ($p < 0.05$), and fold change based on weighted proportions from three replicates per time point.



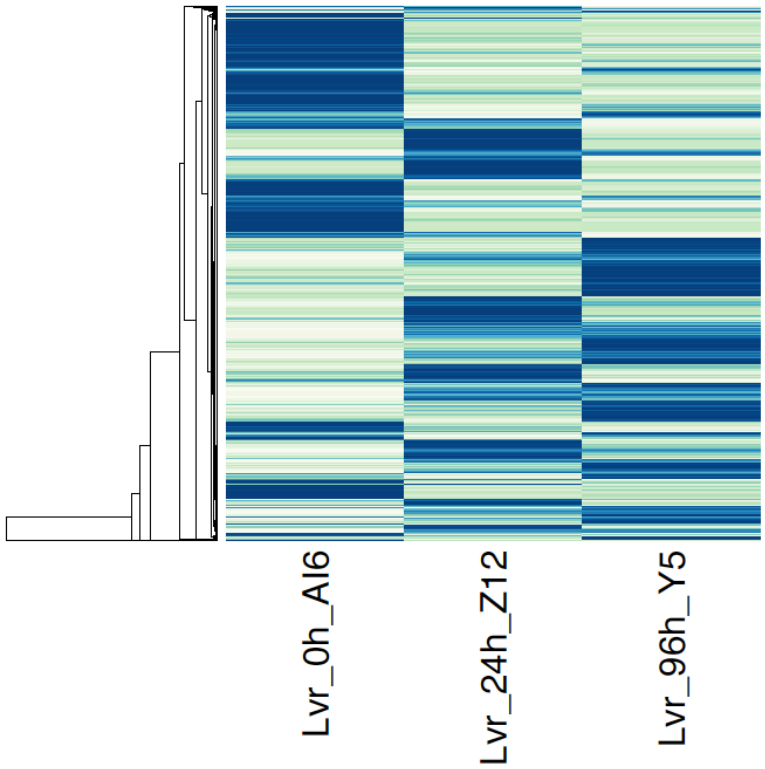
Supplementary Figure S4. Heat map of expression levels for genes that significantly change in the heart. Genes that are significantly differentially expressed between time points are shown, arranged into clusters. Significance is based on Baggerleys test statistic ($p < 0.05$). Relative levels of gene expression are represented by colors, with light green being low expression, and darker blue being high expression.



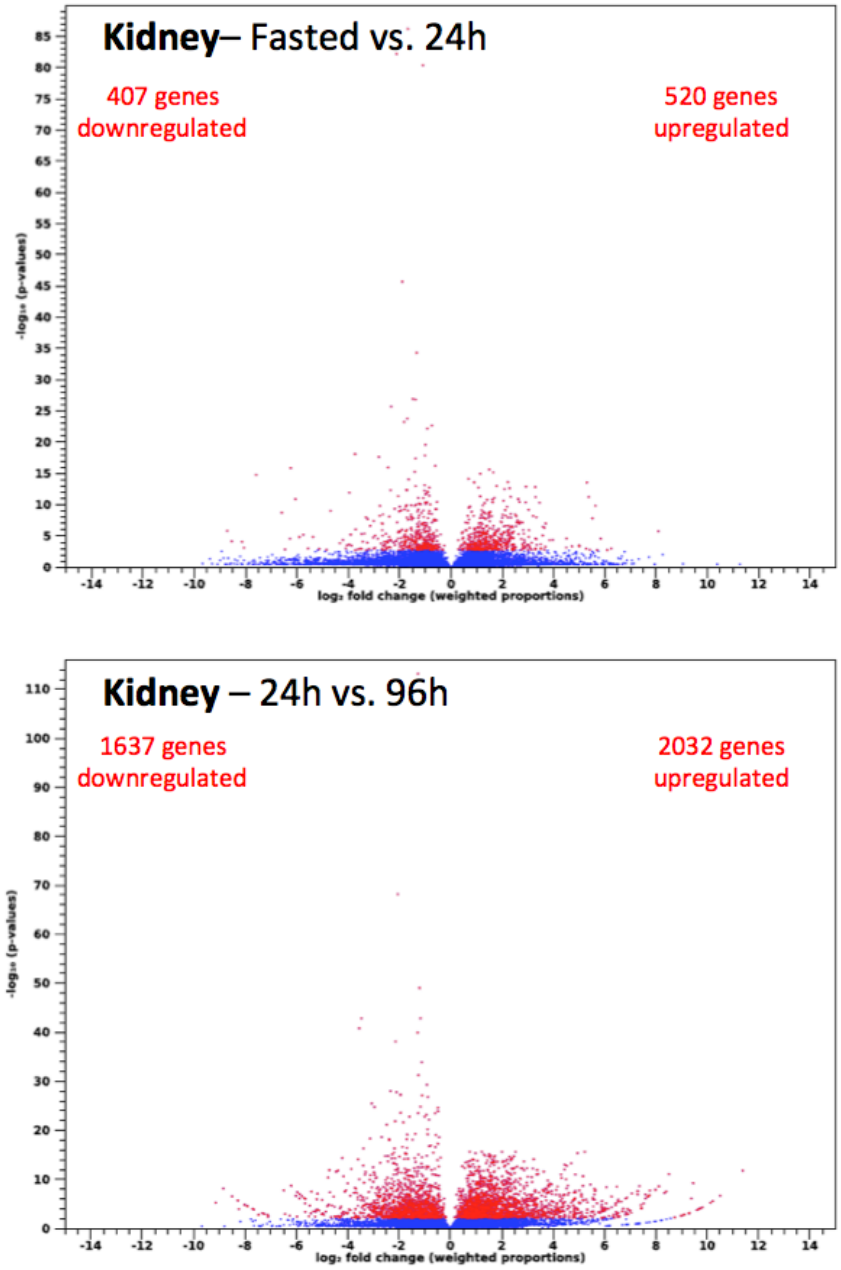
Supplementary Figure S5. Plot of normalized expression counts for housekeeping genes in the heart. Counts are averaged per time point from heart samples, for genes typically considered to be housekeeping genes that should maintain relatively constant expression levels (and are commonly used as loading standards for measuring gene expression). Expression levels were calculated from normalized expression counts averaged across samples per timepoint (=mean[NE]). Values shown are 1-(mean[NE] / mean (mean[NE]) for all time points). This graph indicates that the net expression changes across these genes are approximately zero, suggesting that normalization procedures were effectively at maintaining a flat baseline for transcriptome-wide analysis of gene expression.



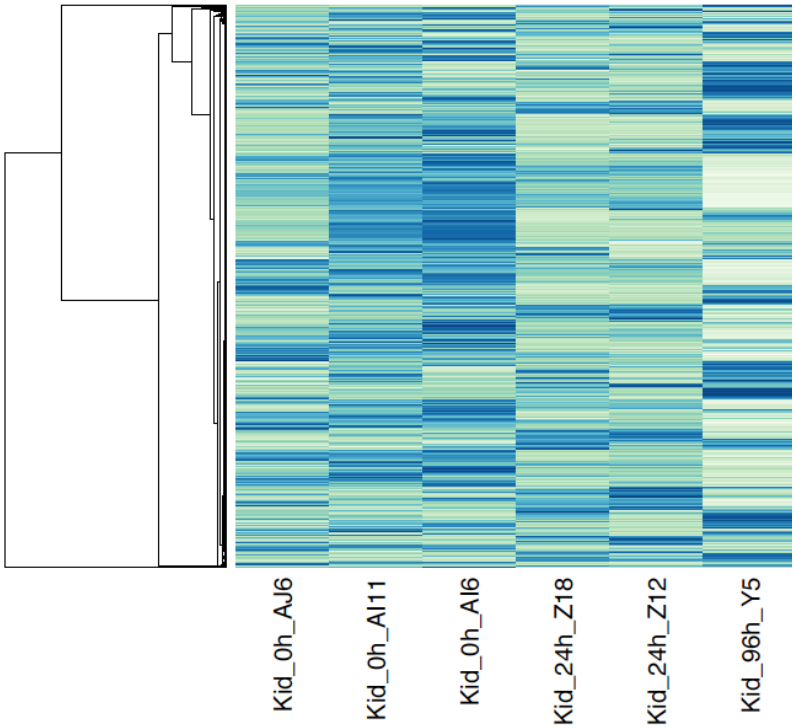
Supplementary Figure S6. Volcano plot comparison of expression changes between timepoints in the liver. Genes that are significantly differentially expressed between time points are shown in red. Significance is based on Kal's Z-test statistic ($p < 0.05$).



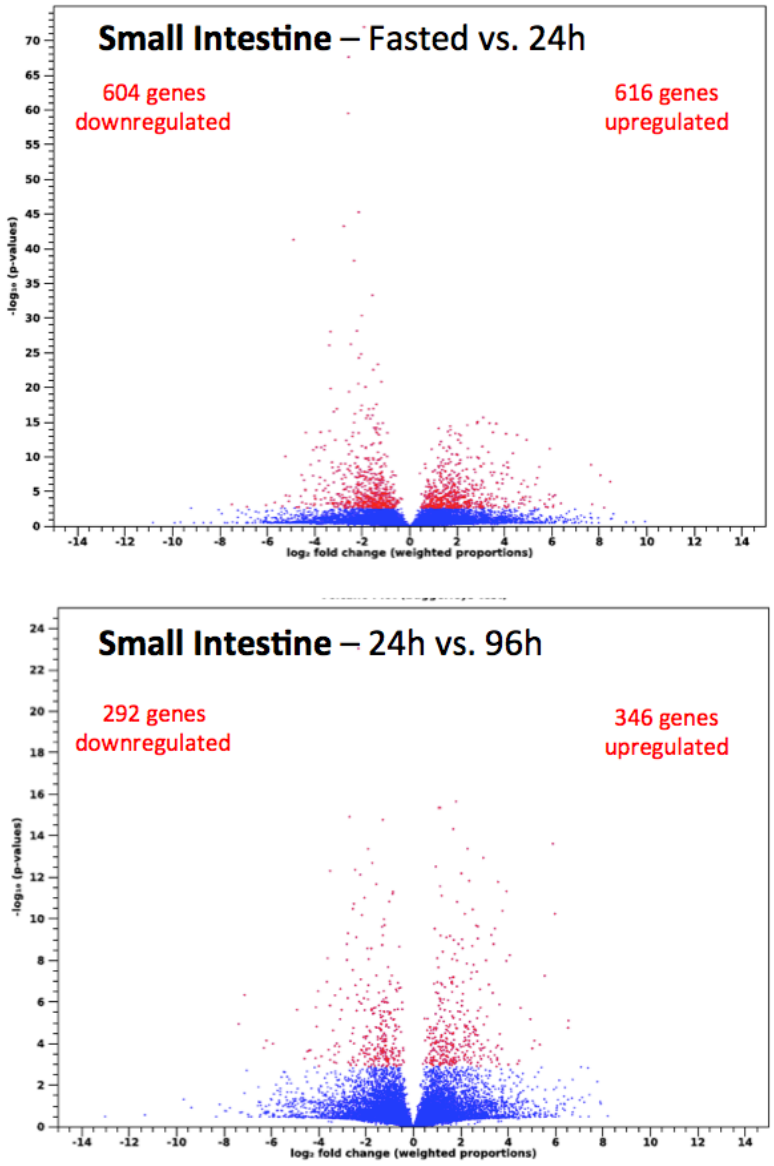
Supplementary Figure S7. Heat map of expression levels for genes that significantly change in the liver. Genes that are significantly differentially expressed between time points are shown, arranged into clusters. Significance is based on Kal's Z-test ($p < 0.05$). Relative levels of gene expression are represented by colors, with light green being low expression, and darker blue being high expression.



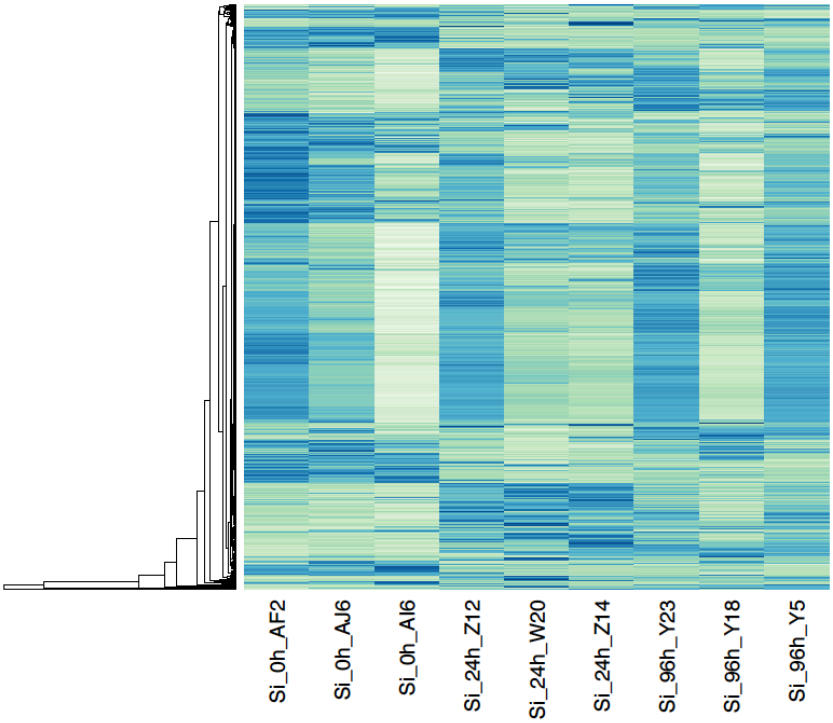
Supplementary Figure S8. Volcano plot comparison of expression changes between timepoints in the kidney. Genes that are significantly differentially expressed between time points are shown in red. Significance is based on Baggerleys test statistic ($p < 0.05$), and fold change based on weighted proportions from three replicates per time point.



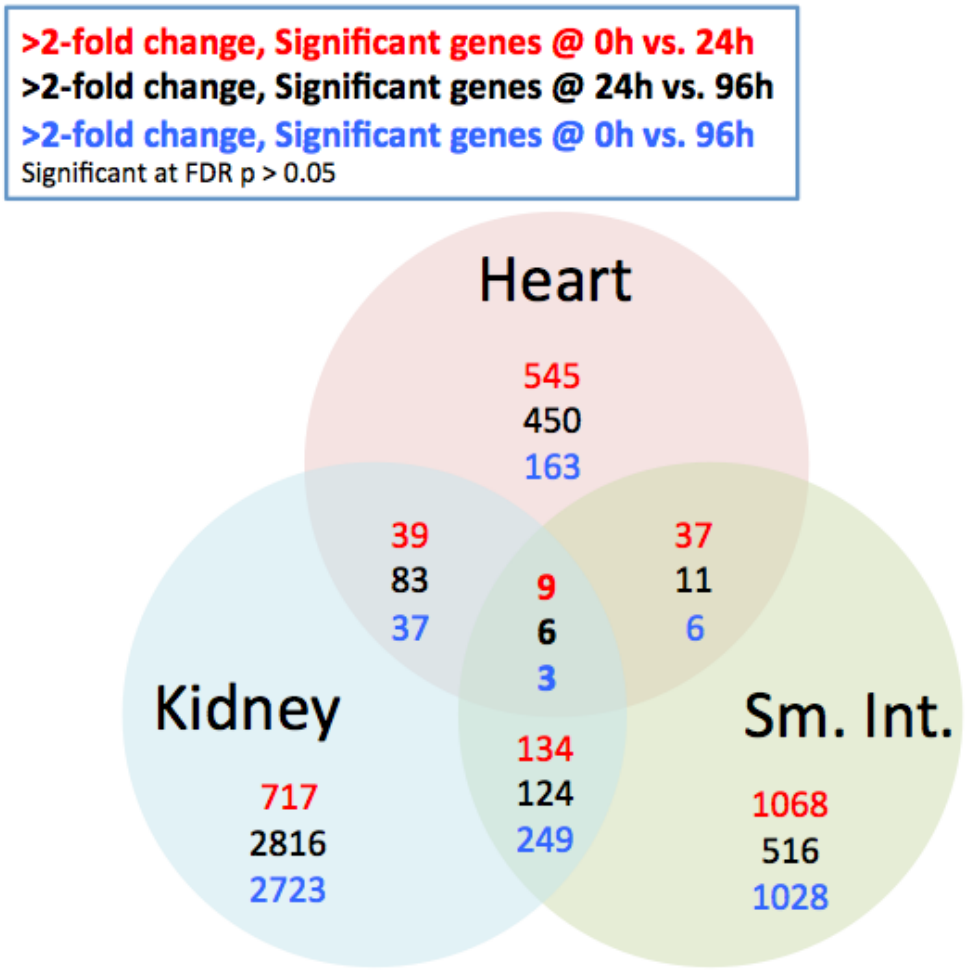
Supplementary Figure S9. Heat map of expression levels for genes that significantly change in the kidney. Genes that are significantly differentially expressed between time points are shown, arranged into clusters. Significance is based on Baggerleys test statistic ($p < 0.05$). Relative levels of gene expression are represented by colors, with light green being low expression, and darker blue being high expression.



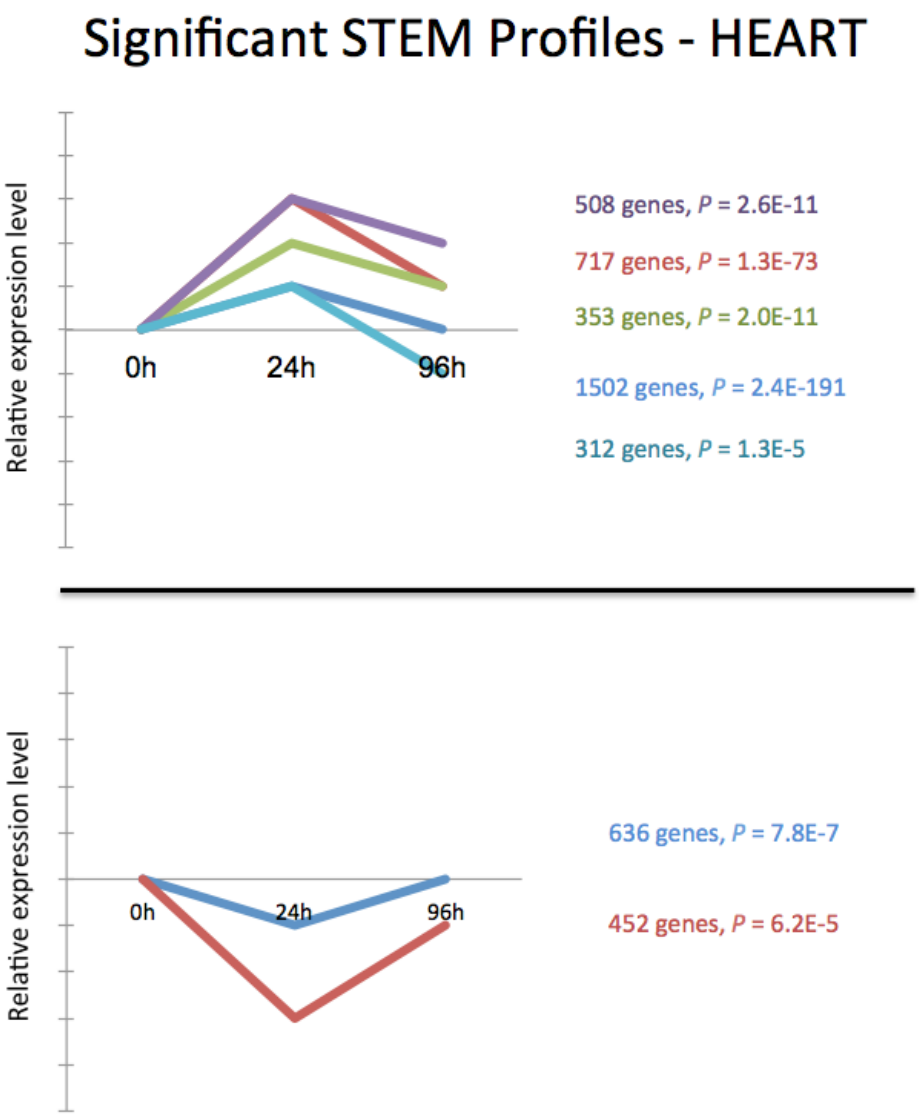
Supplementary Figure S10. Volcano plot comparison of expression changes between timepoints in the small intestine. Genes that are significantly differentially expressed between time points are shown in red. Significance is based on Baggerleys test statistic ($p < 0.05$), and fold change based on weighted proportions from three replicates per time point.



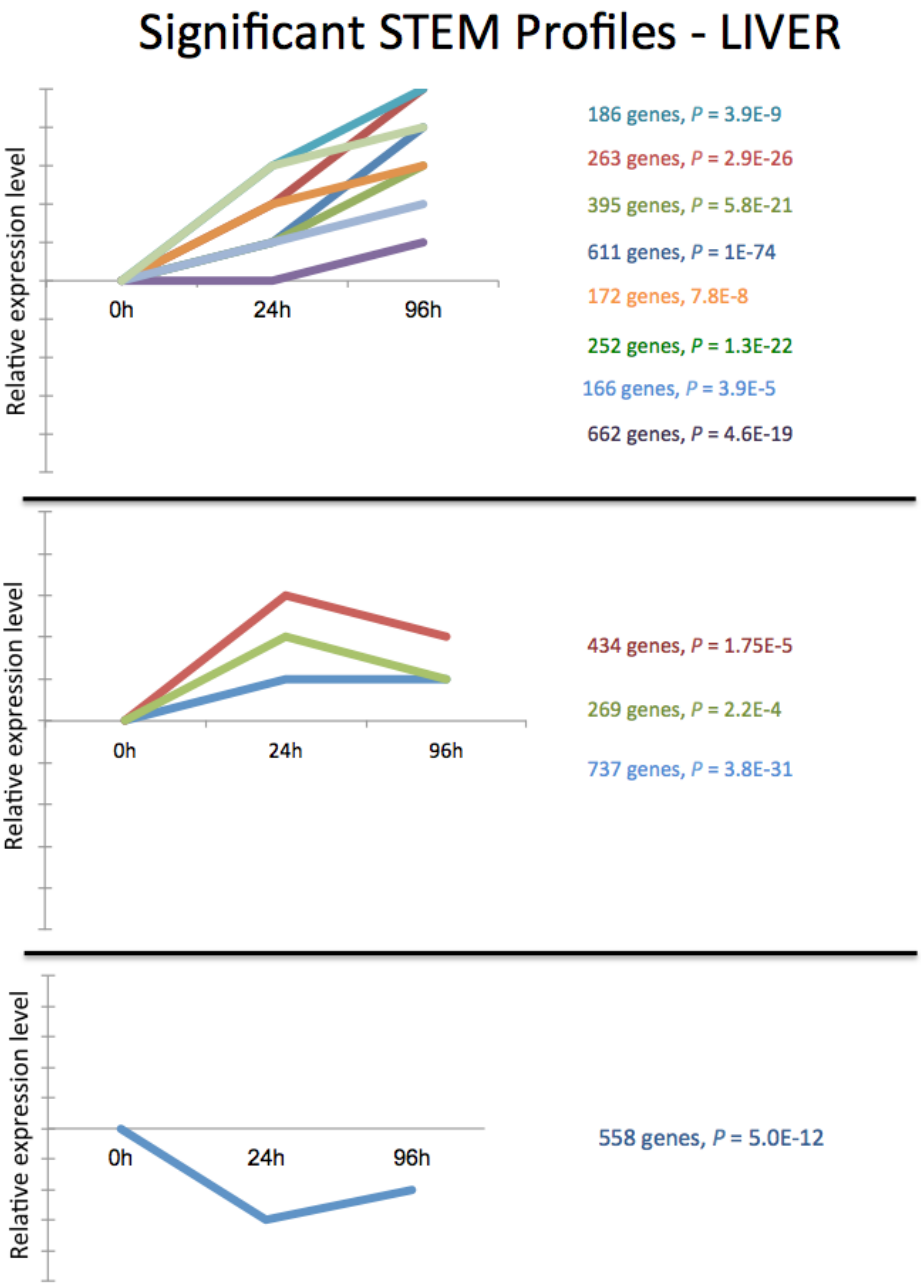
Supplementary Figure S11. Heat map of expression levels for genes that significantly change in the small intestine. Genes that are significantly differentially expressed between time points are shown, arranged into clusters. Significance is based on Baggerleys test statistic ($p < 0.05$). Relative levels of gene expression are represented by colors, with light green being low expression, and darker blue being high expression.



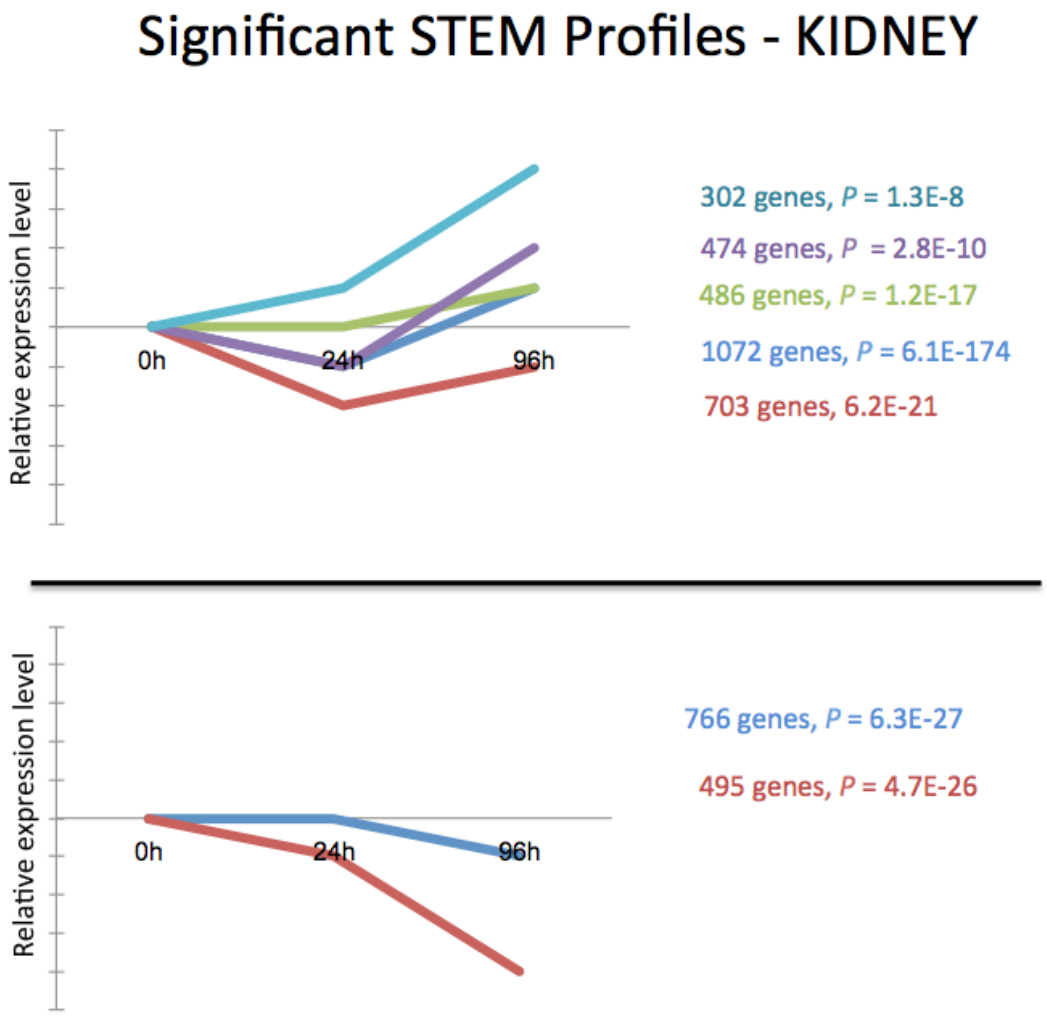
Supplementary Figure S12. Numbers of genes that significantly change in expression levels more than 2-fold in magnitude between 0-24h and 24-96h. The Venn diagram shows the numbers of these genes that are shared across tissues.



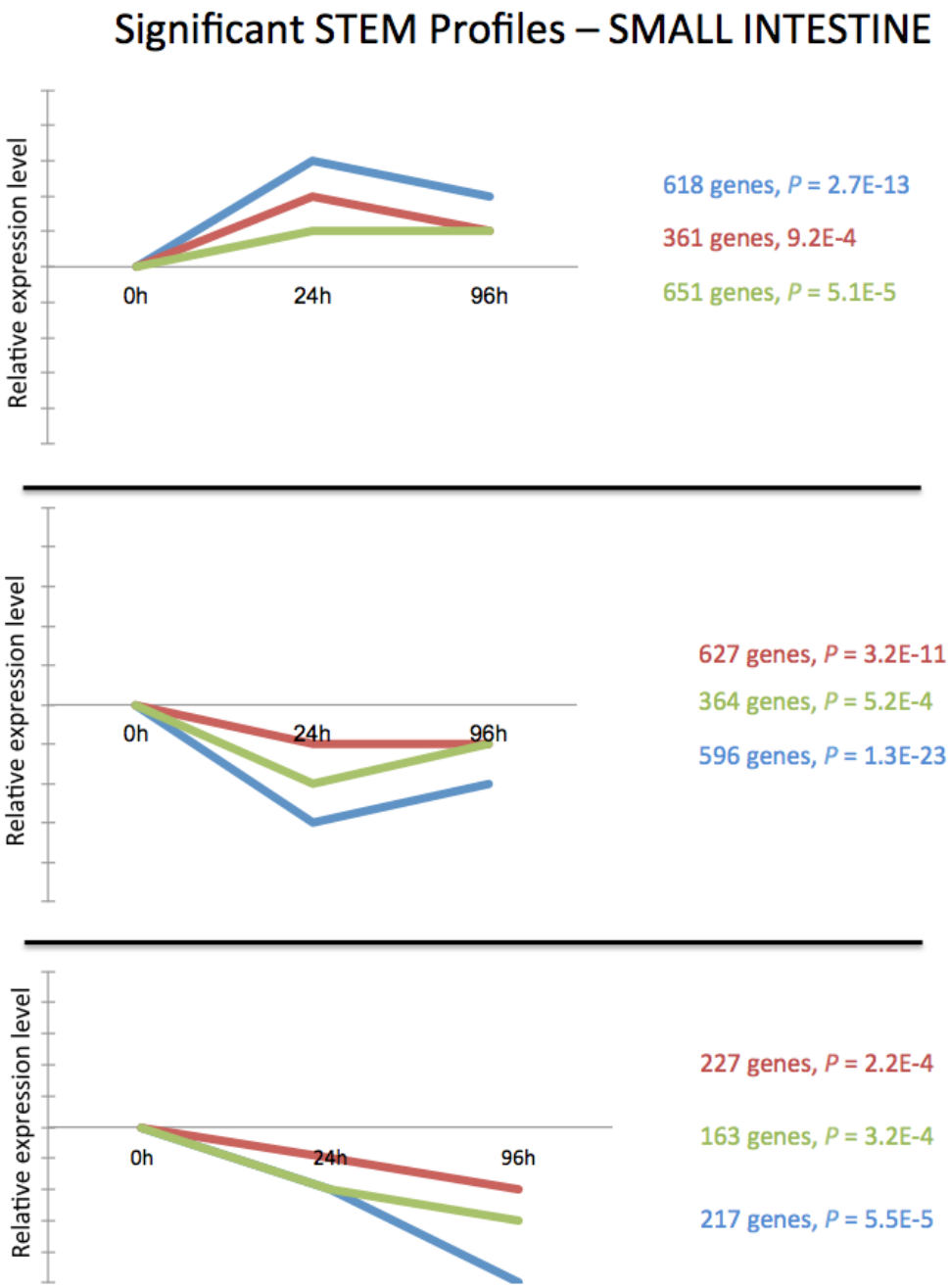
Supplementary Figure S13. Generalized expression profiles significantly over-represented in the heart. Enriched profiles and significance estimated in STEM. Each line represents a generalized expression profile found to be significantly enriched.



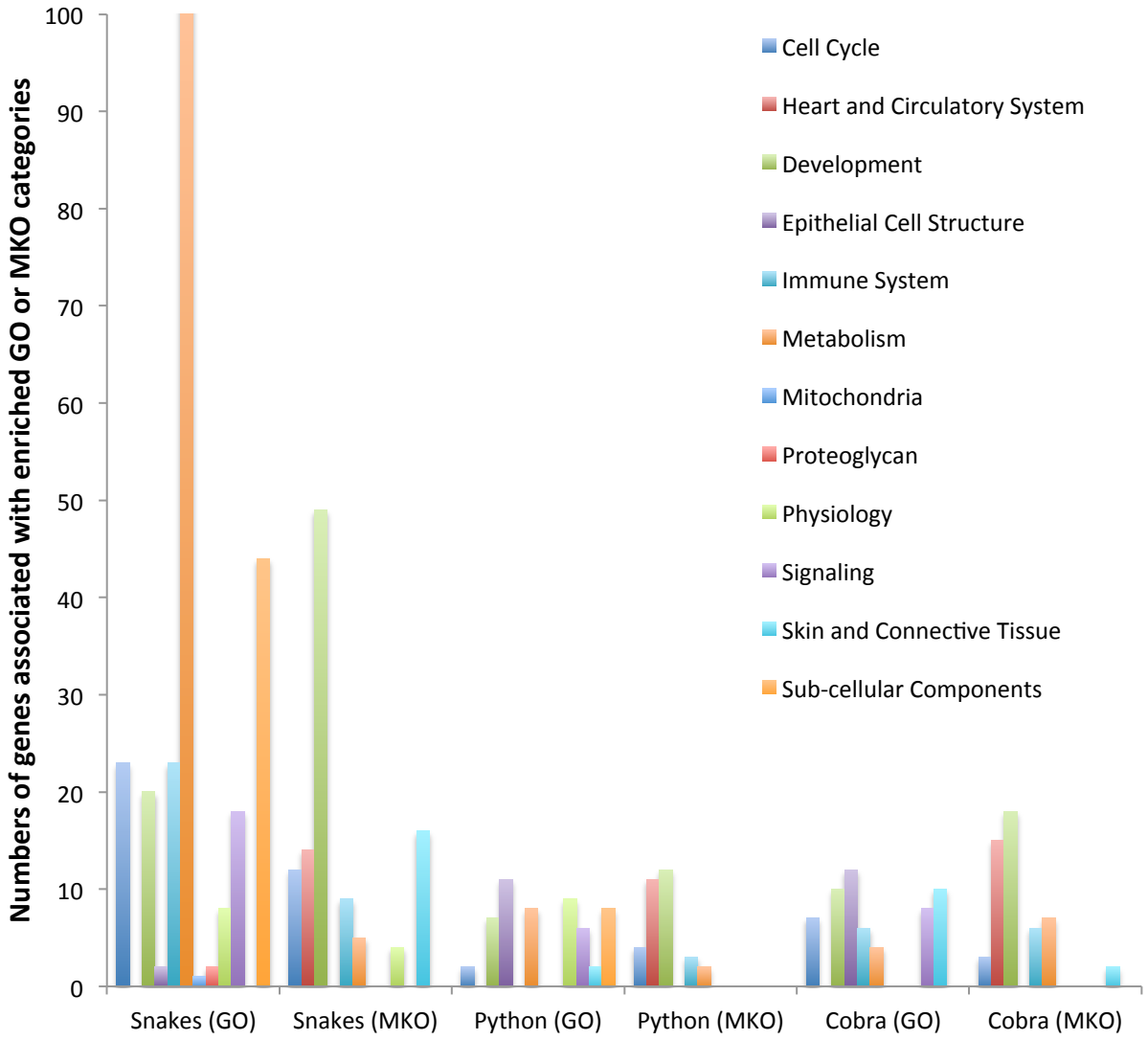
Supplementary Figure S14. Generalized expression profiles significantly over-represented in the liver. Enriched profiles and significance estimated in STEM. Each line represents a generalized expression profile found to be significantly enriched.



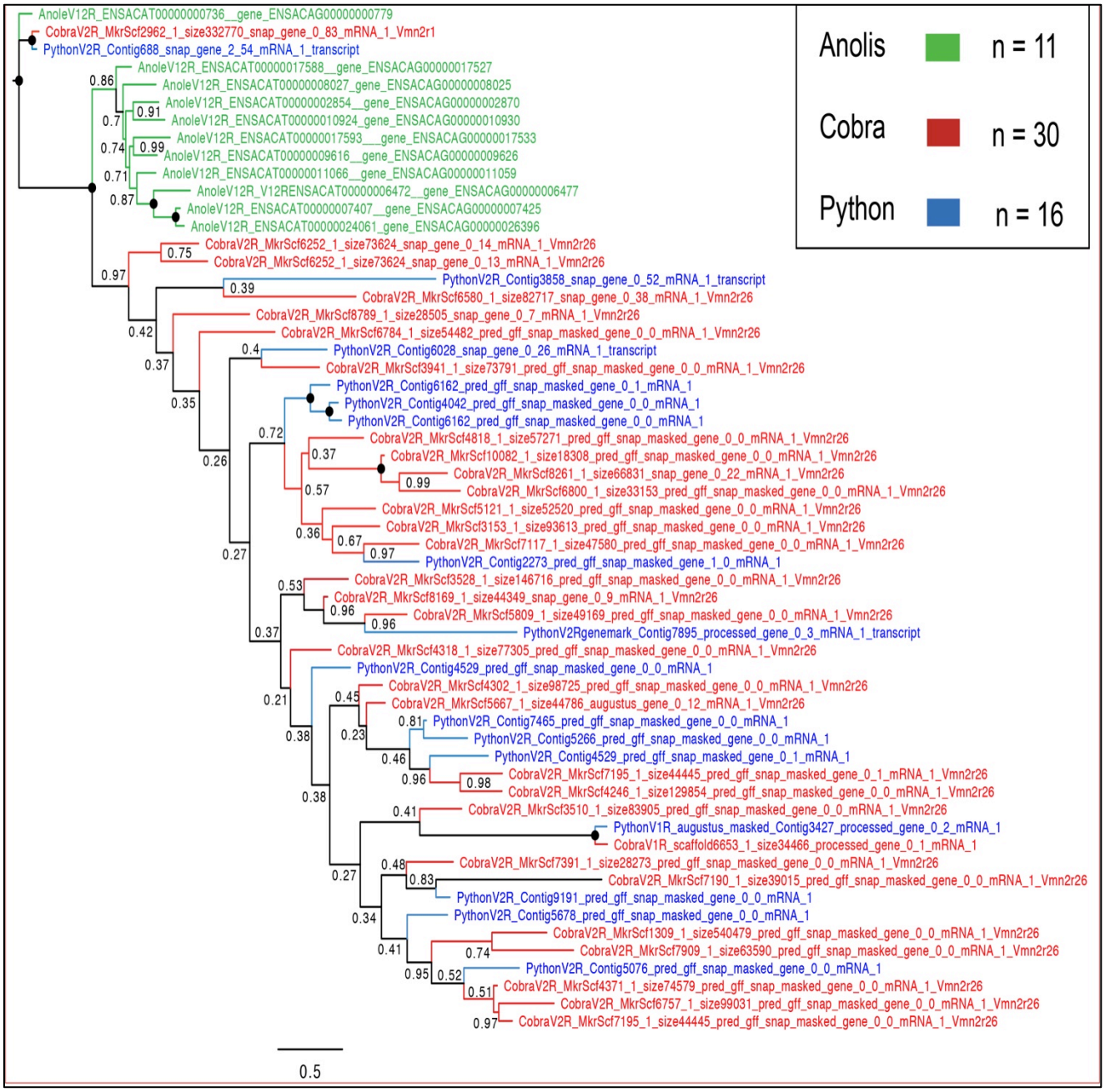
Supplementary Figure S15. Generalized expression profiles significantly over-represented in the kidney. Enriched profiles and significance estimated in STEM. Each line represents a generalized expression profile found to be significantly enriched.



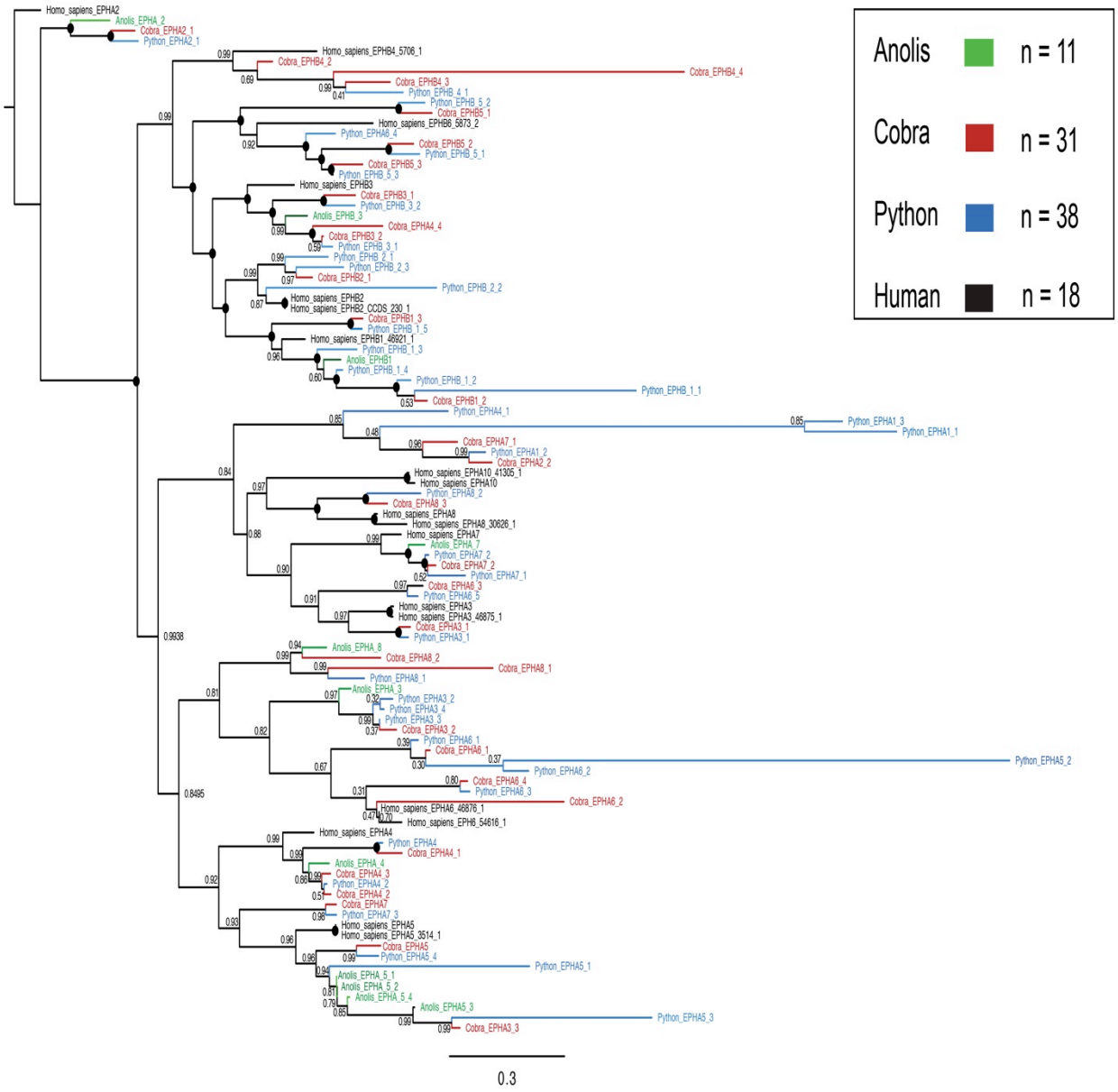
Supplementary Figure S16. Generalized expression profiles significantly over-represented in the small intestine. Enriched profiles and significance estimated in STEM. Each line represents a generalized expression profile found to be significantly enriched.



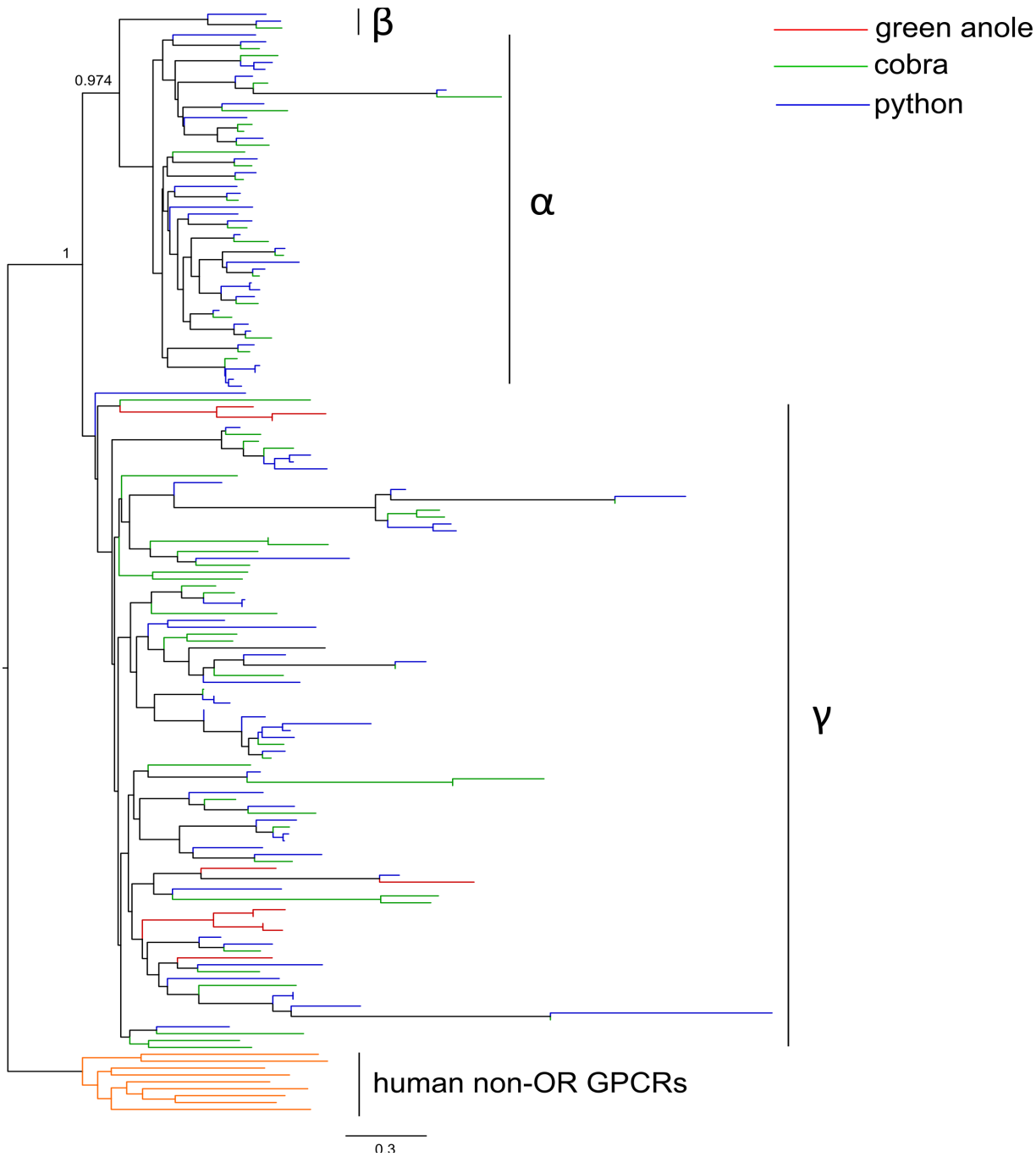
Supplementary Figure S17. Gene ontology terms and mouse knockout phenotypes significantly enriched ($p<0.05$) in positively selected genes ($p<0.001$) in snake lineages. Plot shows the numbers of genes that fall within statistically enriched GO or MKO categories. These categories were clustered by broad biological characteristics into the sets above (X-axis). This process may result in genes being counted multiple times per clustered group when genes have multiple GO terms or MKO phenotypes that are enriched, and these are clustered into a broader category in the graph. Details of GO and MKO term clustering in Dataset 3.



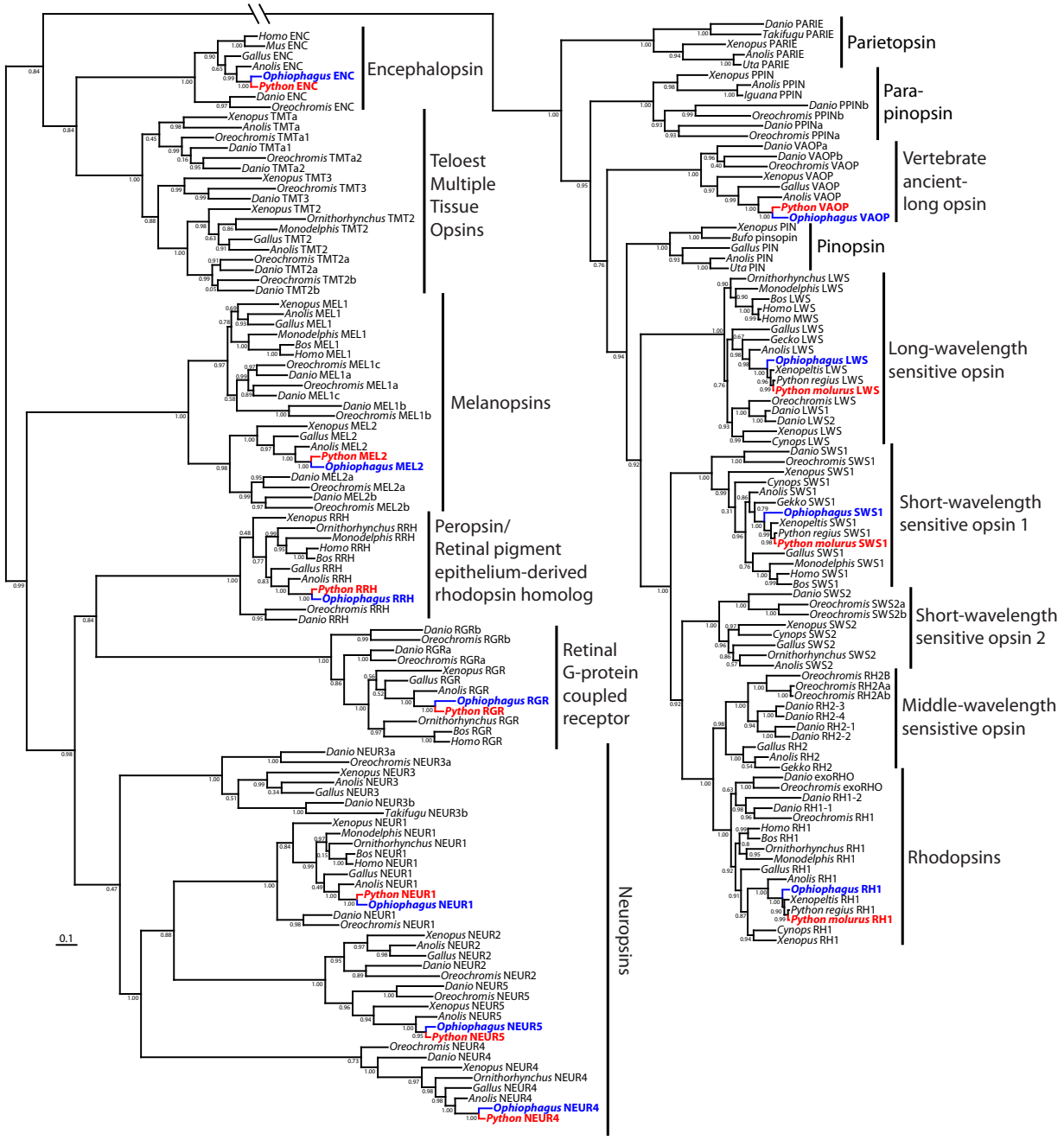
Supplementary Figure S18. Phylogeny of annotated vomeronasal genes in the genomes of the lizard and snakes. Phylogenetic trees estimated using Bayesian inference in MrBayes, with poster probabilities for nodal support indicated by a number or a filled circle if 100%.



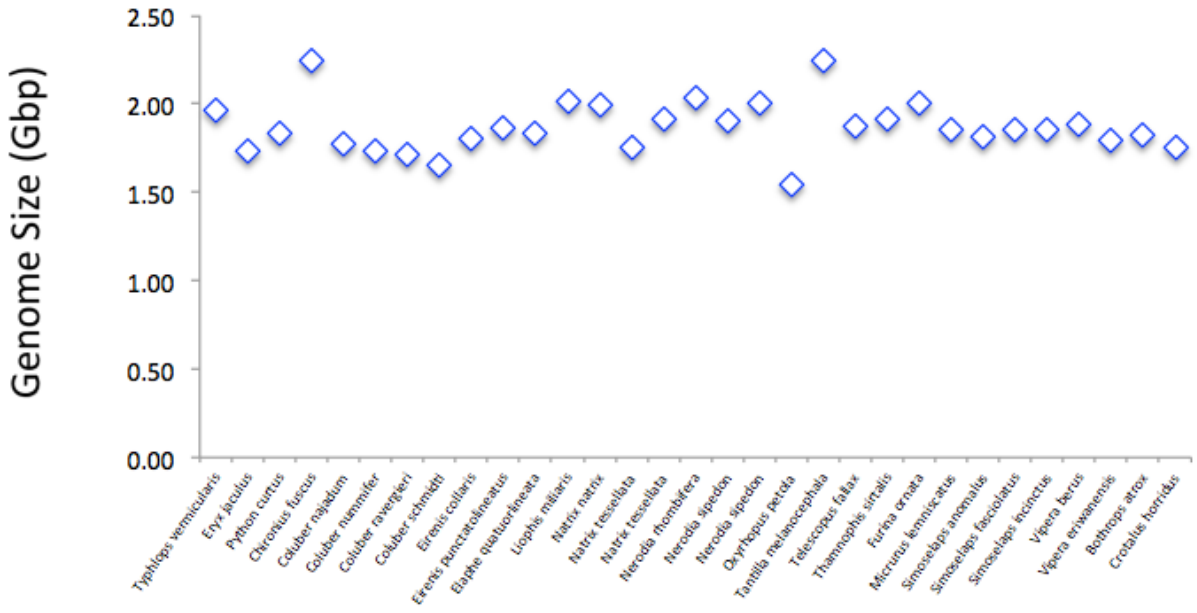
Supplementary Figure S19. **Phylogeny of annotated Ephrin-like receptor genes in the genomes of the lizard and snakes.** Phylogenetic trees estimated using Bayesian inference in MrBayes, with posterior probabilities for nodal support indicated by a number or a filled circle if 100%.



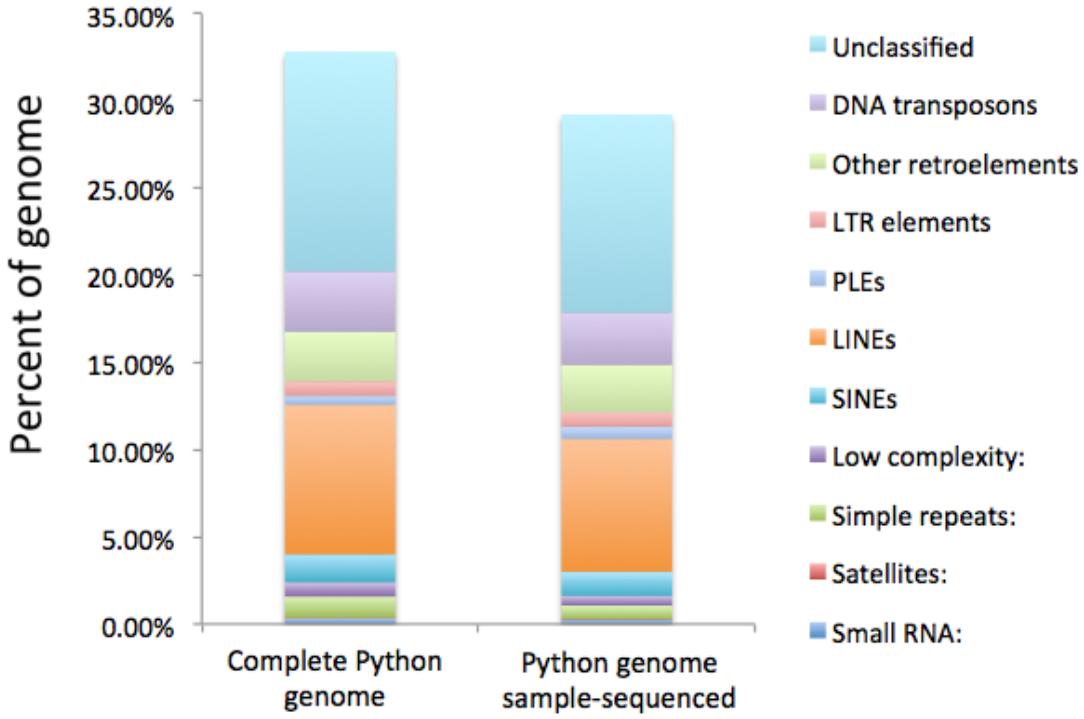
Supplementary Figure S20. **Phylogeny of olfactory repeptor gene families in squamate reptiles and human.** Phylogenetic tree based on analysis of amino acid sequences using FastTree2, with support values for three main clades of receptors (alpha, beta, gamma) based on SH-test implemented in FastTree2.



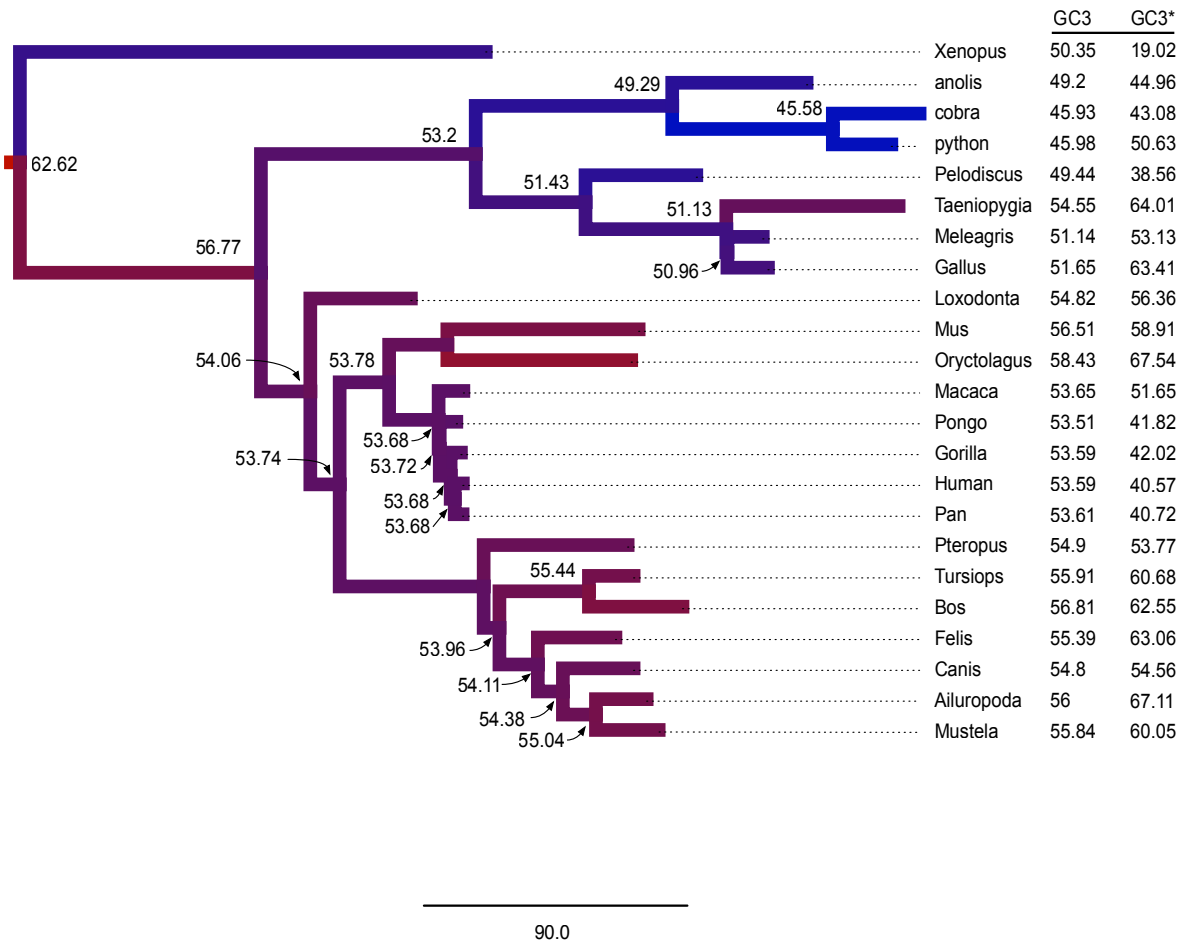
Supplementary Figure S21. **Phylogeny of vertebrate visual and non-visual opsins illustrating the presence and absence of opsin genes in snakes.** Sequences from the python (red) and cobra (blue) genomes are indicated. The phylogenetic tree was estimated by maximum likelihood (PhyML) and rooted with four human non-opsin GPCRs (not shown). Numbers at nodes are aLRT SH-like branch support values.



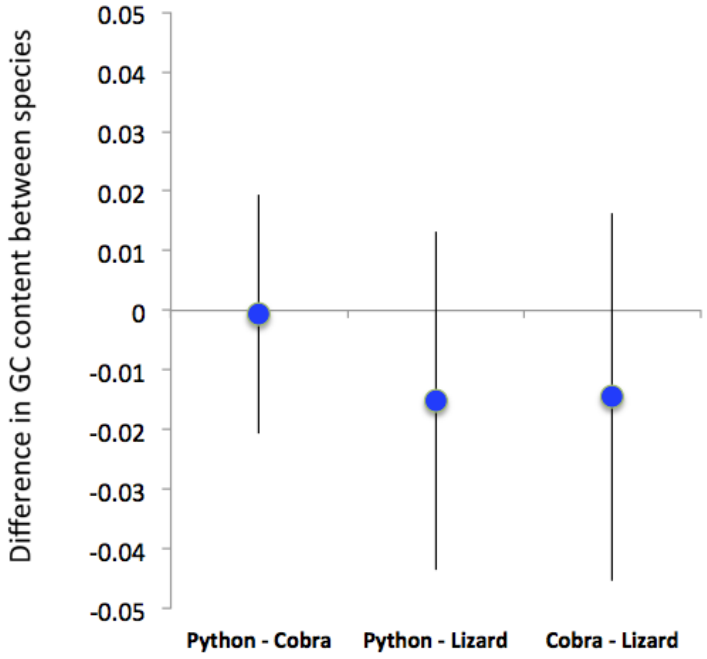
Supplementary Figure S22. Snake genome size estimates based on flow cytometry. Data from Animal Genome Size Database.



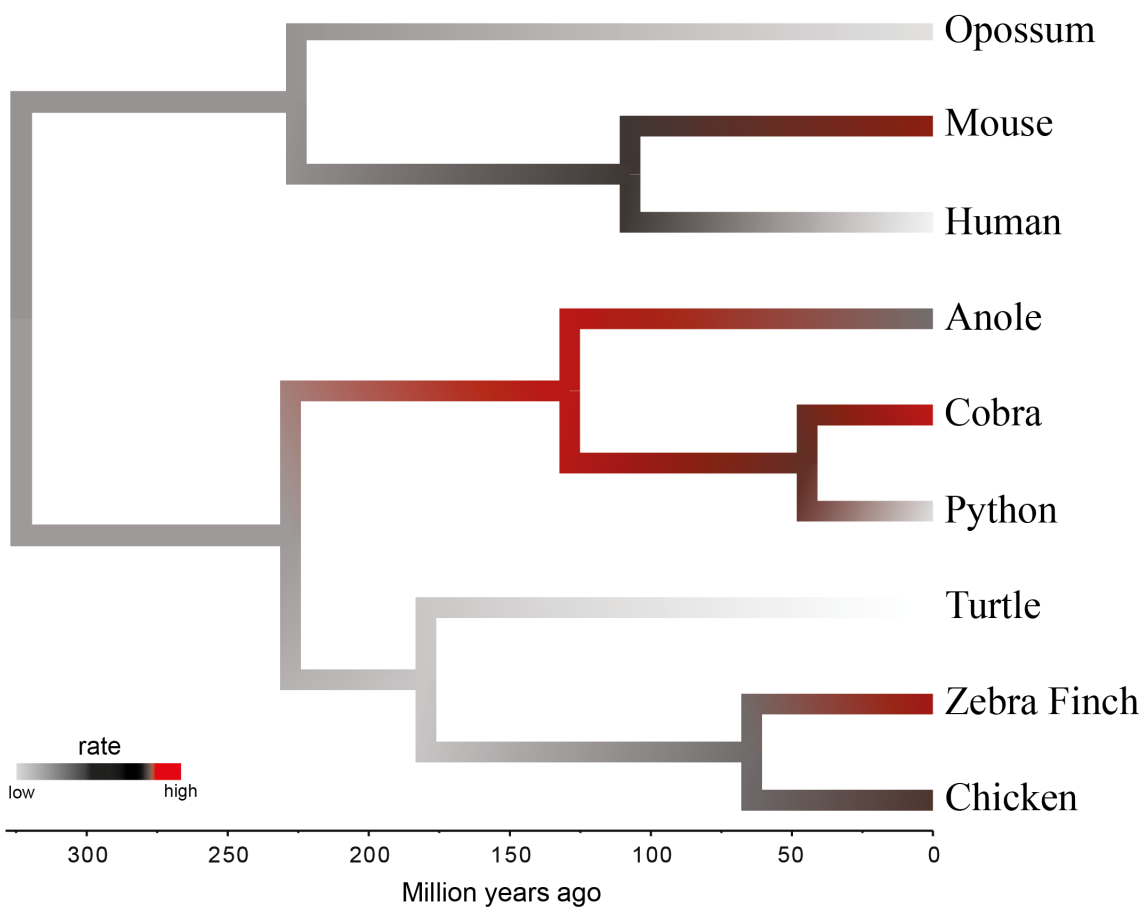
Supplementary Figure S23. Comparison of RepeatMasker estimation of repeat content between the complete python genome and the unassembled sample-sequencing dataset from the python.



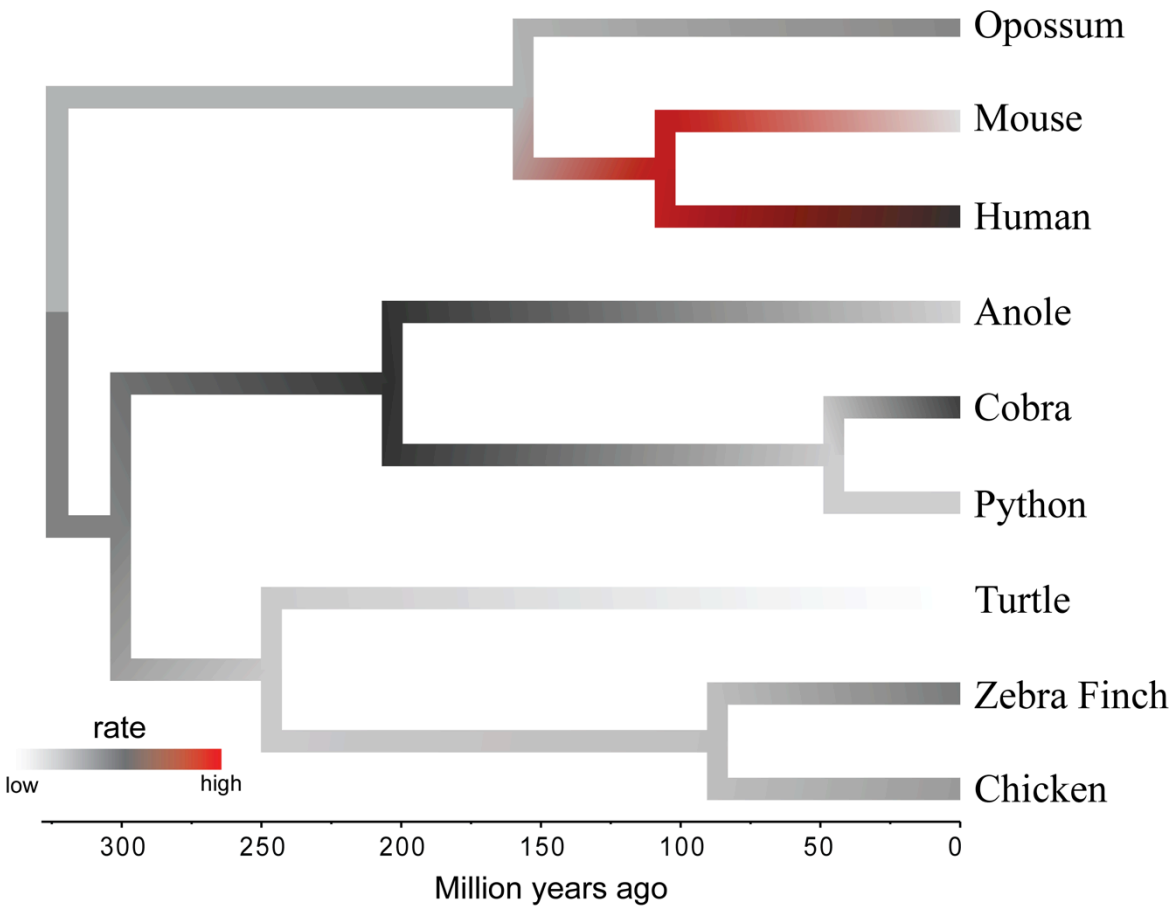
Supplementary Figure S24. Evolution of GC3 composition tetrapod genomes. Data based on Ensembl alignments used for protein evolutionary analysis, filtered to remove taxa with lower gene representation and further filtered to remove all sites that contained any missing data. For each gene, we used the program nhPhyml (47) to calculate ancestral GC3 as well as equilibrium GC3 (GC3*) under a nonhomogeneous model of molecular evolution (four rate categories and estimated transition/transversion ratio and shape parameter). Branch lengths represent Dij, the divergence in GC3 between nodes, to portray the magnitude of GC3 divergence among vertebrates. Colors represent GC-rich (red) through AT-rich (blue) trends.



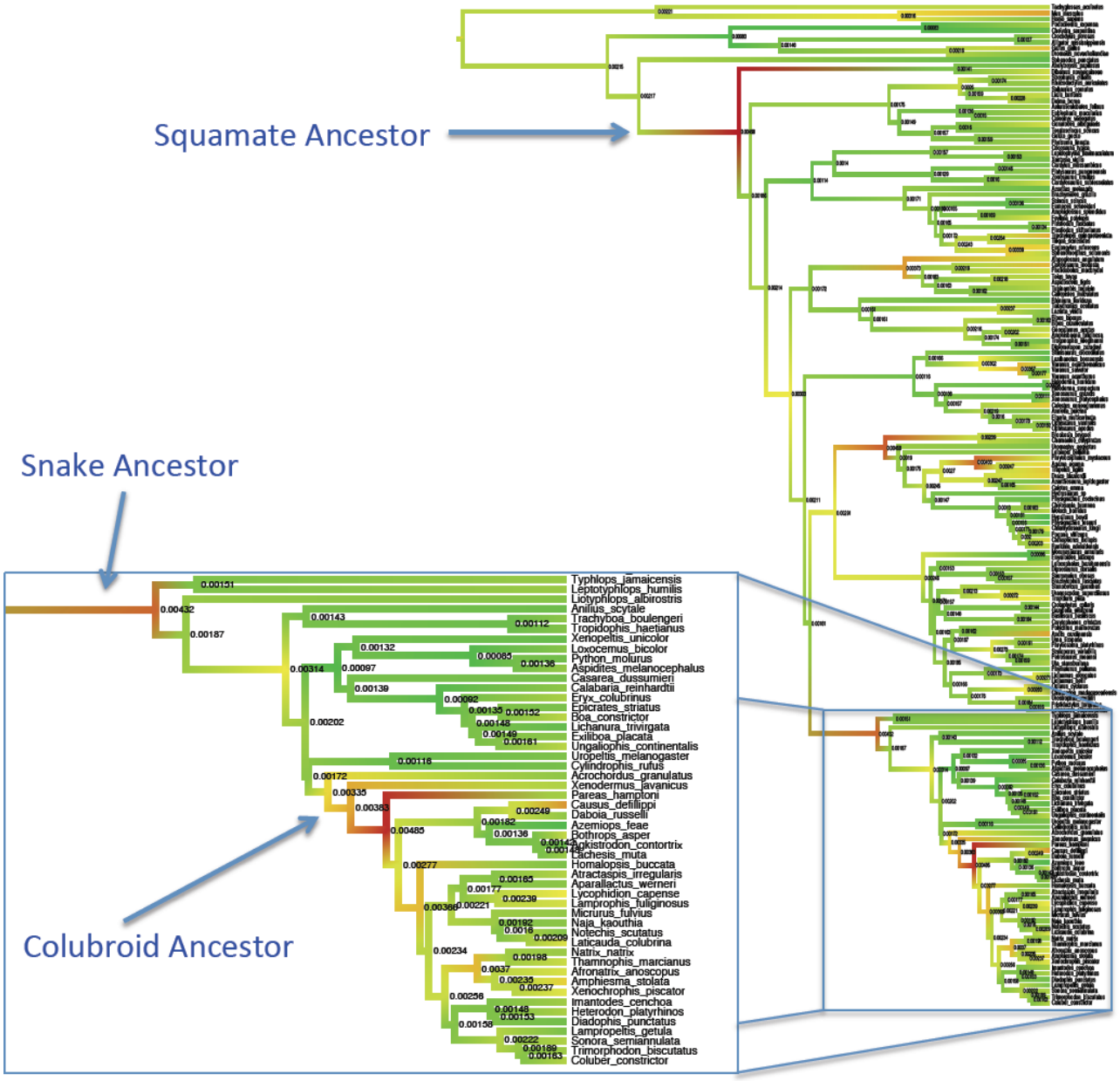
Supplementary Figure S25. Trends in GC composition for aligned regions of squamate reptile genomes. Data based on 3-way genome alignments between the *Anolis* lizard, python, and cobra, indicating the difference in GC composition.



Supplementary Figure S26. Evolutionary rates from 10,000 randomly sampled codons from the ensemble alignments of protein coding gene orthologs (dataset *Ensembl10_10k*).



Supplementary Figure S27. Evolutionary rates from all 62,817 4-fold sites from the Ensembl plus snake alignments of protein coding gene orthologs (dataset *Enselbl10_4-fold*).



Supplementary Figure S28. Estimation of rates of nucleotide evolution across squamate tree based on 4-fold degenerate 3rd codon positions. Data includes 44 genes and 171 taxa. Snake lineage shown magnified in inset. Analyses conducted in BEAST, with two nodes constrained: squamates, and amphisbaenians. Date constraints are only applied at the root (mammal-squamate split constrained to be 310MYA). Colors represent slower rates (green), intermediate rates (yellow-orange), and fast rates (red).

SUPPLEMENTARY REFERENCES

1. Li R*, et al.* (2009) De novo assembly of human genomes with massively parallel short read sequencing. *Genome Research* 20:265-272.
2. Yao G*, et al.* (2011) Graph accordance of next-generation sequence assemblies. *Bioinformatics* 28:13-16.
3. Cantarel BL*, et al.* (2008) MAKER: An easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Research* 18:188-196.
4. Holt C & Yandell M (2011) MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics* 12:491.
5. Yandell M & Ence D (2012) A beginner's guide to eukaryotic genome annotation. *Nat Rev Genet* 13:329-342.
6. Feschotte C, Keswani U, Ranganathan N, Guibotsy ML, & Levine D (2009) Exploring Repetitive DNA Landscapes Using REPCLASS, a Tool That Automates the Classification of Transposable Elements in Eukaryotic Genomes. *Genome Biology and Evolution* 1:205-220.
7. Consortium TU (2010) The Universal Protein Resource (UniProt) in 2010. *Nucleic Acids Research* 38:D142-D148.
8. O'Donovan C*, et al.* (2002) High-quality protein knowledge resource: SWISS-PROT and TrEMBL. *Briefings in Bioinformatics* 3:275-284.
9. Pruitt KD, Tatusova T, Klimke W, & Maglott DR (2009) NCBI Reference Sequences: current status, policy and new initiatives. *Nucleic Acids Research* 37:D32-D36.
10. Korf I (2004) Gene finding in novel genomes. *BMC Bioinformatics* 5:59.
11. Stanke M*, et al.* (2006) AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Research* 34:W435-W439.
12. Castoe T*, et al.* (2011) A multi-organ transcriptome resource for the Burmese Python (*Python molurus bivittatus*). *BMC Research Notes* 4:310.
13. Castoe TA*, et al.* (2011) Discovery of highly divergent repeat landscapes in snake genomes using high throughput sequencing. *Genome Biology and Evolution* 3:641-653.

14. Wall CE*, et al.* (2011) Whole transcriptome analysis of the fasting and fed Burmese python heart: insights into extreme physiological cardiac adaptation. *Physiological Genomics* 43:69-76.
15. Grabherr MG*, et al.* (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotech* 29:644-652.
16. Hunter S*, et al.* (2009) InterPro: the integrative protein signature database. *Nucleic Acids Research* 37:D211-D215.
17. Eilbeck K, Moore B, Holt C, & Yandell M (2009) Quantitative measures for the management and comparison of annotated genomes. *BMC Bioinformatics* 10:67.
18. Ott BD & Secor SM (2007) Adaptive regulation of digestive performance in the genus Python. *Journal of Experimental Biology* 210:340-356.
19. Secor S & Diamond J (1997) Determinants of post-feeding metabolic response in Burmese pythons (*Python molurus*). *Physiological Zoology* 70:202-212.
20. Gregory TR (2013) Animal Genome Size Database. http://genomesize.com.
21. Jurka J (2000) Repbase Update: a database and an electronic journal of repetitive elements. *Trends in Genetics* 16:418-420.
22. Smit AFA, Hubley R, & Green P (2008-2010) RepeatMasker Open-3.0).
23. de Koning APJ, Gu W, Castoe TA, Batzer MA, & Pollock DD (2011) Repetitive Elements May Comprise Over Two-Thirds of the Human Genome. *PLoS Genet* 7:e1002384.
24. Gu W, Castoe TA, Hedges DJ, Batzer MA, & Pollock DD (2008) Identification of repeat structure in large genomes using repeat probability clouds. *Analytical Biochemistry* 380:77-83.
25. Marçais G & Kingsford C (2011) A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* 27:764-770.
26. Quinlan AR & Hall IM (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26:841-842.
27. Sibson R (1973) SLINK: An optimally efficient algorithm for the single-link cluster method. *The Computer Journal* 16:30-34.
28. Tan P-N, Steinbach M, & Kumar V (2005) *Introduction to data mining* (Addison-Wesley Longman, Boston, MA).

29. Niimura Y & Nei M (2003) Evolution of olfactory receptor genes in the human genome. *Proc Nat Acad Sci USA* 100:12235-12240.
30. Niimura Y & Nei M (2005) Comparative evolutionary analysis of olfactory receptor gene clusters between humans and mice. *Gene* 346:13-21.
31. Wang Z*, et al.* (2013) The draft genomes of soft-shell turtle and green sea turtle yield insights into the development and evolution of the turtle-specific body plan. *Nat Genet* 45:701-706.
32. Niimura Y (2009) On the Origin and Evolution of Vertebrate Olfactory Receptor Genes: Comparative Genome Analysis Among 23 Chordate Species. *Genome Biology and Evolution* 1:34-44.
33. Katoh K & Standley DM (2013) MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Mol Biol Evol* 30:772-780.
34. Price MN, Dehal PS, & Arkin AP (2010) FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. *PLoS ONE* 5:e9490.
35. Shimodaira H & Hasegawa M (1999) Multiple Comparisons of Log-Likelihoods with Applications to Phylogenetic Inference. *Mol Biol Evol* 16:1114.
36. Tamura K*, et al.* (2011) MEGA5: Molecular Evolutionary Genetics Analysis using Maximum Likelihood, Evolutionary Distance, and Maximum Parsimony Methods. *Mol Biol Evol* 28:2731–2739.
37. Guindon S*, et al.* (2010) New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0. *Syst Biol* 59:307-321.
38. Anisimova M & Gascuel O (2006) Approximate Likelihood-Ratio Test for Branches: A Fast, Accurate, and Powerful Alternative. *Syst Biol* 55:539-552.
39. Ernst J & Bar-Joseph Z (2006) STEM: a tool for the analysis of short time series gene expression data. *BMC Bioinformatics* 7:191.
40. Du Z, Zhou X, Ling Y, Zhang Z, & Su Z (2010) agriGO: a GO analysis toolkit for the agricultural community. *Nucleic Acids Research* 38:W64-70.
41. Kiełbasa SM, Wan R, Sato K, Horton P, & Frith MC (2011) Adaptive seeds tame genomic sequence comparison. *Genome Research* 21:487-493.
42. Löytynoja A & Goldman N (2008) Phylogeny-Aware Gap Placement Prevents Errors in Sequence Alignment and Evolutionary Analysis. *Science* 320:1632-1635.

43. Zhang J, Nielsen R, & Yang Z (2005) Evaluation of an Improved Branch-Site Likelihood Method for Detecting Positive Selection at the Molecular Level. *Mol Biol Evol* 22:2472-2479.
44. Gharib WH & Robinson-Rechavi M (2013) The Branch-Site Test of Positive Selection Is Surprisingly Robust but Lacks Power under Synonymous Substitution Saturation and Variation in GC. *Mol Biol Evol* 30:1675-1686.
45. Consortium IHGS (2001) Initial sequencing and analysis of the human genome. *Nature* 409:860-921.
46. Fujita MK, Edwards SV, & Ponting CP (2011) The *Anolis* Lizard Genome: An Amniote Genome without Isochores. *Genome Biology and Evolution* 3:974-984.
47. Boussau B & Gouy M (2006) Efficient Likelihood Computations with Nonreversible Models of Evolution. *Syst Biol* 55:756-768.
48. Romiguier J, Ranwez V, Douzery EJP, & Galtier N (2010) Contrasting GC-content dynamics across 33 mammalian genomes: Relationship with life-history traits and chromosome sizes. *Genome Research* 20:1001-1009.
49. Stamatakis A (2006) RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* 22:2688-2690.
50. Shaffer HB*, et al.* (2013) The western painted turtle genome, a model for the evolution of extreme physiological adaptations in a slowly evolving lineage. *Genome Biology* 14:R28.
51. Drummond AJ, Suchard MA, Xie D, & Rambaut A (2012) Bayesian Phylogenetics with BEAUti and the BEAST 1.7. *Mol Biol Evol* 29:1969-1973.
52. Okajima Y & Kumazawa Y (2010) Mitochondrial genomes of acrodont lizards: timing of gene rearrangements and phylogenetic and biogeographic implications. *BMC Evolutionary Biology* 10:141.
53. Pereira SL & Baker AJ (2006) A Mitogenomic Timescale for Birds Detects Variable Phylogenetic Rates of Molecular Evolution and Refutes the Standard Molecular Clock. *Mol Biol Evol* 23:1731-1740.
54. Pyron RA (2010) A Likelihood Method for Assessing Molecular Divergence Time Estimates and the Placement of Fossil Calibrations. *Syst Biol* 59:185-194.
55. San Mauro D (2010) A multilocus timescale for the origin of extant amphibians. *Molecular Phylogenetics and Evolution* 56:554-561.

56. Rambaut A & Drummond AJ (2007) Tracer, v1.4.
57. Wiens JJ*, et al.* (2012) Resolving the phylogeny of lizards and snakes (Squamata) with extensive sampling of genes and species. *Biology Letters* 8:1043-1046.
58. Shedlock AM & Edwards SV (2009) Amniotes (Amniota). *The Timetree of Life*, eds Hedges SB & Kumar S (Oxford University Press, USA), pp 375-379.