doi:10.1038/nature12027

Methods

Genome Assembly and Annotation

Sample collection - Preparation of genomic DNA from *L. chalumnae*

A blood sample from an adult African coelacanth was obtained from Professor Rosemary Dorrington (Rhodes University, Grahamstown, South Africa; CITES permit 043018). This sample was originally from a specimen captured in the Grand Comoros in 2003 by Mr. Ahamada Said of the Coelacanth Education Center (coelacanth accession number SAIAB 97564). The sex of the individual remains unknown. The blood was preserved in PAXgene reagent (Qiagen) and frozen at -20/-80 C and sent on dry ice to the Amemiya lab in 2004. Upon receipt the blood cells were examined microscopically (Supplementary Figure 1a) to insure their intactness and then subjected to flow cytometry in order to determine the genome size (Supplementary Figure 1b). The 2.75 pg/C is in good agreement with the 2.85 pg/C derived from genome sequencing (see main text). The cells were subsequently embedded in low melting point agarose and high molecular weight DNA was prepared[51]. The DNA was subsequently analyzed for quality via pulsed field gel electrophoresis and by partial restriction digestion with EcoRI to judge its suitability for BAC cloning (Supplementary Figure 1cd). We subsequently used this DNA to generate a small number of BAC clones to show proof-of-principle that the DNA obtained in this way was of sufficient quality for BAC cloning (not shown) and, therefore, suitable for the coelacanth genome project. Over 200 μg of this African coelacanth DNA was available for the sequencing project. Genomic DNA was extracted from the agarose using GELase (Epicentre Biotechnologies) and a number of centrifugation and dialysis steps. The resultant genomic DNA was sent to the Broad Institute for preparation of libraries for DNA sequencing.

Sample collection - Preparation of RNA from *L. chalumnae*

RNA isolation was performed *via* a standard lysis and extraction procedure using TRIzol reagent (Invitrogen) on several *Latimeria chalumnae* tissue samples that had been archived at -80 C in Chris Amemiya's lab for > 20 years. Samples were run in an Agilent 2100 Bioanalyzer (Supplementary Figure 2). It was not expected that any of the samples would be usable given the long period over which they had been archived and the fact that most of the tissues were taken from specimens that had expired. Surprisingly, one sample (muscle, bottom left), had a decent RNA integrity value (7.7) and did not show overt degradation as seen in all the other samples. This muscle sample was from a female specimen from the Virginia Institute of Marine Sciences (VIMS 08118, CCC number 141)[52]. The tissue sample was obtained prior to *Latimeria chalumnae* being placed under CITES protection (thus, no CITES permit was necessary at the time). This muscle RNA sample was sent to the Broad Institute for RNA sequencing and analysis.

Sample collection - Preparation of RNA from *Protopterus annectens*

A juvenile specimen of the African lungfish, *Protopterus annectens*, was obtained from a tropical fish distributor. This specimen measured 31 cm in standard length and weighed 127 grams. The specimen was processed under ACUC protocol 06AM01 to Chris Amemiya. The specimen was euthanized using a lethal dose of MS222 and immediately dissected. Several tissues were taken, including brain, blood, gonad, gut, liver, skin, muscle, fins. Tissues were divided in half -- one half was frozen at -80 C and the other half was preserved in RNAlater (Qiagen). Small samples of each tissue were also taken for histological examination and the entire carcass was radiographed in order to count vertebral ribs for definitive taxonomic identification[53]. RNA was isolated *via* a standard lysis and extraction procedure using TRIzol reagent (Invitrogen). RNAs from brain, kidney+gonad and liver+gut were sent to the Broad Institute for RNAseq analysis. This voucher specimen and its frozen tissues will be deposited in the University of Washington Fish Collections. A complete mtDNA was assembled from the RNAseq data from this specimen and Blast searches against GenBank nr collection confirmed its identity as *Protopterus annectens*.

Genome Sequencing and Assembly

The *Latimeria chalumnae* assembly, LatCha 1.0 was constructed from 180 bp paired end fragment libraries (61X coverage), 3 kb jumping libraries (88X coverage), and 40 kb FOSSILLs[54] (1X coverage). All libraries were sequenced by Hi-Seq Illumina machines, producing 101 bp reads. Assembly of the Coelacanth genome was carried out using a pre-publication version of the software program ALLPATHS-LG[55]. Data from the sequencing instruments was imported directly into the program, without any filtering or other preprocessing. In brief, the ALLPATHS-LG algorithm then proceeded by correction of sequencing errors within reads, closure of short-fragment read pairs, formation of an initial de Bruijn from these filled fragments, and disambiguation of the graph using paired ends from the jumping libraries. Unfortunately, we were only able to obtain enough high molecular weight DNA to make the longest jumping libraries (40 kb), and could only use medium weight DNA for the mid-range jumping libraries (3 kb); this is very likely to have adversely affected the genome assembly. The resulting assembly has a heterozygosity rate of 1/357 bp.

RNA-sequencing and Assembly

Total RNA was isolated from muscle tissue of a single *Latimeria chalumnae*. The RNA was checked for quantity and quality using a BioAnalyzer (Agilent) and then 5 micrograms was treated with Turbo DNase (Ambion) according to the manufacturer's recommendations. Samples were shown to be free of residual, detectable genomic DNA with a qPCR assay (data not shown). An Illumina RNA-seq library was generated from this RNA as previously described[56] with the following modifications. We performed four rounds of oligo (dT) selection with the Dynabeads mRNA purification kit (Invitrogen) and then incubated with RNA fragmentation buffer (Affymetrix) at 80°C for 4 minutes. Indexed adaptors for Illumina sequencing were ligated onto end-repaired, A-tailed cDNA followed by two rounds of size selection with 0.7 volumes of AMPure beads (Beckman Coulter Genomics). We used 10 PCR cycles to amplify the library followed by one round of clean up with 0.7 volumes of AMPure beads. The library was sequenced on one lane of a flowcell with a Hi-Seq Illumina machine, producing 210,146,976 101 base, paired end reads. Three lungfish strand-specific dUTP libraries (brain, gonad/kidney, gut/liver) were produced from

Oligo dT polyA-isolated RNA. The libraries were sequenced by Hi-Seq Illumina machines, producing 76 bp reads (3-4 Gb of sequence/tissue). All four RNA-seq datasets were assembled via the genome-independent RNA-seq assembler Trinity[57].

Annotation - Ensembl

The Ensembl gene annotation pipeline was used to create gene models for coelacanth. The genome was repeat masked with RepeatMasker[58] (version 3.2.8) using a custom coelacanth library created by RepeatModeler. In total 41% of the genome was masked by RepeatMasker. Low complexity mapping was performed using Dust[59].

Little protein or cDNA evidence was available for coelacanth; aligning Uniprot[60] protein sequences using Genewise[61] produced only 153 models. Genbank cDNAs were aligned with Exonerate and produced 46 models. The vast majority of the gene annotation came from models built using orthologous proteins and from RNA-seq.

Othologous proteins were placed on the genome by running a BLAST[62] alignment of Uniprot sequences against Genscan exons. Alignments were divided into groups based on taxonomy and the Uniprot Protein Existence (PE) classification. Proteins were selected for re-alignment in such a way as to favor PE level 1 and 2 proteins over other PE levels and mammalian proteins over non-mammalian proteins. The selected proteins were then realigned to the genomic sequence using Genewise. In total 297,885 coding models were created. The orthologous models were clustered and the clusters filtered to select models with the most agreement with RNA-Seq split reads; 50,773 models were selected for use in the final gene set.

RNA-Seq models were created by aligning a set of 375 million paired end Illumina reads to the genome using BWA, this resulted in 225 million reads aligned and properly paired. The Ensembl RNA-Seq pipeline was used to process the BWA alignments and produced a further 30 million split read alignments using Exonerate. The alignments were processed further to produce 23,058 transcript models, one transcript per loci. RNA-Seq transcripts were assessed for quality by a BLAST search of the predicted open reading frames against a Uniprot PE 1 and 2 dataset, models with no BLAST alignment or poor BLAST coverage were discarded. The resulting models were added into the gene set where they produced a novel model or splice variant, in total 4,994 models were added.

Where RNA-Seq models were considered to be too fragmented for use in the final gene set they were used to add UTR to orthologous models.

The final gene set was created by combining transcripts from the three evidence sources. Redundant transcripts were removed and overlapping transcripts were clustered into multi transcript genes. In Total 19,033 protein coding genes were produced containing 21,817 transcripts. A total of 141 pseudogenes were identified, finally 2,894 short non-coding RNAs were added.

Annotation - MAKER

MAKER version 2.22[63] was run on *Latimeria chalumae* using assembled *Latimeria chalumae* mRNA-seq data from muscle, and all SwissProt proteins as evidence (downloaded November 17, 2011 from ftp://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/complete/). Repetitive regions were masked using a custom repeat library provided by Ensembl, all organisms in Repbase[64], and a list of known transposable elements provided by RepeatMasker. Additional areas of low complexity were soft masked[65] using RepeatMasker to prevent the seeding of evidence alignments in those regions but still allowing extension through them when appropriate[65-66]. Genes were predicted using SNAP[67] and Augustus[68-69] trained for *Latimeria chalumae* using MAKER in an iterative fashion as described by Cantarel et al.[66].

The initial MAKER annotation set contained 15,839 protein coding genes supported by protein or mRNA seq evidence. In addition to these high-confidence MAKER generated annotations, SNAP and Augustus produced an additional 74,769 gene predictions that did not overlap MAKER annotated gene models or evidence. These predictions were evaluated with InterProScan[70] to identify those containing protein domains. 9,138 of these predicted genes with protein family domains were added as genes to the final annotation set bringing the final gene count to 24,977 of which 84% contain a protein domain as detected by InterProScan[70], and 60% of which have an annotation edit distance less than 0.5, consistent with a reasonably well annotated genome[63,71]. In addition 455 of the 458 core eukaryotic proteins identified by Parra et al., are represented in the final annotation set[72] and 94% of the annotated genes have similarity to proteins in SwissProt by BLAST (E<.000001)[62]. Next the annotations from Ensembl were passed to MAKER along with newly available liver and testes mRNA-seq data from *Latimeria menadoensis* kindly provided by the labs of Giuseppe Scapigliati, Alberto Pallavicini, and Ettore Olmo, which were assembled using Trinity[57] and aligned to the genome using BLAST[62]. When existing gene models overlapped these aligned mRNA-seq data, the one with the lowest AED score was advanced to the final annotation set. All non-overlapping models from both annotation sets were advanced to the final annotation set, and new gene models were created when supported by the new mRNA-seq evidence. This produced a super-annotation set containing 29,237 protein coding gene annotations of which 68% contain a protein domain as detected by InterProScan[70], consistent with the average domain content of six reference proteomes[63], and 75% of which have an annotation edit distance less than 0.5, suggesting an overall improvement in the annotation with the addition of the Indonesian coelacanth mRNA-seq data[63,71]. Consistent with this conclusion, an additional core eukaryotic protein[72] was also identified and annotated, bringing the total to 456/458, with 84% of the total annotated genes have similarity to proteins in SwissProt identified by blast (E<.000001)[62].

Annotation - lincRNAs

Identification of long-intergenic RNA in coelacanth

Reads were first mapped independently to the *Latimeria chalumnae* v1 assembly using Bowtie to compute the average and the standard deviation of the insert size. These values were used as parameters for mapping reads with TopHat[73]. Transcripts were reconstructed using Cufflink[74] and collapsed into a consensus set using Cuffcompare and the Ensembl pre-annotations.

The coding potential of each transcript from an intergenic locus was assessed using both forward and reverse orientation using Coding Potential Calculator[75]. Only multi-exonic loci for which all transcripts were deemed non-coding (value < 0) were retained for further analysis.

We used the coelacanth centred alignments (see CNE Evolution Methods) of the human, mouse, dog, elephant, opossum, chicken, green anole, African clawed frog, coelacanth and stickleback genomes to compute the conservation scores of each nucleotide within the alignment using PHAST[76]. Conservation scores within exons and introns of both lncRNA loci and both protein-coding genes were compared to each others and to intergenic sequences in coelacanth using a Kolmogorov-Smirnov test.

Identification of positional equivalents in human mouse and zebrafish

We compiled comprehensive lncRNA data sets in human based on long non-coding RNA loci annotated by Ensembl and lncRNA loci identified by Cabili et al[77] (9,085 non-redundant loci in total). The data set in mouse comprises 2,175 non-redundant lncRNA loci called by Ensembl and by Belgard et al.[78] in the necortical layers of mouse, while the zebrafish set is composed of 1,434 non-redundant lncRNA loci identified by Ulitsky et al.[79] and Pauli et al.[80].

We identified positional equivalents between coelacanth and human, mouse or zebrafish, as lncRNA loci that are found in the same relative orientation and in conserved synteny relative to neighbouring protein-coding genes in the two species analysed. The searched focused on genes with one to one orthologs between coelacanth and human, coelacanth and mouse, or colacanth and zebrafish.

In order to obtain further evidence for the conservation of lncRNAs we used the coelacanth-centered multiple sequence alignment to find the exact sequence positions homologous to the splice sites in the latimeria lncRNAs and determined whether an experimentally known splice site is annotated in RefSeq or an EST dataset at this position for any of the other 8 aligned species.

Annotation – ncRNAs

The small ncRNA annotation was complemented by intensive manual curation to tag likely pseudogenes in particular for snoRNAs and to include 5 additional ncRNAs. See Supplementary Dataset 3.

Among the 127 microRNA families annotated in the *L.chalumnae genome*, 47 have homologs in tetrapods. Five microRNA families (mir-1248, mir-3064, mir-3536, mir-3538 and mir-599) are specific to the Sarcopterygii.

Conserved secondary structure elements were determined using RNAz 2.0[81].

Coelacanth duplicated genes

Proteomes for *Caenorhabditis elegans*, *Ciona intestinalis*, *Drosophila melanogaster*, *Gallus gallus*, *Gasterosteus aculeatus*, *Homo sapiens*, *Saccharomyces cerevisiae* and *Takifugu rubripes* were obtained from the Ensembl Core database (Version 65). The *Latimeria chalumnae*'s proteome was obtained from

the Ensembl Core database (Version 66). Isoforms were removed from all proteome sets which would bias subsequent phylogenetic analyses.

The proteomes for the above 9 species were filtered to remove protein sequences less then 100 amino acids that resulted in a dataset of 143,929 protein sequences that was submitted to OrthoMCL [82](Version 2.0.2) for protein clustering. A total of 18,307 OrthoMCL groups (also referred to as OG groups) were obtained using an E-value=1E-5 and an effective database size of 143,929 sequences for BLASTP[83] (Version 2.2.21) with a Markov Chain Clustering (MCL) inflation parameter of 1.5.

These 18,307 OrthoMCL groups were filtered using a custom PERL script to provide a subset of OrthoMCL groups that contained 2 or more *L. chalumnae* proteins and at least proteins for two other species within an OrthoMCL group resulting in 1,762 OrthoMCL groups containing a total of 27,641 proteins identified for further Phylogenetic analysis.

Proteins in each of the 1,762 OrthoMCL groups were aligned using Muscle, MAFFT and ClustalW and the resulting alignments were combined by M-Coffee to produce the final multiple sequence alignments (MSA)[84]. Any MSA alignment columns that contained 50% or more gaps were stripped from the final alignments. Each set of aligned sequences was used as input to reconstruct phylogenetic trees using a Maximum Likelihood approach implemented in PhyML 3.0[85]. ProtTest 3[86] was used to select the amino acid substitution model that best fits the proteins alignments. PhyML 3.0[85] was used to reconstruct maximum likelihood (ML) trees using the selected model from ProtTest with optimized number of invariable sites and optimized across site rate variation. Bootstrap values were calculated using the aLRT model [87]. The 1,762 OG groups were dynamically rooted by species using the outgroup setting implemented in PhyML and the resulting ML trees were analysed and visualized using the ETE2 python toolkit[88].

To get a general overview of all duplications in the complete *L. chalumnae* genome relative to the other species' genomes, gene families of paralogous proteins were identified. The 1,762 ML trees obtained from PhyML analyses were used as input to detect duplication and speciation nodes. The species overlap algorithm (SO) was used as described in the ETE toolkit. In total 226 trees were found to contain *L. chalumnae* –specific duplication events. Among the 226 trees, 115 trees showed low bootstrap values (<50%), whereas 111 trees showed high support (>50%). These 226 trees comprise a total of 404 duplicated gene pairs specific to *L. chalumnae*, of which 336 gene pairs have a bootstrap support of ≥ 50%.

A genome background of non-redundant InterPro IDs and their corresponding annotations was obtained from the Ensembl core database for all *L. chalumnae* genes that have an annotation. These non-redundant InterPro domains were used to map their corresponding GO IDs and annotations from the InterPro2Go mapping file (April/May 2012 release) to the *L. chalumnae* genome wide Ensembl IDs. The *L. chalumnae* genome background of GO terms was further filtered to remove any redundant GO terms for the same gene that arises due to a single GO term mapping to multiple InterPro terms for the same gene. In total, 1,846 unique GO terms could be mapped to 13,123 *L. chalumnae* genes.

Gene enrichment analysis for the 336 *L. chalumnae* paralogous gene pairs (bootstrap values ≥ 50%) against the genome wide distribution of GO and InterPro terms using both a Hypergeometric and a Fisher's test was conducted in R using the coRNA package[89] with Bonferroni correction for multiple hypothesis testing.

Determining the closest living fish relative of the tetrapod ancestor

Construction of the phylogenomic datase

To resolve difficult phylogenetic questions, the dataset has to be of high quality, i.e., composed of orthologous genes, devoid of sequence contaminations, frameshifts, annotation errors, etc[90-91]. We therefore applied stringent criteria and performed multiple controls to assemble our supermatrix.

The complete proteomes of 16 jawed vertebrates were downloaded from the Ensembl database (release 66). Selected organisms included eight mammals (one monotreme, *Ornithorhynchus anatinus*; two marsupials, *Monodelphis domestica* and *Macropus eugenii*; and five placental mammals, *Dasypus novemcinctus*, *Loxodonta africana*, *Canis familiaris*, *Homo sapiens* and *Mus musculus*); three birds (*Gallus gallus*, *Meleagris gallopavo* and *Taeniopygia guttata*); the lizard *Anolis carolinensis*; the frog *Silurana (Xenopus) tropicalis*; the coelacanth *Latimeria chalumnae,* and two ray-finned fishes (*Danio rerio and Takifugu rubripes*). Proteins shorter than 100 amino acids were discarded and, for alternatively spliced genes, only the longest splice variant of each gene was retained. A pseudo-complete proteome of the West African lungfish *Protopterus annectens* was generated based on RNA-seq data from three different tissues (brain, liver/gut and kidney/gonad; see above). Briefly, pooled reads were assembled with Trinity as above and the resulting transcripts were translated with Bioperl[92]. For each transcript, only the longest ORF (at least 150 amino acids) was retained, while redundancy was reduced using the dereplication mode of USEARCH 5.2.32[93]. Groups of orthologous proteins were determined with OrthoMCL 2.0.3[82] using USEARCH instead of BLAST and an inflation parameter value of 1.5. This procedure led to 7,764 OrthoMCL groups for which both *Latimeria* and *Protopterus* were present. To maximise orthology, we eventually retained the 373 groups for which exactly one copy had been identified in each of the 17 species and for which the length of the coelacanth ortholog was at least 300 amino acids.

In a second step, the 373 OrthoMCL groups were aligned with MAFFT v6.717b and orthologous sequences from three chondrichthyan (*Leucoraja erinacea*, *Scyliorhinus canicula* and *Callorhinchus milii*), a second frog (*Rana chensinensis*) and a third ray-finned fish (*Oreochromis niloticus*) were added using HaMStR[94] v8b. For the chondrichthyans, short-read nucleotide assemblies provided by the authors were used (http://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?study=SRP004911), while for *Rana*, publicly available short reads (http://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?study=ERP001146) were locally assembled using Mira[95] with the kind permission of the data generators. In contrast, newly released *Oreochromis* sequences were directly added as predicted proteins downloaded from Ensembl 67, again using HaMStR.

Due to the high level of errors in the critical genome sequence of *X. tropicalis*, representing the most basal group in tetrapods, all transcripts from both *X. tropicalis* and *X. laevis* available at the NCBI were

downloaded and added to the alignments using the program Forty (Denis Baurain, unpublished). A second improved version of the assembly of *Rana chensinensis* generated in the meantime by Mira was also added with Forty to improve its sequence coverage. The resulting alignments were manually verified; in particular, highly variable sequences, usually corresponding to distant paralogs, and probable contaminant sequences (e.g., *Xenopus* sequences clustering within mammals) were eliminated.

To automatically detect sequencing errors (particularly when resulting in frameshifts) and annotation errors, which are frequent in high-throughput data [90,96], we used the software HmmCleaner (Raphael Poujol, unpublished). Briefly, for each sequence of an alignment, a Hidden Markov Model (HMM) profile is computed for the alignment minus the sequence using HMMER [http://hmmer.org;[97]. Then, every region of the sequence having diverged more than a specific accumulative score from the HMM profile is discarded from the final alignment. The score has to be estimated empirically and depends on both the taxon sampling and the divergence among the species.

Ambiguously aligned positions were then removed using Gblocks[98] with stringent parameter values (b2 = 85%, b3 = 8, b4 = 10, b5 = none). To maximize the information content of our alignments, we used custom Perl scripts to retain only the positions having a known character state (nor missing nor gap) for the three key taxa of our analysis: *Latimeria*, *Protopterus* and and a least one chondrichthyan, as these are the slowly evolving outgroups. After all these filtration steps, only the genes with at least 200 positions and the sequences with at least 100 amino acids were kept. The supermatrix was then assembled using SCaFoS[99], considering only the genes with at most 1 missing species, which yielded a dataset of 252 proteins and 100,984 unambiguously aligned positions.

To ensure that our dataset did not contain paralogous or contaminant sequences, we ran a congruence test[100]. Briefly, single gene trees were inferred with the LG+F+Γ_4 model and 100 bootstrap replicates using RAxML[101] and we looked for bipartitions in all single gene trees with a bootstrap support ≥ 70% that were incongruent with the supermatrix based tree. This analysis only revealed a contamination of *Xenopus* by a human sequence, the other incongruences being most likely explained by stochastic or systematic errors. The final dataset thus consisted of 251 proteins and 100,583 positions.

Inference of the phylogeny

Phylogenetic trees were inferred using RAxML[101] with the site-homogeneous LG+F+Γ_4 and GTR+Γ_4 models, and using PhyloBayes[102] with the site-heterogeneous CAT+GTR+Γ_4 model[103]. We used cross-validation to determine the best fitting model, as described in Lartillot and Philippe 2008[104]. The analysis was performed in PhyloBayes v3.3, using ten randomly generated replicates, in which the original data set was divided into training data sets (9/10 of the positions) to estimate the parameters of each model and test data sets (1/10 of the positions) to calculate likelihood scores with the corresponding parameters. To estimate the statistical support, bootstrap on positions was used for the LG and GTR models, and jackknife of proteins for the CAT+GTR model (with random sampling of 66% of the proteins, which is similar to the number of characters discarded in a bootstrap replicate).

How slowly evolving is the coelacanth?

Protein-coding gene evolution

Relative rate of gene evolution

To test the rate of evolution of coelacanth relative to other species we performed two types of analyses, Tajima relative rate test[105] and Two-Cluster test [106], on the carefully curated dataset used for the phylogenomic analysis (see section "Determining the closest living fish relative of the tetrapod ancestor").

Tajima Relative Rate Test

First, we applied Tajima relative rate test (RRT) on the sequence alignments of a dataset consisting of approximately 250 genes. Each gene-set was separately aligned and sites with gaps or unknown amino acids were excluded. Each comparison included two ingroups and one outgroup. For each such triplet, we concatenated all the aligned gene-sets that included all three species and performed the Tajima RRT using in-house perl scripts. The relative rates of evolution between coelacanth and other species (lungfish, human, mouse, chicken and dog) were evaluated using each of the three chondrichthyan species as outgroup (*Leucoraja erinacea*, *Callorhinchus milii*, *Scyliorhinus canicula*). Tajima RRT analysis shows that coelacanth is not only evolving significantly slower than any of the tetrapod species used but also more slowly than lungfish ($p < 0.05$; Supplementary Dataset 6). An only slightly different picture is revealed on the respective analysis between lungfish and tetrapods. Lungfish is evolving significantly slower than human, mouse and dog, but seems to evolve as fast as the chicken. As can be seen in **Figure 1**, the substitution rate observed on the coelacanth lineage is approximately half that of tetrapods. Because branch lengths may be underestimated in regions of a tree that have few species, here potentially confounding the analysis of the coelacanth branch, we examined the node-density effect[107-108] in each tree of the Bayesian posterior distribution but found no evidence for this artifact.

Two-Cluster analysis

As a further step, we performed an analysis based on the estimated phylogeny described in the phylogenomics section. We calculated pairwise distances between taxa from the branch lengths of the inferred phylogenetic tree (Figure 1) using the R modules *ape*[109] and *geiger*[110]. Based on those distances we performed the "Two-Cluster" test proposed by Takezaki *et al*.[106]. We compared the mean distance of coelacanth – representing a single-taxon cluster – to the outgroup and the mean distance of the monophyletic cluster of mammals to the outgroup. The outgroup cluster consisted of the three chondrichthyan species used in the phylogenomics dataset. The variance for each distance was obtained by comparing them to the respective distances calculated from three different tree datasets (method described in Kumar & Filipski[111]). These tree datasets included the consensus trees of 100 jackknife datasets, and two datasets retrieved from the CATGTR chains consisting of 100 and 400 sampled trees respectively (sampling every 10 trees with a burn-in of 100). Finally, we tested through Z statistics whether the difference between the distances of the two clusters (coelacanth vs. mammals) is significantly different from zero. The same analysis was conducted for the cluster pairs lungfish-mammals, coelacanth-lungfish, coelacanth-chicken and lungfish-chicken. Results from the Two-Cluster test (Supplementary Dataset 5) consistently showed once again that the rate of evolution of mammals is

significantly higher than that of coelacanth as inferred also with Tajima RRT. Coelacanth is slower than lungfish and the latter is slower than mammals. Testing the rate difference between coelacanth-chicken and lungfish-chicken showed that chicken is evolving faster in both comparisons. Interestingly, the Two-Cluster test estimated a faster rate for chicken compared to lungfish, which was not inferred with Tajima RRT. This inconsistency between the two methods can be attributed to the fact that the distances used for the two-cluster test are based on a phylogenetic model and can therefore reflect more accurately the evolutionary distance between taxa and clusters of taxa.

Overall, comparison of the relative rate of evolution between proteins of coelacanth and tetrapods or lungfish confirms that in terms of protein evolution coelacanth is indeed slowly evolving. However, lungfish follows the same pattern but to a lesser extent.

Transposable element analysis

Annotation of repeat elements - Repeat library construction

A repeat library was built using RepeatScout (version 1.0.5) with an lmer size of 16. Due to memory requirements, 1/3 of *Latimera* scaffolds (1216 out of 4053) were used to detect high-frequency mers. The whole genome was masked with RepeatMasker (version 3.3.0). Elements that occurred less than 10 times were filtered out. The remaining sequences were annotated with three different methods: (1) RepeatMasker using RepBase version 14.11, (2) TBlastX against RepBase 14.11, and (3) BlastX against a custom non-redundant collection of proteins belonging to transposable elements from NCBI (keywords: “retrotransposon”, “transposase”, “reverse transcriptase”, “gypsy”, “copia”). The best annotation between the three methods was chosen based on alignment coverage and score. Sequences were manually curated to remove spurious matches.

The automatic annotation was followed by a detailed precise manual annotation which allows us to detect more divergent and less frequent transposable elements. The manual annotation characterizes specific features such as LTR, TIRs or TSDs.

Both automatic and manual annotations were combined to generate the repeat library.

Repeat content estimation

The repeat content and the copy number of each family were estimated by the following three steps.

(1) The genome was masked by using RepeatMasker with the corresponding repeat libraries.

(2) An in-house Perl script was used to parse the results from RepeatMasker. The Perl script reports (a) on the base pair level, total number of the base pairs that were masked and the percentage of the whole genome (b) on the transposable elements (TE) family level (assigned based on Repbase classifications), the copy number, total base pairs and the percentage of the whole genome for each TE family.

(3) An in-house script was used to parse the output file from RepeatMasker. The number of hits per element was counted, and results were grouped per family. The size of each element was assigned as reference size in order to estimate whether the hits are complete elements or not. The total number of hits was counted, followed by the number of hits that make 30%, 50% and 80% of the reference size.

Transposable element analyses in the coelacanth transcriptome

The transcriptome of *Latimeria chalumnae* was masked using RepeatMasker with the same library used for the repeat content estimation. The number of expressed copies per family was estimated by the method described above.

Genome rearrangement

Analyses of conserved synteny were performed using the coelacanth-anchored LastZ multispecies alignments. For these analyses, a scaffold was considered homologous to a particular chromosomal region if ≥500 bp of the scaffold aligned to that region and if the alignment spanned ≥20% of the scaffold (i.e. excluding small transposition events). To identify candidate synteny breaks, all coelacanth scaffolds that aligned to two different chromosomes (within each reference species) were concatenated and the number of base pairs corresponding to each pair of chromosomes was tabulated. Statistical analyses of candidate synteny breaks were performed to compare observed patterns to those that could be randomly generated by missassembly. These compared (a) the number of observed 500 bp intervals corresponding to each chromosome pair to (b) the number expected for that pair within a random sample equal in size to the number of break-informative intervals observed in the pair of genomes (876,489 break-informative intervals in the chicken/coelacanth alignment), expressed as Bonferroni corrected Chi-square tests.

Coelacanth species comparison - transcriptomes

The identity percentage on a nucleotide level between *Latimeria chalumnae* and *Latimeria menadoensis* was calculated within the coding DNA sequence only, from the initial ATG to the STOP codon, whenever available.

A total of 5,608 coding sequences with a minimum length of 500 codons were selected from the *Latimeria menadoensis* transcriptome obtained from liver and testis RNA-seq sequence data assembly. The open reading frames were predicted using the “Find Open Reading Frames” tool included in CLC Genomic Workbench v5.1 (CLC Bio, Katrinebjerg, Germany), using the “open-ended sequence” option.

The *L. menadoensis* sequences were aligned with the *L. chalumnae* genomic scaffolds with BLASTn. Significant alignments were concluded at an e-value of 1e-25 and only alignments of at least 80 nucleotides were considered for the comparison between the two species. The identity percentage was calculated as the percentage of identical nucleotides over the total number of nucleotides aligned.

Coelacanth species comparison – BACs and genome

A total of 26 *L. menadoensis* BAC contig sequences, totaling 5.3 Mb were used as BLAST queries to identify orthologous *L. chalumnae* scaffolds. The *L. menadoensis* BAC contigs were then aligned to their orthologous *L. chalumnae* scaffold sequences using Megablast optimized for highly similar sequences (bl2seq)[112]. Regions of direct homology between each *L.chalumnae* scaffold and its orthologous *L. menadoensis* contig were estimated by examining a dot-plot generated from the blast results, and the corresponding alignments were examined for evidence of mis-assembly within the *L. chalamnae* scaffolds. Aligned regions totaling 3.8 Mb of sequence, excluding runs of N within the scaffolds, were subsequently examined to provide a direct estimation of the overall sequence similarity between *L. chalumnae* and *L. menadoensis*.

Coelacanth informing the vertebrate adaptation to land

Tetrapod gene loss

A list of predicted tetrapod (human, opossum, platypus, chicken, lizard, frog) and teleost (zebrafish, stickleback) orthologs of coelacanth genes was downloaded from Ensembl66 using Biomart. Coelacanth genes that had no predicted tetrapod ortholog but a predicted teleost ortholog were kept as candidates for tetrapod-specific gene losses. Genes that had no associated gene name in zebrafish were removed from the list. The remaining genes were checked manually for EnsemblCompara tree[113] topologies and synteny data from the Synteny Database[114] and Genomicus[115] consistent with a pattern of tetrapod gene loss. See Supplementary Figures 9 and 11 as an example. Finally, we used tblastn searches (E-value ≤1e-10) of the longest predicted coelacanth protein sequence against the genome assemblies of human, mouse, chicken, lizard, and frog and analyzed the five best hits to exclude cases of missing tetrapod gene annotations in Ensembl66. Gene expression and phenotypic data for zebrafish were obtained from ZFIN: The Zebrafish Model Organism Database (www.zfin.org).

CNE evolution

Alignments

Using human or coelacanth as reference genome, whole-genome alignments of 10 vertebrates (human hg19, mouse mm9, dog canFam2, elephant loxAfr3, opossum monDom5, chicken galGal3, lizard anoCar2, *Xenopus tropicalis* xenTro3, stickleback gasAcu1) were carried out. For human-centric multiple alignments, pairwise alignments for all species except coelacanth were downloaded from UCSC Genome Browser. Pairwise alignments were generated using LASTZ-1.02.00[116] (parameters B=2 C=0 E=30 H=2000 K=2200 L=6000 O=400 T=2 Y=3400). The alignments were reduced to single-coverage with respect to the respective reference genome using UCSC tools for 'chaining' and 'netting'. Multiple alignments were generated using MULTIZ.v11.2/roast.v3[117] and the tree topology "(((((((((Human, Mouse), Dog), Elephant), Opossum), (Chicken, Lizard)), Xenopus), Coelacanth), Stickleback)".

Analysis of conserved noncoding elements

A neutral model was built by running PhyloFit (general reversible "REV" substitution model) on four-fold degenerate sites (as defined by human CCDS genes obtained from UCSC Genome Browser) of human

chromosomes 1-22 extracted from the 10-species alignment. A second neutral model was similarly built for chromosome X. Conserved sequences in the human genome were predicted using PhastCons. The parameters were target coverage of input alignments = 0.3 and average length of conserved sequence = 45 bp; the conserved model was defined as rho=0.3 times that of the neutral (non-conserved) model. A total of 1.52 million conserved elements that span 132 Mb (4.3%) of the human genome (chromosomes 1 – 22, X; 3,036 Mb) were identified. To assess the sensitivity of this approach to functional elements, the PhastCons elements were compared against the human CCDS gene set. 92% of CCDS exons (243,893 / 265,689 exons on the autosomes) were overlapped (minimum coverage 10%) by a PhastCons element.

The conserved elements were compared against the genomic locations of human genes from Ensembl Release 65 (21,136 protein-coding genes, 12,930 pseudogenes, 12,441 RNA genes, 1,838 RNA pseudogenes) and classified as sequences of protein-coding genes, UTRs, RNA genes, pseudogenes, intronic and intergenic regions. The intronic and intergenic elements were further filtered against human mRNAs (~361,000 sequences) and spliced ESTs (~4 million sequences) yielding 996,331 conserved noncoding elements (CNEs). We then excluded CNEs shorter than 30 bp and focused our analysis on 739,646 CNEs covering 81,041,077 bp or 2.7% of human genome, ranging in size from 30 bp to 2,707 bp with an average length of 110 bp (median length, 74 bp). To pinpoint the evolutionary branch of origin of these CNEs, a CNE that was at least 30% alignable to another genome was deemed to be present in that genome and then the most recent common ancestor that contained the CNE was set as the branch of CNE origin. In addition, PhyloP (using likelihood ratio test method) was used to see if a CNE was under statistically significant constraint (p-value < 0.01) at more recent branches in the species tree. If there was a significant onset of constraint at a more recent branch, the branch of origin of that CNE was revised. The branches of origins of various CNEs are given in Supplementary Table 17. Note that the number of CNEs identified in the sarcopterygian lineage (53,985 CNEs) is likely to be an overestimate because previous studies have shown that a significant proportion of CNEs that originated in the gnathostome ancestor has diverged beyond recognition in teleost fishes[118-119]. Thus, some CNEs identified in the sarcopterygian lineage using stickleback as basis for comparison may have actually originated in the gnathostome ancestor but diverged in the stickleback.

To determine enrichment of regulatory elements in CNEs of tetrapod and sarcopterygian origin, 5,119 p300 binding sites predicted in E11.5 mouse forebrain, midbrain and limb, were obtained from Visel et al.[120]. The p300 sites were “lifted over” from mouse mm9 assembly to human hg19 assembly (resulting in 4,528 sites), and overlapped with the human CNEs (minimum 30 bp overlap). Enrichment of p300 sites in CNEs was found using a binomial distribution of the overlaps between p300 sites and 1,000 sets of randomly selected noncoding regions in the human genome. One-tailed P-values were calculated.

To identify the genes putatively regulated by tetrapod and sarcopterygian CNEs, we assigned each CNE to its single nearest gene within 2 Mb in the human genome. While the tetrapod CNEs were found to be associated with 8,886 genes, sarcopterygian CNEs were associated with 8,376 genes. For functional enrichment of genes that are associated with CNEs, GREAT tool (http://great.stanford.edu)[121] was used. Significantly enriched functional categories among Gene Ontology (GO) “biological process” and “molecular function” terms and HGNC Gene Families were identified based on a binomial test of genomic regions (Bonferroni-corrected P-value < 0.01).

Autopod CNE evolution

Transient transgenic analysis was performed using a coelacanth <1 kb element orthologous to the mouse Island I combined with a minimal promoter (HSP68) and LacZ. Injection and staining was performed by Cyagen Biosciences (Cyagen.com). Two transgenic embryos were obtained and displayed similar expression patterns.

For interspecies DNA sequence comparison, we retrieved genomic sequences of the human, chicken, frog, coelacanth, pufferfish, medaka, stickleback, and zebrafish Hoxd loci from the Ensembl database (last update 05/12). The extended sequence of the elephant shark Hoxd locus was determined by sequencing a BAC clone (#46A6; IMCB Eshark BAC library, GenBank accession number JX519116). Alignments were performed using the rVISTA program Shuffle-LAGAN[122-123] (window size 50 bp; homology threshold 70%).

Evidence for selection in the urea cycle during the evolution of tetrapods

Coding sequences were aligned using PRANK[124-125], which has an option to preserve codon reading frames that has been shown to outperform other methods for alignments of codon blocks[126]. We used a branch-site model in the HYPHY package[127], which estimates dN/dS (ω) values among different branches and among different sites (codons) across a multiple species sequence alignment. The approach avoids partitioning branches into "foreground" positive selection and "background" negative/neutral selection. Instead it uses a random effects likelihood framework in which ω can take one of three values along branches ($\omega^{-}_{b} \leq \omega^{N}_{b} \leq 1 \leq \omega^{+}_{b}$) to explore every branch-site combination. Sequential likelihood ratio testing is used to identify branches with amounts of episodic diversifying selection (final p-values are adjusted using Holm's multiple testing correction).

The coelacanth and placental evolution

Chick electroporations

A 222 bp chicken HA14E1 sequence encompassing bases 233,860 to 233,639 of GenBank AC163712 sequence was inserted into the pTK-EGFP vector (from Dr. Hisato Kondoh). This vector consists of an eGFP reporter gene with a minimal thymidine kinase promoter from Herpes Simplex virus[128]. This plasmid (along with a plasmid expressing nuclear mCherry ubiquitously) was microinjected *in ovo* into the neural tube of a chick embryo at HH4 and electroporated as per methods described in[129] Embryos were harvested at HH11 and imaged for expression of eGFP and mCherry using a Zeiss Axioscope2 Plus fluorescence microscope. Five independent experiments were performed and all five showed expression of GFP in the extraembryonic regions only (blood islands and developing vasculature).

Mouse BAC transgenic experiments

The coelacanth BAC that was used for constructing a mouse transgenic line was described in Smith et al. 2012[130]. Briefly, the BAC insert included bases 1 to 168,364 of GenBank FJ497005.1 sequence (*Latimeria menadoensis* HOX-A cluster, 319,360 bp). The BAC included *Hoxa14* and spanned *Evx1* to *Hoxa9*. The *Latimeria Hoxa14* gene has been supplanted with an RFP (DsRed) coding sequence that included a

requisite sequence for polyadenylation (Clontech). The entire BAC insert was moved into the pPACGFP vector (Amemiya, unpublished), which has a P1 origin of replication and a GFP gene with ubiquitous promoter[131]. The GFP gene was used as a marker to identify transgenic founders after microinjection into mouse embryos. This transgenic line is known to correctly splice and express two *Latimeria*-specific genes via RT-PCR[130]. The DsRed signal was weak and not readily detectable in embryos transgenic for the BAC construct and necessitated immunohistochemistry from paraffin sections of the dissected embryos.

Paraffin sections of day 8.5 (E8.5) mouse embryo (including placenta) were cut at 5 microns onto positively charged slides and heated at 59 C for approximately 30 minutes. Slides were then placed on the Leica "Bond" Robotic immunostainer for immunohistochemistry (IHC) staining. We used a rabbit anti RFP primary antibody (abcam 28664) and a Leica MicroSystems "Bond Polymer Refine Detection" kit, which is a biotin-free, polymeric horseradish (HRP) linker antibody conjugate system. After processing on the Bond immunostainer, slides were dehydrated through graded alcohols and coverslipped from xylene with Surgipath MicroMount. Images were taken on a Leica DM2500 microscope. Two independent E8.5 mouse embryos were examined along with two non-transgenic controls. Staining (Fig. 4b) was consistently observed in a subset of cells in the extraembryonic membrane (developing labyrinth) in transgenic embryos but not in non-transgenic controls.

Coelacanth lacks IgM

A *Latimeria menadoensis* BAC library[132] was screened initially using *Latimeria* and lungfish V_H and C_H probes[133-134]. Clones were validated as containing IgH hybridisation fragments and fingerprinted using an automated system[135-136]. Five clones were strategically selected and sequenced by the Joint Genome Institute using ABI 3730xl sequencers to roughly 10X coverage. Phred and Phrap were used for editing and assembly[137], and manual annotation was performed using Vector NTI software (Invitrogen). The *Latimeria* BAC library and a 100X coverage lambda genomic library (unpublished) were screened subsequently with several other V_H and C_H probes (including degenerate oligonucleotides against highly conserved transmembrane regions of Cμ) in order to identify any other IgH-containing clones that escaped initial detection. In addition, various PCR strategies were employed with *Latimeria* genomic DNA to amplify putative Cμ -containing fragments; none of these efforts were successful.

Supplementary Notes

Genome Assembly and Annotation

Supplementary Note 1 - Assembly accuracy

To assess the base quality in the assembly we compared the sequence in the assembly to the assembled RNA-Seq data. Briefly, the 15,763 Human-Coelacanth 1:1 orthologous genes from Ensembl, were BLAST-ed against the RNA-Seq transcripts. The best hit for each gene was selected and matches and mismatches counted. A mismatch rate of 1/357 bp was observed (0.28%). Any discrepancies not due to polymorphism could results from errors either in the assembly or the RNA-Seq transcriptome (Errors in the RNA-Seq appear more likely due to lower sequence coverage).

We next assessed the polymorphism rate in the assembly using k-mer counts based on the raw read. This resulted in an estimated SNP rate of 1/445 bp (0.23%). The maximal error rate is therefore 0.28-0.23%= 0.05%. Assuming an equal error rate between the assembly and the RNA-Seq transcripts the error rate in the assembly would be ~ 0.025%.

When ALLPATHS-LG was validated on mouse and human the error rate was estimated to 0.03% and 0.05%.

The total repeat content of the coelacanth is 25%, not an especially high value. However, one of the most common repeat elements is the > 8 kb *Lati-Harb* transposon which is highly conserved in the coelacanth genome and present at least once per 100 kb[130]. This element, because of its size and high sequence identity, likely causes problems in the assembly process, especially when using shorter reads.

Supplementary Note 2 - Long intergenic non-coding RNAs

Identification of 1,214 long intergenic non-coding RNAs

A total of 110,659 transcripts predicted by Cufflink were made non-redundant using Cuffcompare resulting in the identification of 41,288 intergenic loci of these 2,966 are multiexonic. After testing for coding potential a total of 31,731 single exon and 1,493 multiexon intergenic loci were deemed to be non-coding. Only intergenic non-coding multi-exonic loci were retained for further analysis.

In order to remove any loci that could map, on the same strand, to a currently unannotated protein coding genes or unannotated UTR in the coelacanth genome, we searched the coelacanth centred alignments with 9 other vertebrate species (human, mouse, dog, elephant, opossum, chicken, green anole, African clawed frog, and stickleback) for candidate lncRNAs in coelacanth that are aligned with protein-coding exons in at least one of the other species. Of the 680 loci that are found among the alignments (either partial or full length), we discarded 287 loci that overlap an annotated exon in at least one species, leading to a conservative set of 1,214 lncRNAs of which 71 overlap an intron in at least one

species, and 24 overlap a non coding transcript (lncRNA or processed transcript) in mouse and/ or human (Supplementary Dataset 2).

Of these 1,214 lncRNA, 719 (59.61%) have more than one transcript (average: 2.18 ± 0.1). These non-coding transcripts have an average size of 2114 ± 93.68 nt containing an average of 1.97 (± 0.06) exons. Like lncRNA identified in other species, the loci in coelacanth are significantly shorter, have less exons and are expressed at a significantly lower level than protein coding genes (Kruskal-Wallis test, P<2.2x10-16 in all comparisons).

LncRNAs in coelacanth are selectively constrained

We used the alignments to compute the nucleotide conservation scores across protein coding and lncRNA exons and introns as well as within intergenic sequences. We found that non-coding exons are significantly more conserved than intergenic sequences, coding and non-coding introns (Kolmogorov-Smirnov test, P<0.001 in all comparisons after Bonferroni correction). However as observed in all other organisms analyzed thus far, non-coding RNA exons appear to be under weaker selective constraints than are protein-coding exons (P<0.001 after Bonferroni correction).

Identification of positional equivalents in human, mouse and zebrafish

We searched the 1,618 protein coding genes in coelacanth that are flanked by a lncRNA for orthologous genes also found in the vicinity of a lncRNA in human, mouse or zebrafish. We collected a total of 520, 242 and 102 of such genes in human, mouse and zebrafish respectively. In every pairwise comparison, the set of genes with neighbouring lncRNA in coelacanth and ortholog in another species is significantly enriched for genes also flanked by a lncRNA in one of the three species (hypergeometric test, P<0.001 in human, mouse and zebrafish). Of these 520, 242 and 102 protein-coding genes, 176, 75 and 28 in human, mouse and zebrafish respectvely are flanked by a lncRNA transcribed on the same strand and with the same orientation relative to the protein-coding gene in coelacanth. We found a total of positionally equivalent 23 lncRNA loci in coelacanth, mouse and human.

We tested genes in human with orthologs in coelacanth and a lncRNA in their vicinity for any functional bias relative to all the genes with a 1:1 ortholog in coelacanth using Fatigo[138]. As expected from previous analyses, genes flanking lncRNA are enriched in genes involved in sequence-specific DNA binding transcription factor activity ($P=1.891x10^{-10}$). However we found no significant functional enrichment for 176 genes flanking positionally equivalent lncRNA between coelacanth and human when compared to human genes with a one-to-one ortholog in coelacanth and flanked by a lncRNA locus in human.

Additional evidence comes for the conservation of splice sites. About 5.9% of the splice sites in the lncRNA set (230 sites in 173 transcripts) are alignable to sequence in at least one of the other 9 vertebrate genomes included in the latimeria-centered MSA. Of these 20% exactly correspond to an annotated splice site in at least one of these species, providing direct evidence for the partial conservation of 33 lncRNAs.

Supplementary Note 3 - Genes duplicated in the coelacanth lineage

L. chalumnae-specific duplication events

A total of 336 *L. chalumnae* specific duplicate gene pairs (672 genes) supported by high bootstrap values (> 50%) were identified in 226 phylogenetic trees reconstructed from 1,762 OrthoMCL protein families. A total of 246 out of 674 duplicate genes (36%) are located in tandem along a contig separated by one to five unrelated genes.

GO Gene Enrichment Analysis

After correcting for multiple hypotheses testing, 15 unique GO terms were found to be statistically enriched ($P \leq 0.05$). The most overrepresented GO term is GO:0004930 (G-protein coupled receptor activity) that maps to 37 paralogous gene pairs from 6 OrthoMCL groups (P = 1.83E-47). G-protein coupled receptors (GPCRs) are a large, ancient family of integral transmembrane proteins found in all eukaryotic organisms which facilitate signal transduction by binding structurally diverse ligands such as photons, odorants, biogenic amines, peptides and glycoproteins. GPCRs mediate extracellular environmental signals by coupling via guanidine nucleotide-binding proteins (G-proteins) to various secondary pathways involving ion channels, adenyl cyclases and phosholipases[139-140]. The typical structure of GPCRs comprises of 7 transmembrane alpha-helices which assists in ligand binding by linking the extra-cellular N-terminus plasma membrane receptor to the intra-cellular C-terminus[141-142]. GPCRs comprise ~1-2% of the total gene complement of vertebrate genomes analysed, similar to the 1.5% (238 of 19,032 genes) of *L. chalumnae* genes annotated as having G-protein coupled receptor activity[143]. Of the 6 OrthoMCL GPCRs groups analysed, two appear to have undergone large scale duplications within (i) *L. chalumnae*[*] and *Homo sapiens* (OG29) and (ii) numerous *L. chalumnae*, *G. aculeatus* and *T. rubripes* paralogues with *L. chalumnae* paralogues segregating separately from the other two fish GPCRs (OG23).

L. chalumnae GPCRs within OG23 belong to the C family of GCPRs whose receptors are involved in extracellular calcium sensing, gamma-amino-butyric acid and metabotropic glutamate (inhibitors and excitatory neurotransmitters respectively). Specifically for OG23, the vomeronasal type-2 receptors are involved in pheromone sensing and may have a possible role in binding basic amino acids that function as fish odorants as demonstrated in goldfish[144].

InterPro Domain Gene Enrichment Analysis Results

After correcting for multiple hypothesis testing, 38 unique InterPro terms were statistically enriched ($P \leq 0.05$).

In *L. chalumnae*, 79 duplicate gene pairs map to the GPCR Rhodopsin superfamily InterPro group (IPR017452), the most statistically enriched interpro term (P = 4.68E-32). The rhodopsin class of GPCRs

[*] http://www.sanbi.ac.za/wp-content/uploads/coelacanth/tree_images/OG23.txt.fa.clustalw.phylip.root_phyml_tree.tree.png

http://www.sanbi.ac.za/wp-content/uploads/coelacanth/tree_images/OG29.txt.fa.clustalw.phylip.root_phyml_tree.tree.png

alone is the most highly represented protein family in mammals. It includes hormone, neurotransmitter and light receptors, all of which transduce extracellular signals through interaction with guanine nucleotide-binding (G) proteins. Although vertebrates from pufferfish to human share a similar gene inventory, recent analyses demonstrated that a whole genome duplication occurred before the divergence of teleosts and osteoglossomorphs more than 230-350 million years ago, whereas other ray-finned fish (actinopterygians) and all sarcopterygians (tetrapods and coelacanthiforms) experienced no such event. Several recent studies have estimated that only 8 to 15 % of WGD- derived duplicates were retained in *T. rubripes* and *T. nigrovodiris.* Semyonov et al[145] showed that over 65% of the pufferfish nGPCR genes consists of lineage-specific duplicates which is an unexpected result because vertebrates from teleosts to tetrapods share a similar gene inventory and only 8-15% of whole genome duplicates survive in pufferfish. This implies that a selection pressure different from that for the rest of the genome could have affected evolution of nGPCRs after the occurrence of WGD in teleosts. Among the five groups of visual pigments in vertebrates, the rhodopsin type 2 (RH2) group shows the largest number of gene duplication events. Molecular phylogeny of vertebrate RH2 opsins shows that the RH2 opsins in tetrapods are more closely related to those in coelacanth, euteleosts, and lamprey. Furthemore, the RH2 opsins in lampfish, medaka, tilapia, and those in zebrafish and goldfish are clustered in two separate groups[146].

Determining the closest living fish relative of the tetrapods

Supplementary Note 4 – Phylogenomic analyses

The phylogenomic dataset consisted of 22 vertebrate taxa and 100,583 amino acid positions (49,122 variable and 35,090 parsimony-informative) concatenated from 251 orthologous protein alignments (see Methods). It was analyzed using both maximum likelihood (ML) and Bayesian approaches under two site-homogeneous models [LG+F+Γ and GTR+Γ (general time reversible)] and two site-heterogeneous models (CAT+Γ and CAT+GTR+Γ). Site-homogeneous models assume that a single amino acid replacement matrix can adequately describe all alignment positions, whereas the site-heterogeneous models create categories by regrouping sites with similar profiles of stationary amino acid frequencies and estimates a single exchangeability matrix for all the categories (in the case of CAT+GTR+Γ).

It has been shown that site-homogeneous models are much more sensitive to LBA artefacts than site-heterogeneous models and that site-heterogeneous models usually have a much better fit given enough data[10-15]. To estimate the fit of each model to the data, we used cross-validation (see Methods). In agreement with these previous studies[10-15], the ranking of the models was as follows: CAT+GTR+Γ > CAT+Γ > GTR+Γ > LG+F+Γ. Moreover, Log-Likelihood differences were all significant (CAT+GTR versus LG: 3130 ± 102, CAT versus LG: 1947 ± 108 and GTR versus LG: 1637 ± 56).

In spite of their different fits, all three models recovered all nodes with maximal statistical support (100% BS for LG and GTR; PP=1 and a Jackknife value of 100% for CATGTR) but two: (1) relationships between the coelacanth, the lungfish and the tetrapods, and (2) relationships between the elephant, the

armadillo and the remaining placental mammals. This confirms that when the question at hand is easy to resolve, even poor fitting models are able to find the correct solution, whereas the sophistication of the model becomes important when the phylogenetic signal is weak and/or blurred by heterogeneous evolutionary processes[90,147].

When using the worst fitting model (LG+F+Γ), the inferred phylogeny weakly suggests that the lungfish and the coelacanth form a clade that is sister to the tetrapods (see Supplementary Figure 4), with a bootstrap support (BS) of 59%. However, this grouping of lobe-finned fish might result from a long-branch attraction artefact (LBA) due to the high evolutionary rates of ray-finned fish and tetrapods relative to the slow evolutionary rate of both the lungfish and the coelacanth. Instead, the correct solution would be paraphyletic lobe-finned fish with the lungfish as the sister of the tetrapods, a hypothesis that only receives a BS of 41% under this model.

Such an interpretation is strengthened by the observation that under a better-fitting model (GTR+Γ), this alternative hypothesis becomes the ML solution and receives a BS of 63% (vs. 37% for the former). Moreover, under the best fitting model (CAT+GTR+Γ), Bayesian inference also recovers this lungfish and tetrapod sistergroup relationship with a maximal posterior probability (PP=1) and a jackknife-estimated statistical support of 100% (Figure 1). Therefore, we conclude that the lungfish is indeed the closest living fish relative of the tetrapods and that the monophyly of lobe-finned fish obtained under the LG+F+Γ model results from an LBA artefact due to the high evolutionary rates of ray-finned fish and tetrapods that are not adequately handled by this poor fitting model.

To further study the potential effect of LBA, taxon sampling was manipulated to increase the LBA artefact (by excluding the slowly-evolving chondrichthyans, i.e. by using only the fast-evolving actinopterygians as an outgroup) or to decrease it (by excluding the fast-evolving actinopterygians). As shown in the table for this Note 4, the results are in perfect agreement with our hypothesis of an LBA artefact and with the fit of the models estimated by cross-validation. The site-heterogeneous models (CAT+Γ and CAT+GTR+Γ), which fit best the data, recover lungfish+tetrapods whatever the taxon sampling, albeit with a reduced support when the LBA artefact is exacerbated. In contrast, the site-homogeneous models (LG+F+Γ and GTR+Γ) recover this topology only when the outgroup is composed of slowly evolving species, and recover the erroneous lungfish+coelacanth group when the outgroup is composed of fast evolving species. This confirms that a long branch attraction between the fast-evolving actinopterygians and tetrapods disturbs phylogenetic inference and that the simultaneous use of a rich taxon sampling and of a realistic model of sequence evolution is necessary to obtain an accurate phylogenomic inference[90,147].

In the Bayesian tree computed under the CAT+GTR+Γ model, the only node that cannot be resolved with confidence lies within placental mammals and puts the elephant as the sister of the four other placentals, with a PP of 0.53 and a jackknife value of 53%. The two (lesser fitting) site-homogeneous models also favour the same solution (LG BS=61% and GTR BS=69%). However, it should be noted that our taxon sampling was not designed to address this issue.

For short internal branches, it is expected that, due to incomplete lineage sorting, single gene phylogenies are different from the species phylogeny. In such a case, the analysis of a concatenation might be misleading[148]. This is why[149] Shan and Gras (2010) used the coalescent model implemented in BEST[150] to study the relative position of coelacanth, lungfish and tetrapods. Unfortunately, these approaches relied on very simple site-homogeneous models that fit the data very poorly (see above). When studying ancient phylogenetic questions (here >400 MYa), our opinion is that it is better to use a more accurate model of sequence evolution (i.e., CAT+GTR+Γ) than to use an approach that accounts for incomplete lineage sorting but under a very simple model of sequence evolution. In the long run, the relative position of lungfish and coelacanth should be evaluated by phylogenetic software able to jointly handle these two important biological complexities.

Table for Supplementary Note 4: Bootstrap and posterior probability supporting the best scenario for each data set and statistical model combination.

Data set vs Model	LG+F+$\wp$	GTR+$\wp$	CAT+$\wp$	CATGTR+$\wp$
All species	P+L (BS=59)	**P+T (BS=63)**	**P+T (PP=1.0)**	**P+T (PP=1.0)**
- 3 teleosts	**P+T (BS=93)**	**P+T (BS=94)**	**P+T (PP=0.92)**	**P+T (PP=1.0)**
- 3 chondrichthyans	P+L (BS=75)	P+L (BS=88)	**P+T (PP=0.57)**	**P+T (PP=0.99)**

P=*Protopterus*; L=*Latimeria*; T=Tetrapoda; BS=Bootstrap Support; PP=Posterior Probability

How slowly evolving is the coelacanth?

Supplementary Note 5 – Transposable Elements

Transposable elements (TEs) constitute less than 25% of the coelacanth genome (Supplementary Table 7), which is somewhat lower compared to similar sized genomes, such as clawed frog or mammals that contain more than 35% of TEs. The repertoire of TEs, as assessed by the number of TE superfamilies, is much wider than is observed for birds and mammals but much less diverse than ray-finned fish (Supplementary Table 14). The most abundant classes are Long (LINE) and Short (SINE) Interspersed elements. The majority of active families are CR1 and L2 LINEs (2.9% and 1.3%), LatiHarb1 transposon (1.45%) and Deu SINE (1.8%) (Supplementary Tables 8-10). LTR retrotransposon diversity is poor, showing only the presence of Gypsy and DIRS retrotransposons. Epsilon retroviruses were identified using phylogenetic analyses. By phylogenetic analysis, they cluster with infectious retroviruses of two freshwater fishes, snakehead and walleye. A potentially active Miniature Inverted-repeat transposable element (MITE) was discovered. Its insertion was identified in three different polinton DNA transposon sequences. More than 66840 of the polinton copies contained the MITE.

Four waves of TE bursts were detected in the Coelacanth genome (Supplementary Figure 5). However, there is no indication of a recent burst of TE (0.01~0.05, Supplementary Figure 5), which indicates that most TEs were in a stable state during at least the last 10 million years.

The coelacanth genome provides key information about the ancestral TE repertoire of tetrapods. With respect to transposable elements the structure of the genome holds an intermediate position of the coelacanth between ray-finned fish and tetrapods. It contains fewer superfamilies than ray-finned fish but more than birds and mammals, indicating loss of certain TE types after divergence from ray-finned fish, both before and after the split between tetrapods and coelacanth. On the other hand, although teleost genomes have a higher diversity of TE families, these families show a much lower copy number compared to tetrapods. Latimeria with respect to this feature is more "tetrapod" like, because many of its TE families display high copy numbers.

Harbinger transposable element

An approximately 8 kb stretch of sequence containing a full length copy of the *Latimeria* Harbinger-1 transposon, LatiHarb1[130], was used as a Blast query to search the *L. chalumnae* genome assembly. Default Blast parameters with word_size set to 7 bp were used for the search to allow for expected regions of high variability within the transposon. The search generated approximately 103,000 alignments across 9591 individual scaffolds (combined length 2.74 Gb), out of the 22,818 total scaffolds in the assembly. A *de novo* python script was written to further refine these highly fragmented Blast hits. Matches to conserved regions of LatiHarb1 were identified by clustering short matches and subsequently ensuring that orientation and respective order were consistent. A naïve run of this algorithm with no filtering applied generated 59,000 putative matches for LatiHarb1 transposons, covering a total of 48 Mb of sequence with an average individual coverage of 818 bp; 1,965 of these putative matches were over 5 kb in length. Filtering these results with a minimum length threshold of 1 Kb dramatically reduced the total number of transposon matches to 12,000 but with only a marginally smaller total coverage of 42 Mb, indicating that only small, fragmented matches were being removed. The average length of retained transposon matches after filtering was 3,500 bp. Given that the final genome assembly is composed of ~212K contigs with a contig N50 of 12.6 Kb, it is reasonable to assume that many of the partial matches correspond to a real full length transposon matching the end of a contig. This generates a lower bound estimate for genomic coverage of 1.45% for this transposon.

Supplementary Note 6 - Active transposable elements

The active transposable elements (TE) in the coelacanth genome were detected by two methods; first, based on the sequence similarity of the identified TE copies in the genome with the consensus sequences in the TE library (0.00~2.5%[151]) and based on analysis of the RNA-seq data set.

Both methods show similar results, namely that the active TEs in the coelacanth genome belong to at least 14 TE super-families (Supplementary Table 10), which is consistent with previous results that fish genomes have a higher number of active TEs than mammalian genomes[152].

To detect active based on sequence similarity based on the consensus sequence, the coelacanth genome was first masked with a de novo constructed TE library (lch.lib-v3.fa). An in-house Perl script was developed to detect active TEs based on RepeatMasker outputs.

We developed a pipeline to detect active TEs from the assembled RNA-seq data set. Briefly, the assembled RNA-seq sequences were masked by RepeatMasker (version 3.3.0) with the de novo constructed TE library (lch.lib-v3.fa); the masked sequences with Smith-Waterman score 225 (suggested by RepeatMasker website) and with at least 80% of base pairs masked were categorized as potential copies of the transcribed TE. We excluded those exonized TEs by blasting the potential copies of the transcribed TE against the coelacanth protein sequences (evidence based gene model prediction, Ensembl 66). A contig consisting of fragments from the TEs and "normal" protein-coding genes was classified as a transcribed gene with an exon, which originated from a TE insertion/transposition event. We also excluded those contigs that contained fragments from different TE superfamilies that may have resulted in errors in the RNA-seq assembly process. To evaluate the expression profile of the TEs, we first mapped the raw reads by using RSEM [153], and the expected fragments per kilobase of transcripts per million fragments mapped (FPKM) [154] was calculated for each TE contig within each RNA-seq sample following the instructions of the Trinity[155] website.

Supplementary Note 7 – Large-scale synteny

The evolution of the coelacanth karyotype mirrors trends in morphological evolution, exhibiting both extensive conservation of ancestral features and significant changes that occurred specifically within the coelacanth lineage. When compared to other vertebrate lineages, most coelacanth scaffolds show conserved synteny with a single chromosome (Supplementary Figure 6), suggesting that many gene arrangements have been conserved from a common ancestor. A smaller fraction of coelacanth scaffolds (~15%) reveal physical linkages between homologs that are located on two different chromosomes in one or more tetrapod species. Several of these linked segments correspond to arrangements that were present in the ancestor of the sarcopterygians and have been conserved in the coelacanth genome (Supplementary Figure 6). The availability of the coelacanth genome assembly and strong conservation of synteny in the coelacanth lineages therefore permits the resolution of several key features of the ancestral sarcopterygian genome and further resolves subsequent alterations to this structure within tetrapod lineages. The coelacanth assembly provides the sensitivity necessary to detect fusions and other intrachromosomal rearrangements in the coelacanth lineage, and fissions in the other tetrapod lineages, but is less sensitive to other types of rearrangement

Other linkages appear to represent changes that occurred after the coelacanth lineage diverged from the ancestral lineage that gave rise to tetrapods. Interestingly, many of these changes correspond to fusion events involving regions that are homologous to chicken microchromosomes (fused to both micro- and micro- chromosomes). Microchromosomes are common among non-mammalian sarcopterygians (including coelacanth[156]) and various theories have been proposed as to the origin of these chromosomes[156]. Recent comparative studies have revealed that many chicken microchromsomes correspond to individual chromosomes that were present in the genome of the ancestral tetrapod lineage[157]. Patterns of microchromosomal fusion in coelacanth are consistent with these studies but

differ with respect to the specific chromosomes that are involved in fission events, revealing extensive parallelism in microchromosome fusion among the sarcopterygians. Thus, the microchromsome complement in coelacanth, basal amphibians and reptiles appear to represent subsets of a larger complement of ancestral microchromosomes, with other microchromosomes having experienced independent fusions in several lineages (including coelacanth). Overall, our analyses indicate that karyotypic evolution in the coelacanth lineage has proceeded in a manner that is very similar to that of nonmammalian tetrapods. Mammalian rates have been reported to be substantially higher[158].

Supplementary Note 8 - *L. chalumnae* vs. *L. menadoensis* transcriptome comparison

Supplementary Figure 7a shows the cumulative percentage of contribution to the transcriptome, ordering the transcripts from the most to the least expressed. The values were calculated based on the number of reads mapping on each transcriptome contig, in relation with the total number of reads mapping on the entire assembled transcriptome. Only intact paired-end read mappings were considered, allowing a minimum 95% identity percentage over a minimum of 75% of read length alignment. The *L. menadoensis* transcriptome generated by joining data from both the Trinity and CLC assemblies was used as a reference for the mapping of liver and testis reads, whereas a slightly improved version of the *L. chalumnae* muscle assembly (the improvement consisted mainly in the removal of redundant contigs) was used as a reference for the mapping of muscle reads.

The graph is “zoomed” over the 1,000 most expressed genes as, due to the high proportion of lowly expressed contigs, the curves quickly reach the asymptote, making the full graph poorly informative.

The figure clearly shows remarkable differences among the three tissues. Namely, muscle appears to invest most of its transcriptional activity in a limited set of genes, as 80% of the total paired-end reads map to just about 200 genes contigs. On the contrary, testis represents an opposite case of a tissue transcriptomically very rich, as in order to reach the 80% of the total expression, we have to take into consideration almost 3,000 genes. Liver is an intermediate case (about 700 genes contribute to 80% of the total transcription).

Therefore muscle, as expected, can be seen as a rather “transcriptomically poor” tissue, highly specialized in the intense expression of genes related to muscular structure and its contraction activity. In fact, not surprisingly, about 25% of the total expression is made up of actin and myosin alone. Nevertheless it has to be taken in consideration that the sequencing depth applied to the muscle was much higher than the one used in the other 2 tissues, so RNA-seq was still able to gather sequence information about a rather large set of genes expressed at very low levels in this tissue.

Liver is a highly specialized tissue as well, since the majority of transcripts showing the highest expression levels are related to liver-specific functions, but it still expresses quite a broad range of transcripts, according to the variety of synthetic and metabolic processes this tissue is involved in.

Testis definitely appear as the tissue, among the three, which gave the best overall contribution to the genome annotation due to the higher chances of obtaining transcripts with a high coverage, resulting increased chances of correct gene predictions. This is likely related to the active status of germ cells and to the high replication rate expected in testis.

Supplementary Figure 7b shows the top 1000 transcripts of each tissue (ordered by FPKM expression values), as the comparison of the entire transcript sets is highly dependent on the sequencing depth applied. Although there is a "core set" of genes expressed at high levels in all the tissues (172), which can be considered as housekeeping genes, the majority of the transcripts expressed at high levels appear to be related to tissue-specific functions. Obviously this is just a highly simplified vision of the comparison among the three transcriptomes, as the expression of most of the transcripts identified as "tissue specific" in the diagram can be still detected, although at lower levels, in the other two tissues.

The bottom line is that, despite the better coverage of the transcriptome offered by testis, there is a relevant number of genes subject to strict regulation and expected to be expressed almost exclusively in a specific tissue and therefore there is no doubt that the RNA-seq of each of the three tissues was very useful, offering a broader coverage of the genes expressed in coelacanth.

Coelacanth informing the vertebrate adaptation to land

Supplementary Note 9 – Genes involved in adaptation

Molecular chaperone machinery

Heat shock protein 70 (Hsp70) and heat shock protein 40 (Hsp40; also called DnaJ) play a major role in maintaining protein homeostasis under both normal and stress conditions[159]. Hsp70 is the prototypical chaperone, with a limited number of isoforms present in most cellular compartments[160] (13 in humans), where they facilitate the productive folding and assembly of their substrate proteins. On the other hand, the DnaJ family is highly diverse, occurring in multiple isoforms[160] (49 in humans), and providing specificity to their Hsp70 partners by regulating their chaperone activity and targeting certain protein substrates to them[161]. We previously isolated and characterized a coelacanth gene encoding an inducible form of Hsp70[162-164] (HSPA1A/Hsp72; Supplementary Table 13). This protein sequence was used to search the coelacanth genome, as well as the 13 human Hsp70 protein sequences. Coelacanth homologues for 10 of the 13 human Hsp70s were identified (Supplementary Table 13; Supplementary Figure 8), including the cytosolic inducible (HspA1A; HspA1B; and HspA1L) and constitutive isoforms (HspA8), the endoplasmic reticulum (ER) isoform (HspA5), the mitochondrial isoform (HspA9), and certain specialized isoforms (HspA12A; HspA12B; HspA13; and HspA14).

The J domain enables DnaJ proteins to interact with Hsp70 partner proteins, and has certain invariant features that make it a signature sequence for this family of proteins (e.g. invariant HPD motif). Using a consensus J domain sequence[165], the coelacanth genome was searched for genes encoding DnaJ homologues. Four type I/DnaJA, 16 type II/DnaJB and 20 type III/DnaJC isoforms were identified, making a total of 40 distinct protein members (Supplementary Table 13). The canonical DnaJ proteins (type I/DnaJA) that typically interact with the canonical Hsp70s in facilitating protein folding in the cytosol and

in mitochondria were identified. Six DnaJ proteins of the endoplasmic reticulum were identified (homologues of human ERdj1/DnaJC1, ERdj2/DnaJC23/Sec63, ERdj3DnaJB11, ERdj4/DnaJB9, ERdj5/DnaJC10, and DnaJC3/p58^{IPK}), suggesting that all the machinery was in place for translocation and folding into the ER, as well as retrograde translocation for degradation.

Hsp90 also occurs as a number of isoforms[160] (five isoforms in humans, including two in the cytosol), and is a highly abundant and essential molecular chaperone that regulates the conformational state of signal transduction proteins involved in fundamental cellular processes such as proliferation, differentiation, development, and the stress response[166]. There are over 300 different Hsp90 client proteins, consisting mainly of transcription factors and kinases, including certain oncogenic proteins (androgen/estrogen receptors and proto-oncogenic protein kinases) and prion proteins. Searching the coelacanth genome using human Hsp90 identified 5 distinct sequences which all appeared to be homologues of the inducible (Hsp90α) and constitutive (Hsp90β) cytosolic isoforms. The Hsp70/Hsp90 organizing protein (Hop) coordinates the functional cooperation between the Hsp70 and Hsp90 protein folding pathways, so as to ensure efficient delivery of client proteins from Hsp70 to Hsp90[167]. A coelacanth homologue of human Hop was identified. Overall, the findings suggest that the coelacanth has all the major features required for functionally integrated Hsp70 and Hsp90 protein folding pathways.

PAS genes

The Per-Arnt-Sim (PAS) protein domain is found throughout the eukaryotic and prokaryotic phyla[168]. PAS domains are often part of signal transduction proteins, capable of binding to small and chemically diverse ligands. A subset of PAS proteins also contain the basic-helix-loop-helix (bHLH) domain, which allows DNA-binding. Proteins in the bHLH-PAS protein family function in the sensing of internal or external stimuli, such as oxygen, xenobiotic chemicals, steroids, and light[169]. They are components of important pathways that allow adaptation of the organism to changes in the environment.

The coelacanth genome contains 28 members of the PAS gene family, as compared to 23 human members and 34 in zebrafish (Supplementary Table 14). In general, the number of coelacanth genes in each subfamily matches that of humans, where there is usually a single copy of each gene. The presence of additional copies in zebrafish and other teleosts likely reflects the genome duplication specific to ray-finned fishes that occurred after the divergence of lobe-finned fishes[170]. AHR1, AHRR, HIF1, SIM1, BMAL1, and PER1 are examples of genes for which the zebrafish has multiple copies. On the other hand, coelacanth has duplicates of AHR1, HIF2, CLOCK2, and NXF. Interestingly, in each case one member of the pair is more closely related to the human ortholog, while the other clusters with orthologs from other fish species or is basal to the human and fish orthologs (data not shown). AHR genes are of particular interest because of their roles in response to xenobiotics as well as in the immune system[169]. AHR1/AHR2 paralogs have arisen due to a tandem duplication that occurred prior to the divergence of ray-finned and lobe-finned fishes[171]. The two AHR1 genes in coelacanth arose independently of the duplicated AHR1 genes in teleosts. Most mammals have lost the AHR2 gene, whereas coelacanth and all other fish studied to date have retained both AHR1 and AHR2. In addition to AHR1 and AHR2, there is evidence for an AHR3 in some shark species[171]; it is not clear if the additional AHR in coelacanth (denoted as AHRx in Supplementary Table 14) is orthologous to this shark AHR or

represents yet another AHR form. Overall, the coelacanth gene diversity is similar to that in humans, with a few duplications that suggest some diversification in functions involving AHR, HIF, CLOCK, and NXF.

Sex determination and differentiation genes

In vertebrates many genes affecting sex development have been described so far. Sex development consists of two main processes: sex determination (committed by environmental or genetic factors) and sex differentiation. A complex network of molecular interactions implements these phenomena[33-35]. During evolution this scenario has been influenced by changes in gene sequence, type and number.

Thirty-three genes involved in sex determination and differentiation, twenty-five male-specific and eight female-specific, were identified and characterized in the two species of *Latimeria* from the *L. chalumnae* genome and the *L. menadoensis* transcriptome. Due to the close evolutionary relationships between the coelacanths, the integration of the two datasets let us obtain a better definition of both gene and transcript structures.

The comparison between sex developmental gene sequences in the two coelacanths confirmed their close relationship showing a maximum divergence percentage of 2.05%. Phylogenetic analyses or comparisons of the conserved syntenic blocks in vertebrates were carried out on genes that, according to literature, play a crucial role in this pathway. In particular, *DMRT1* (Doublesex and Mab 3-Related Transcription factor 1), a key gene in male sex determination and in testis differentiation, showed a conserved synteny in *Latimeria*, as Brunner and colleagues[36] have stated for other vertebrates.

Another important gene, *Sox9*, a transcription factor belonging to the SoxE subfamily, was further investigated. During male sex differentiation, this gene activates the pathway for the production of the Müllerian Inhibiting substance (AMH)[37]. The evolutionary relationships of the three genes composing the SoxE group (Sox8, Sox9, Sox10) were evaluated through Bayesian Inference and Maximum Parsimony. *FGF9* (Fibroblast Growth Factor 9), a gene involved in tetrapods male sex development missing in teleost species, was found in the *L. chalumnae* genome. The syntenic arrangements of the genes surrounding of the FGF9/16/20 subfamily were compared. The transcriptome analysis in the liver and testis of the adult male specimen of *L. menadoensis*, revealed a preferential expression of several male determining genes in testis.

In conclusion, this scenario in *Latimeria* confirms the pathways described across vertebrate species. These findings provide an important contribution in order to understand the evolution of sex determination and differentiation genes.

Identification of tetrapod and sarcopterygian specific genes

The goal of this search was the identification of tetrapod genes that lack any non-sarcopterygian homologs. In order to identify these tetrapod- and sarcopterygian specific genes *in silico* a dataset including available peptide sequences of tetrapods (human, mouse, dog, platypus, opossum, zebra finch, chicken, turkey, anole and *Xenopus*) was downloaded from the Ensembl genome database [version 64;

[172]]. Local blastp searches [83] using all downloaded tetrapod peptides against non-tetrapod peptides (Ensembl peptides of *Ciona intestinalis* and *C. savignyi*, *Caenorhabditis elegans*, fruitfly, yeast, lamprey, zebrafish, medaka, stickleback, Fugu and Tetraodon, and all available NCBI peptides of chondrichthyes and *Branchiostoma floridae*) were performed. Tetrapod peptides that did not produce a significant hit (a cut-off of 50 was applied to bit score values) were further processed. First, the redundancy created by multiple tetrapod orthologs was reduced by collapsing gene families to one human peptide. Second, splicing variants belonging to a single gene were removed and only the longest peptide was retained in the dataset.

This pipeline also detects genes that are restricted to a small number of tetrapod taxa, but this group of genes was not target of our analysis. Therefore, the tetrapod peptides produced by the pipeline described above were divided into two groups. The amphibian peptides were used as queries in blastp searches against non-amphibian peptides (extracted from the initial dataset) and non-amphibian peptides were used as queries in blastp searches against amphibians (all available peptides downloaded from Ensembl and NCBI). Another blastp search was conducted in order to reduce the number of false positives, which have distant homologs in invertebrates. We used the peptides producing significant hits (based on a bit score cut-off value of 50) in blast searches against amphibian vs. non-amphibian as queries. The target of these search were NCBI peptide sequences of several invertebrates (*Apis melifera, Bombyx mori, Hydra, Nematostella vectensis, Schistosoma mansoni* and *Strongylocentrotus purpuratus*). A final search was performed in order to identify possible homologs of tetrapod-specific genes in lungfish or coelacanth. Peptides identified in the pipeline described above were used as queries in tblastn searches against the lungfish (*Protopterus annectans*) transcriptome and the coelacanth (*Latimeria chalumnae*) genome. For every identified peptide, a careful phylogenetic analysis was conducted in order to reconstruct the phylogenetic distribution. Supplementary Table 15 summarizes the tetrapod- and sarcopterygian-specific genes identified in this bioinformatic pipeline.

Identification of tetrapod and sarcopterygian specific gene loss

In this analysis we aimed to identify genes which are present at least in one non-osteichthyan taxon and teleosts, but absent from tetrapod genomes. For this purpose, all available peptide sequences of teleosts (zebrafish, medaka, stickleback, Fugu and Tetraodon) were downloaded from the Ensembl genome database [version 64; [172]]. Blastp searches were performed using teleost peptide sequences as queries against tetrapod peptides downloaded from Ensembl (human, mouse, dog, platypus, opossum, zebra finch, chicken, turkey, anole and *Xenopus*). All peptides producing a bit score lower than 50 in this search were further analyzed as described above: First, the redundancy created by multiple teleost orthologs was removed by collapsing all orthologs to one zebrafish peptide. Second, if multiple splicing variants of a single gene exist, only the longest peptide was retained in the dataset. In order to remove teleost-specific genes from this analysis, a blastp search of the candidate peptides against non-osteichthyan peptides (Ensembl peptides of *Ciona intestinalis* and *C. savignyi*, *Caenorhabditis elegans*, fruitfly, yeast, lamprey and all available NCBI peptides of chondrichthyes and *Branchiostoma floridae*) was conducted. Teleost peptides which did not produce a significant hit (based on a bit score cut-off value of 50) were discarded. Phylogenetic trees of the remaining teleost peptides were manually constructed and the presences of homologs in the coelacanth genome and the lungfish transcriptome

were investigated by running tblastn searches. Supplementary Table 16 summarizes the tetrapod- and sarcopterygian-specific gene losses identified in this bioinformatic pipeline.

Supplementary Note 10 - Tetrapod and sarcopterygian CNEs

The unique phylogenetic position of coelacanth between the ray-finned fishes and tetrapods allows the prediction of gene regulatory elements that evolved during the evolutionary transition from lobe-finned fishes to tetrapods, and that might be associated with the morphological and physiological novelties of tetrapods. Noncoding sequences that are under evolutionary constraint are strong candidates for gene regulatory elements and can be computationally predicted by comparing related genome sequences. Functional assay of computationally predicted conserved noncoding elements (CNEs) have shown that at least 50% of them act as enhancers directing tissue-specific expression at various stages of embryonic development[120,173-174], while some function as repressors and insulators[175-177]. To identify putative gene regulatory elements that originated in the most recent common ancestor of tetrapods, we first predicted CNEs that evolved in various bony vertebrate lineages and assigned them to different branch points based on the most recent ancestors in which they were brought under evolutionary constraint. We then validated the functional significance of CNEs by checking their overlap with experimentally identified gene regulatory regions.

To identify CNEs that originated in various bony vertebrate lineages, we generated pairwise alignments (LASTZ) and multiple alignments (MULTIZ) of human, mouse, dog, elephant, opossum, chicken, lizard, frog, coelacanth and stickleback genomes with human as the reference. The conserved sequences in the human genome were then predicted using PhastCons (see Methods). Protein-coding sequences, UTRs, and RNA genes were excluded from these sequences to identify the remaining elements as conserved noncoding elements (CNEs). We then focused on 739,646 CNEs that are ≥30 bp. These CNEs are on average 110 bp long and cover 2.7% of the human genome. A human CNE that showed at least 30% overlap with an orthologous sequence in another genome was deemed to be present in that genome and the most recent common ancestor in which a CNE was found to be under statistically significant constraint (*P*-value <0.01) was inferred as the branch point of origin of that CNE (Supplementary Table 17). This analysis identified 44,200 CNEs (average length 139 bp) that originated in the lineage leading to tetrapods after the divergence of the coelacanth lineage. They represent 6% of CNEs that are under constraint in the bony vertebrate lineage (Supplementary Table 17). Our analysis also identified 53,985 'sarcopterygian CNEs' (average length 151 bp) that evolved in the most recent common ancestor of sarcopterygians. Since we are using a teleost fish (stickleback) as the basis for comparison, this number is likely to be an overestimate, because previous studies have shown that a significant proportion of ancient gnathostome CNEs has diverged beyond recognition in teleost fishes[118-119].

To demonstrate the biological significance of tetrapod and sarcopterygian CNEs, we examined the overlap between these CNEs with experimentally identified enhancers based on ChIP-seq analysis of p300 binding in the forebrain, midbrain and limb of E11.5 mouse embryos[120]. For comparison, similar overlap analysis was carried out using randomly selected genomic regions of similar numbers and sizes. There is a 7.0-fold enrichment for p300 binding sites in tetrapod CNEs ($P < 2.73 \times 10^{-217}$), and an 11.5-

fold enrichment in sarcopterygian CNEs ($P < 8 \times 10^{-322}$) than in random genomic regions. These analyses indicate that the tetrapod and sarcopterygian CNEs are enriched for gene regulatory elements.

To verify the functional categories of genes associated with tetrapod and sarcopterygian CNEs, we assigned each CNE to its closest gene in the human genome and determined the Gene Ontology (GO) terms associated with that gene and the HGNC gene family to which it belongs. We tested for enrichment of GO terms and gene families using the GREAT enrichment tool (http://great.stanford.edu) based on a binominal test of genomic regions. Interestingly, the tetrapod CNEs were found to be most enriched in genomic regions containing genes involved in the perception of smell ("olfactory receptor activity", "sensory perception of smell", "OR" gene family; all with *P* value = ~0) (Supplementary Tables 18-20). This finding indicates that a major regulatory innovation occurred in olfactory receptor genes during the origin of tetrapods, and is consistent with the observation that there has been a significant expansion in the family of OR genes, particularly the α and γ types, in tetrapods compared to aquatic vertebrates such as teleost fishes[178-179]. Additional evidence for this hypothesis is that there are only 41 OR genes in the coelacanth assembly. Even though many genes are likely to have been missed in this assembly, due to the underrepresentation of gene families such as the ORs in draft assemblies, a comparison with the draft assembly of *Xenopus tropicalis* (824 genes) and the finished human genome assembly (387 genes)[179] is revealing. This may reflect the necessity of a more tightly regulated, larger and diverse repertoire of ORs for detecting airborne odorants as part of a terrestrial lifestyle of tetrapods. Besides olfactory receptor genes, enrichment was also seen in genomic regions containing transcription factors and developmental genes controlling morphogenesis (e.g., "hindlimb morphogenesis" $P < 1.7 \times 10^{-42}$), and cell differentiation (e.g., "lymphatic endothelial cell differentiation $P < 1.5 \times 10^{-41}$). An interesting instance of enrichment outside the transcription factor-developmental ('trans-dev') category of gene loci is that encoding immunoglobulin ("V(D)J recombination" $P < 1.2 \times 10^{-14}$; "DNAJ" gene family $P < 2.3 \times 10^{-20}$). This indicates that novel regulatory networks involving adaptive immune system genes were invented during the origin of tetrapods.

In contrast to tetrapod CNEs, the sarcopterygian CNEs are predominantly enriched in genomic regions containing transcription factors and developmental genes (Supplementary Tables 21-23). In fact, there is considerable overlap in the "trans-dev" category of genes that show enrichment of sarcopterygian and tetrapod CNEs. For example, genes belonging to HGNC families ZFHX, IRX, TALE, HOXL, ZFHX, PAX, PRD and CUT are found in regions enriched with sarcopterygian as well as tetrapod CNEs (Supplementary Tables 20 and 23). Interestingly, many gene loci have served as common targets for innovation of tetrapod and sarcopterygian CNEs. The following genes are found in the list of top 10 genes with the highest number of both tetrapod and sarcopterygian CNEs: *LPHN2*, *ZEB2*, *ODZ3* and *ROBO2* (Supplementary Table 24). These overlapping genomic regions between tetrapod and sarcopterygian innovations suggest that the same genes have been recruited for sarcopterygian and tetrapod innovations through the build-up of a more complex regulatory network. These CNEs are strong candidates for experimentally investigating the genetic basis of morphological innovations at the origin of sarcopterygians and tetrapods.

Supplementary Note 11 - Actinodin genes and a possible loss of the superficial muscles along with the loss of dermal fin rays in the fin to limb transition

Zhang et al.[180] proposed that the loss of fish actinotrichia proteins called actinodin (*and*) from the fin actinotrichia might have contributed to the evolution of limbs from fins. They presented the expression patterns of actinodin1-4 in zebrafish and compared and analyzed the domains of amino acids among representative fishes. Based on their comparisons, they could not find *and* orthologues in tetrapod lineages. The present study has extended their analysis by using *in silico* methods to investigate the presence of *and1* and its syntenic region including the genes *adipoq* (adiponectin), *myeov2* (myeloma overexpressed2) and *otos* (otospiralin), in tetrapod lineages (Supplementary Figure 12). Zebrafish has four *and* genes - *and1, and2*, *and3* and *and4*. The latter three genes appear to be products of the teleost specific whole genome duplications. As Zhang et al.[180] demonstrated, functional *and1* is not present in the *and* syntenic region of *Xenopus tropicalis* (Supplementary Figure 12). Furthermore, when the *and1* coding region in zebrafish and the potential *and1* region in *Xenopus* deduced from mVista comparisons (http://genome.lbl.gov/vista/index.shtml) of their *and1* syntenic regions were compared, there were no detectable functional exons comparable to those of *and1* in zebrafish; moreover, intervening sequences were dispersed within the *Xenopus* sequence when aligned with each exon of *and1* in zebrafish (data not shown).

When the investigation is extended to other tetrapods, *and1* was not found in the syntenic regions. In addition, the syntenic regions appear to have been re-organized, and a part of the segment was located in a different chromosomal region of each amniote species examined. Therefore, it is my proposal that (1) *and1* has functionally been lost in tetrapod lineages, (2) the syntenic region had further been re-organized in amniote lineages and (3) the loss of actinodin, and thus the loss of dermal fin rays, might be indicative of a possible event where the superficial abductor and adductor muscles[181] had also been lost in stem tetrapods when dermal fin rays were replaced with digits[182] due to the loss of connectivity between these muscles and dermal fin rays[183]. The superficial muscles, including the abductors and the adductor muscles of the pectoral fin in *Latimeria chalumnae* originate from the shoulder girdle and insert onto the dermal fin rays[181], whereas the superficial muscles are not present in extant amphibians[184-186]. Therefore, it is a speculation that the superficial muscles might have been lost in the fin to limb transition and that the stem tetrapods might have evolved re-organized muscular patterns of limbs similar to those of living amphibians. The fossil *Tiktaalik* had its reduced form of dermal fin rays[187] and thus its superficial muscles might have been a transient state toward their disappearance when the fin to limb transition had progressed toward landing.

Coelacanth lacks IgM

Supplementary Note 12 - Immune genes - summary

Genes encoding molecules of the adaptive and innate arms of the immune system were analyzed from the genomic and transcriptomic databases of *Latimeria* and will be described elsewhere. Briefly, the coelacanth immunome contains large numbers of immune receptors of the immunoglobulin superfamily, including immunoglobulins, T-cell receptors, major histocompatibility complex, and typical collections of known innate immune receptors, differentiation antigens and immune regulatory molecules as well as an additional large multigene family that is difficult to place. Phylogenetic analysis

of several of these genes place coelacanth in positions more closely related to tetrapods than to bony fishes, a general trend seen for many *Latimeria* genes.

Immune genes – complement systems

Genes homologous to the ones involved in mammalian complement cascades were detected in the coelacanth transcriptome and genome data, which included ones of the complement components (e.g., C3, C4, C5, C6, C7, C8[A,B,G], C9), proteases (e.g., C1r, C1s, C2/Factor B, Factor I, MASP1, MASP2), pattern recognition molecules (e.g., C1q[A,B,C], pentraxins, properdin, ficolins), regulators (e.g. C1-Inh, CLU, VTN, CPN, CD59) and receptors(e.g., CD11, CD18. C3aR, C5aR/C5L2, C1QBP, CD93). Presence of most components of complement cascades indicates essential role of complement in the innate immune system of coelacanth. Nonetheless, some genes, such as of factor D or mannose-binding lectin (MBL) have not been detected yet, which may represent either the absence of gene in the current data so far revealed or in the coelacanth genome. Factor D is one of critical components in the alternative pathway of complement in mammals, whose absence impair the complement based immune system. Further investigation is required to examine if factor D is absent in the coelacanth genome.

In human, the genomic region called RCA complex, contains genes of regulators or receptors of complement systems. These includes C4BPβ, C4BPα, DAF(CD55), CR2, CR1, CR1L, and MCP(CD46), all of which contains so called CCP (complement control protein) module. The structure of these molecules varies extensively, where C4BPβ contains 4 CCP modules, C4BPα contains 8 CCP modules, DAF is a glycosyl-phosphatidylinositol (GPI)-anchored membrane protein with 4 CCP modules, CR2 is a transmembrane proteins with 15 CCP modules, CR1L is a proteins with 8 CCP modules, CR1 is a transmembrane protein with 30 CCP modules and, and MCP is a transmembrane protein with 4 CCP modules. The syntenic genomic region of RCA complex bound by CD34 and PFKFB2 genes in coelacanth is present and contains a gene coding transmembrane protein with 4 CCP modules and a pseudogene whose transcripts contain a CCP module attached to fragment of unrelated protein. Factor H and its related molecules also contain CCP modules and, in the search of coelacanth transcriptomes, only partial transcripts homologous to human factor H/coagulation factor XIII B chain was found. Overall, the coelacanth genome contains the fundamentals of complement cascades but the molecules involved its regulation differ to some extent from those of mammals.

Immune genes – receptors

The organization and somatic mechanisms of reorganization of immunoglobulin genes varies markedly throughout the radiations of jawed vertebrates and can serve as an informative general characteristic of systematic relatedness.

An early study in Coelacanth[188] and subsequent work (Amemiya, personal communication) suggests that the immunoglobulin heavy chain gene locus is organized in a manner that shares features with that of both cartilaginous fish and higher vertebrates.

Our efforts with the Coelacanth genome have focused on identifying several large multigene families that we previously described in teleost fish. Our initial question was focused on the NITRs[189-192], which

function in a manner similar to natural killer receptors in higher vertebrates. We have extensive knowledge of these diversified genes in multiple species of teleost fish as well as in gar and have very effective query strategies for identifying homologs. We have concluded that the NITRs are not present in Coelacanth.

We did identify (only) two homologs of another large multigene family of immune-type genes (the DICPs), which are found in teleost fish. The systematic distribution of DICPs in other vertebrate lineages is not known and we cannot comment further on their significance at this point.

We were able to detect large numbers of MDIRs in Coelacanth, representing a third large family of immune-type receptors that are seen in cartilaginous fish as well as in other lineages of jawed vertebrates. MDIRs may relate to other families of immune receptors in mammals (specifically CD300, TREM, TIM and pIgR).

Receptors that bind immune molecules on the surface of cells as well as molecules that are related to them are central elements in physiologically complex networks that mediate immunity; emerging data from multiple vertebrate species point to the apparent expansion of various large IgSF multigene families that serve as part of a vast regulatory network that regulates the development of immune responses based on environmental or cellular context. We feel that the presence or absence of specific variants of these multigene families may be more informative than the effector molecules themselves as to broad patterns of phylogenetic relatedness.

In Coelacanth, we have identified large families of genes encoding such molecules. The genes to which the highest degree of relatedness is seen bind (or are related to those that bind) immunoglobulin on the cell surface in a cell lineage-specific manner as well as perform other receptor functions such as interaction with MHC Class II[193-194]. Furthermore, other receptors and ligands are present in Coelacanth that exhibit strong relatedness to those seen in amphibians.

We conclude that in terms of immunoglobulin receptors as well as several other types of immune receptors, the genes identified in Coelacanth exhibit a higher affinity with amphibian genes than with genes identified in other vertebrates.

Immune genes – toll like receptors

Searching for TLR genes and those of their downstream signaling pathways identified diversified members of TLRs and most of the components of signaling pathways among the genome of African coelacanth and the transcriptome of the Indonesian coelacanth. Namely, TLR1/6/10, TLR2, TLR3, TLR5, TLR7, TLR9, TLR13, TLR14/TLR18, TLR15, TLR21, and TLR22 homologs are shown to be present. Nonetheless, only a pseudogene was found for TLR8, whereas homologs for TLR4, MD2, CD14, and TICAM2(TRAM) were not detected by various homology search strategies. The absence of these components implies overt differences in TLR dependent innate immunity. Specifically, in higher vertebrates, TLR4, CD14, and MD2 are known to play an imperative role in the recognition of lipopolysaccharide (LPS, a potent bacterial immunogen), and TICAM2 (TRAM) is used for the activation of TRIF-dependent NF-kB and IRF3/IRF7 through TLR4. For LBP, another important component for the

recognition of LPS, only a primordial-type LBP/BPI homolog was detected in the coelacanths, as seen in teleost fishes, suggesting that the diversification of LBP and BPI occurred after the coelacanths diverged from the higher vertebrates. Collectively, these observations suggest that the molecular recognition of LPS by TLR4 and its associated signaling cascade emerged in the lineage leading to tetrapods after the evolutionary divergence of the coelacanth.

Discussion

Supplementary Note 13 – Gene families with strong directional selection

Globins

The coelacanth genome provides a unique opportunity to investigate the diversification of gene families at the base of the tetrapod radiation. The globins are a classical model system for studying the function and evolution of genes and proteins. These small heme-proteins may reversibly bind atmospheric O_2 and thus are an interface between the organism and its environment. Hemoglobin (Hb) and myoglobin (Mb) are the best known globins, but recently six additional vertebrate globins have been identified: Neuroglobin (Ngb)[195-196] support the survival of neurons, cytoglobin (Cygb) is mainly expressed in fibroblast-related cells [197], globin E (GbE) has only been found in the retina of birds [198], globin Y (GbY) exhibits a broad expression pattern in *Xenopus* tissues [199], globin X (GbX) is expressed in parts of the CNS and has been discovered in lampreys, sharks, bony fishes and amphibians [199-200], and androglobin (Adgb) is a chimeric protein with an internal globin domain [201]. While Hb, Mb, Ngb, Cygb and Adgb may be present in nearly all vertebrate species [201-202], the other globins are restricted to certain taxa. While some may have an O_2 supply function similar to Hb and Mb, other globins may carry out a variety of different functions, such as detoxification of NO or reactive oxygen species (ROS) [203].

The coelacanth is the only known vertebrate that includes all eight globin types. Thus the coelacanth can be considered as a globin "fossil", providing the toolbox for tetropod globin evolution. This suggests an early divergence of distinct globin types in the vertebrate evolution before the emergence of tetrapods (Supplementary Figure 22). Furthermore, the presence of *GbE* in *L. chalumnae* demonstrate a multiple independent loss of this gene in teleost fish, amphibians, reptiles and mammals, while this gene has been retained in the coelacanth and in birds. Likewise, Mb, GbX and GbY have been lost in some lineages; in humans and other placental mammals, only Hb, Mb, Ngb, Cygb and Adgb survived.

Multiple sequence alignment and phylogenetic inference

A multiple sequence alignment of the amino acid sequences of vertebrate globins was constructed employing MAFFT 6 [204] with the G-INS-i routine and the BLOSUM 45 matrix at http://mafft.cbrc.jp/alignment/server/. Bayesian phylogenetic analysis was carried out with MrBayes 3.1.2 [205] assuming the WAG model [206]. A gamma distribution of substitution rates was assumed. Metropolis-coupled Markov chain Monte Carlo (MCMCMC) sampling was performed with one cold and three heated chains. Two independent runs were performed in parallel for 10,000,000 generations. Starting trees were random and the trees were sampled every 1000th generation. Posterior probabilities were estimated on the final 40,000 trees (burnin = 10,000).

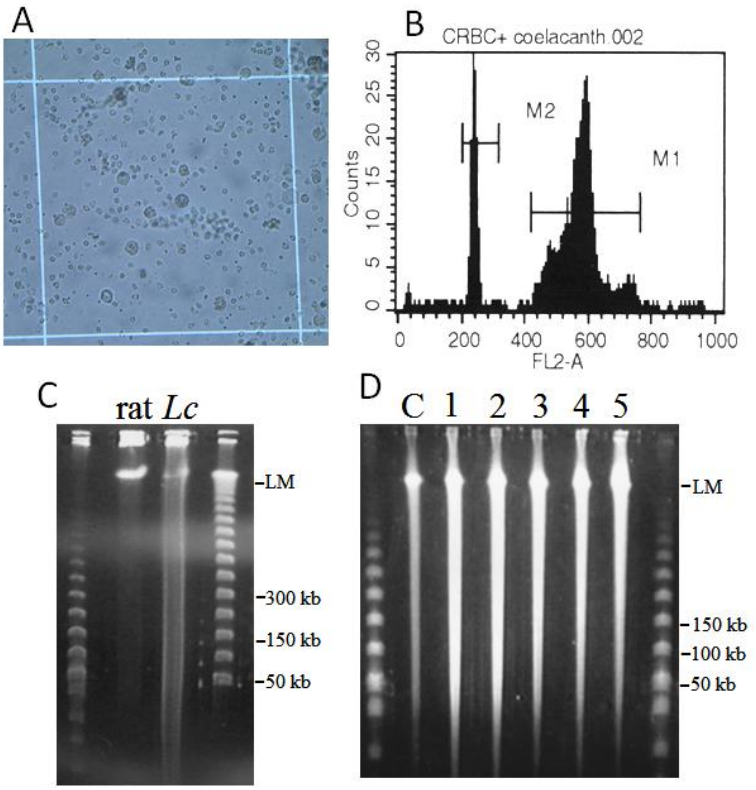
Cytochrome P450 gene superfamily and the defensome

The cytochrome P450 (CYP) gene superfamily is among the most complex enzyme superfamilies. CYPs are involved in the oxidative transformation of endogenous regulatory molecules (steroids, retinoids, fatty acids) governing processes in development and reproduction. CYP involvement in metabolizing foreign chemicals in the diet or environment can determine health outcomes and disease from chemicals in the embryo and adult. *Latimeria chalumnae* has 55 CYP genes, distributed in 18 families as do other vertebrates, and in 33 subfamilies. *Latimeria* also has multiple genes in other gene families that function in chemical-biological interactions; the glutathione transferases (11 genes), sulfotransferases (43 genes), glucuronyl transferases (n genes), aldo-keto reductases (13 genes), and ABC transporters (52 genes). These gene families together with the CYPs comprise the chemical defensome[207], which can determine both adaptation to the environment, and adverse effects resulting from chemical exposures.

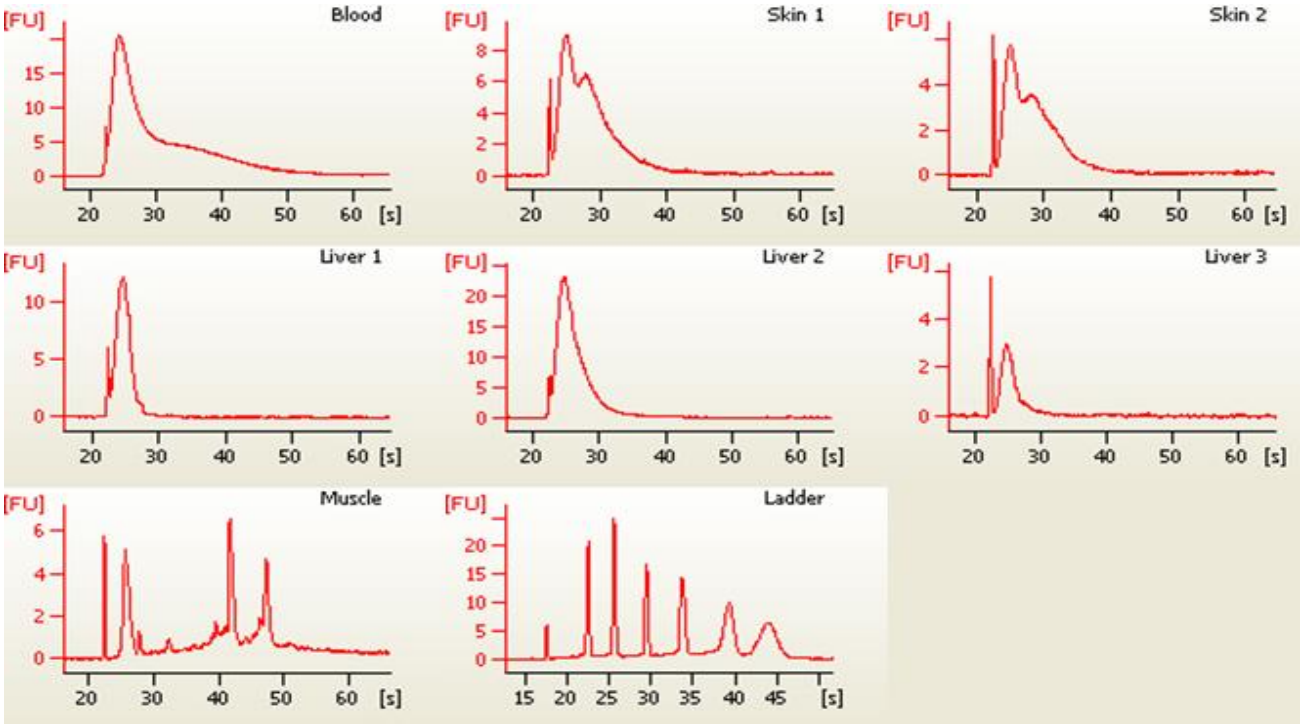
Gene families for CYPs with endogenous functions have members that are recognizable orthologues or co-orthologues to those in other vertebrates. The coelacanth CYP1 family, typically involved in response to hydrocarbons and other toxic aryl hydrocarbon receptor agonists, has four subfamilies, CYP1A, CYP1B, CYP1C and CYP1D, found also in the ray-finned fishes. In coelacanth, there is a single gene in each of these subfamilies, with gene structure like orthologous genes in other fishes, including the single exon CYP1C gene. Many teleosts have adjacent single exon paralogous genes in the CYP1C subfamily, resulting from tandem duplication rather than whole genome duplication.

A surprising feature of the coelacanth CYP complement is the relative dearth of CYP2 family genes. Coelacanths show only five CYP2 gene subfamilies with (at present) 14 genes, compared to 2 to 5 times as many (or more) in ray-finned fishes[208], mammals[209], other tetrapods or even early deuterostomes such as sea urchin[207], which has 73 CYP2 or CYP2-like genes. The reason for fewer CYP2s is unknown, but could involve differences in physiology, or evolution in habitats relatively free of chemicals which might induce or be metabolized by CYP2 enzymes. It is difficult to distinguish between low gene numbers as a preserved ancestral condition or a reduction caused by gene losses. Comparison to a complete shark genome may answer this question. The small number of genes and differences in "blooms" in the CYP2s suggests that coelacanth CYP2s may hold important clues to conserved endogenous functions and toxicological relevance of this usually highly diverse and often-confusing drug-metabolizing CYP family.

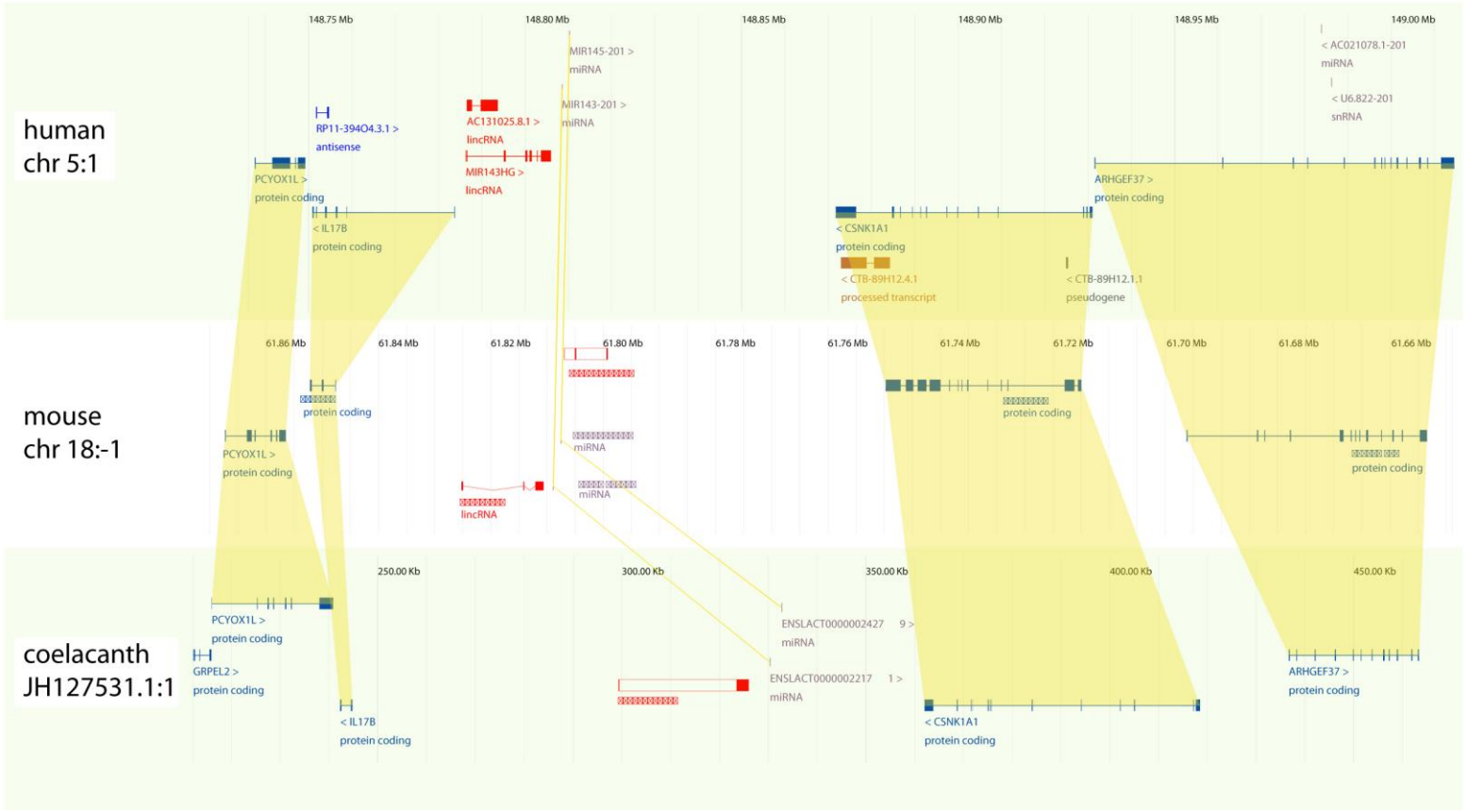
Supplementary Figures



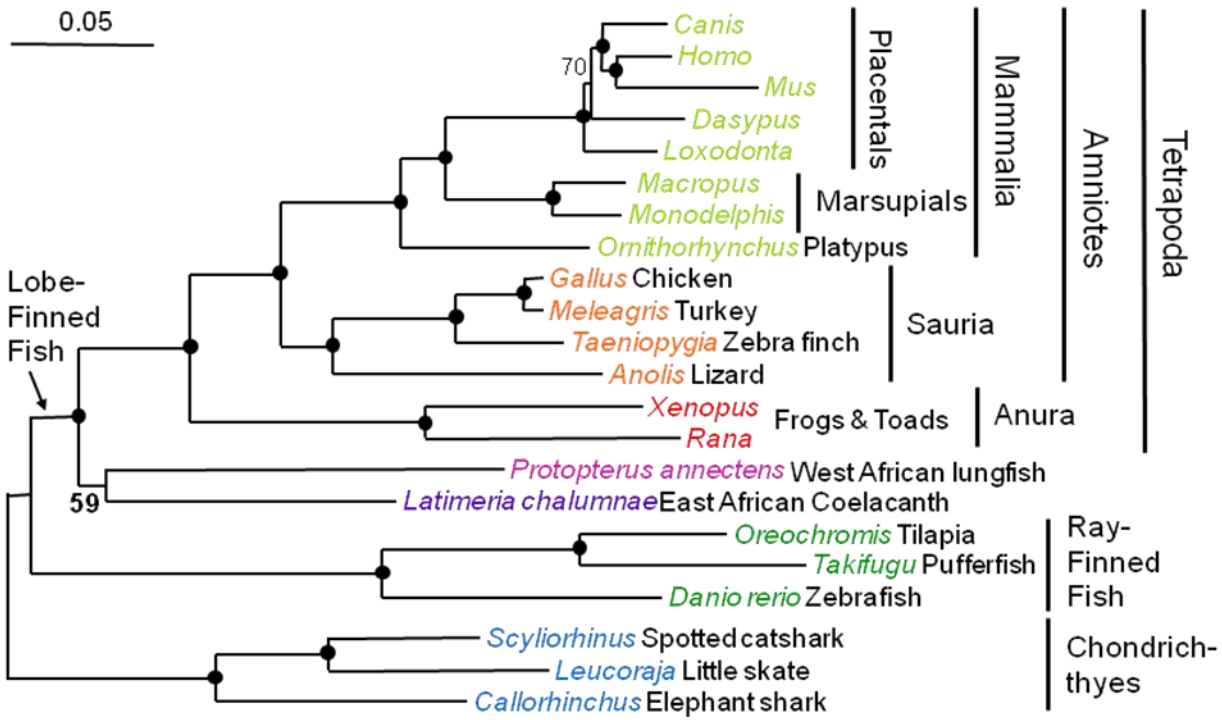
Supplementary Figure 1. Assessment of the quality of preserved *Latimeria chalumnae* blood and its genomic DNA. A blood sample from a Comoran specimen was analyzed by microscopy (A) and flow cytometry (B), and was embedded in agarose for preparation of high molecular weight DNA (C,D). (A) Microscopic phase-contrast examination showing many intact RBCs. (B) Flow cytometric analysis of blood sample. Coelacanth cells were washed in PBS (with 50 mM EDTA), combined with chicken red blood cell nuclei, stained in propidium iodide, and analyzed by flow cytometry. Several thousand cells were analyzed in four separate experimental runs. The left peak represents the chicken red blood cells (2.33 pg per 2C nucleus) and the right peak represents the coelacanth sample (primarily erythrocytes). Number of events counted is given on the left (x 1000). Based on these results we can conclude that the cells that we received from the coelacanth were of good quality (i.e., not overly hemolysed) and that the estimated genome size is ~ 2.75 pg/C. (C,D) Analysis of coelacanth genomic DNA. (C) Agarose-embedded *Latimeria* genomic DNA was run on a pulse field gel along with a similarly prepared sample from brown Norway rat. Some degradation was evident in the coelacanth (*Lc*) sample, however, the majority of the DNA was still in the well (i.e., was of very high molecular weight). (D) Agarose-embedded DNA was subjected to an EcoRI-EcoRI methylase competition reaction prior to electrophoresing. In this experiment, DNA was partially digested with a standard amount of EcoRI and increasing amounts of methylase (tracks 1-5). Track C represents an untreated control sample. This experiment showed that the DNA was sensitive to competition by EcoRI methylase, which blocks available EcoRI sites (note the increased amount of DNA in the limiting mobility (LM) band in track 5). We conclude from these experiments that the DNA is of good integrity and that the method of blood preservation used by Dr. Dorrington was effective.



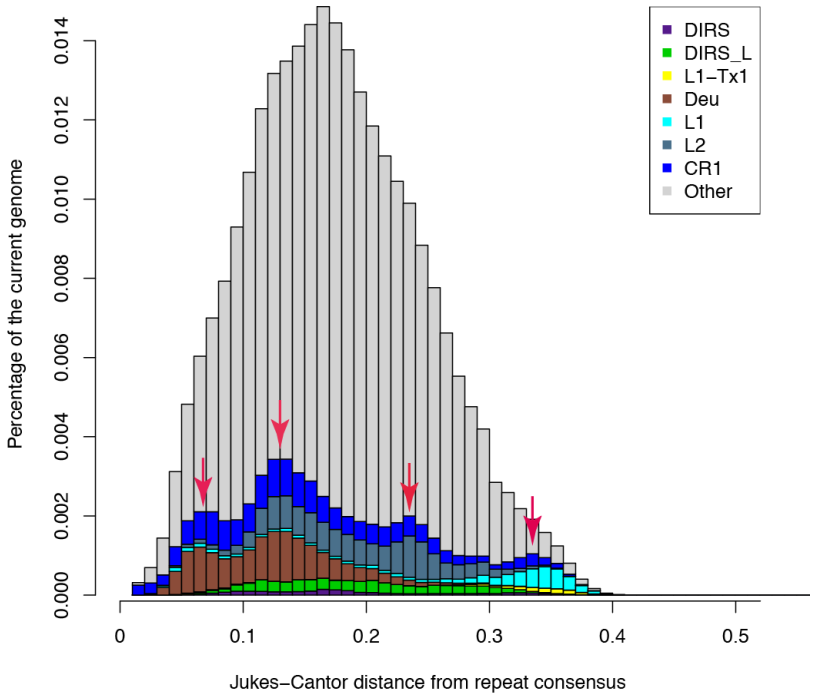
Supplementary Figure 2. Agilent 2100 Bioanalyzer quality plots of RNA from seven tissue samples of *Latimeria chalumnae.* Only muscle RNA (bottom left) was usable; all other samples were too degraded to be used for RNAseq.



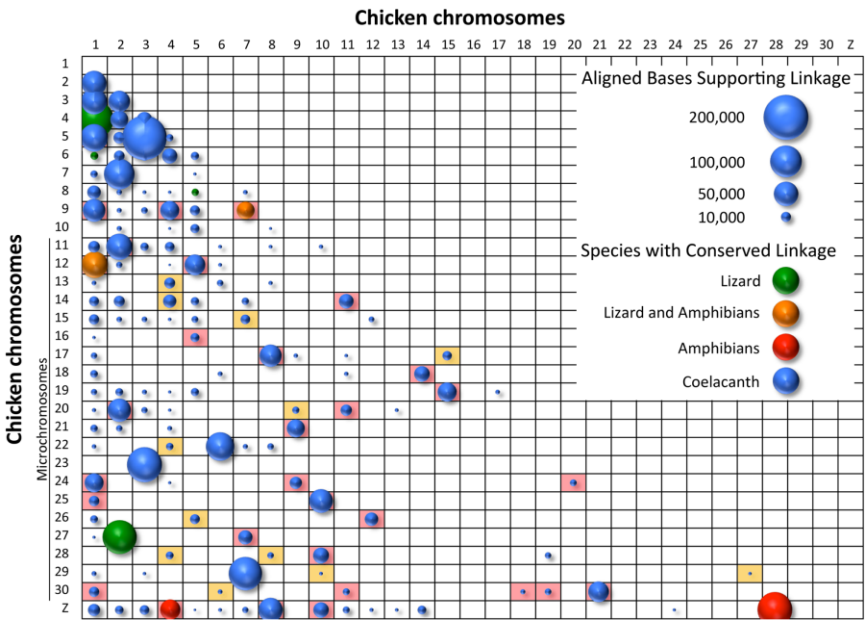
Supplementary Figure 3. Synteny conservation between lincRNAs in human, mouse and coelacanth. LincRNAs are displayed in red, protein-coding genes in blue, microRNA in violet. The shaded area across species represents one to one orthologous relationships.



Supplementary Figure 4. Phylogenetic tree inferred from the same phylogenomic dataset as in Figure 1 but using the worst fitting model LG+F+G4. In this maximum likelihood tree obtained with RAxML, the lungfish and the coelacanth form a clade that is sister to the tetrapods. Confidence estimates were derived from 100 bootstrap replicates and bullets denote nodes receiving maximum support. The scale bar indicates the number of substitutions per site.

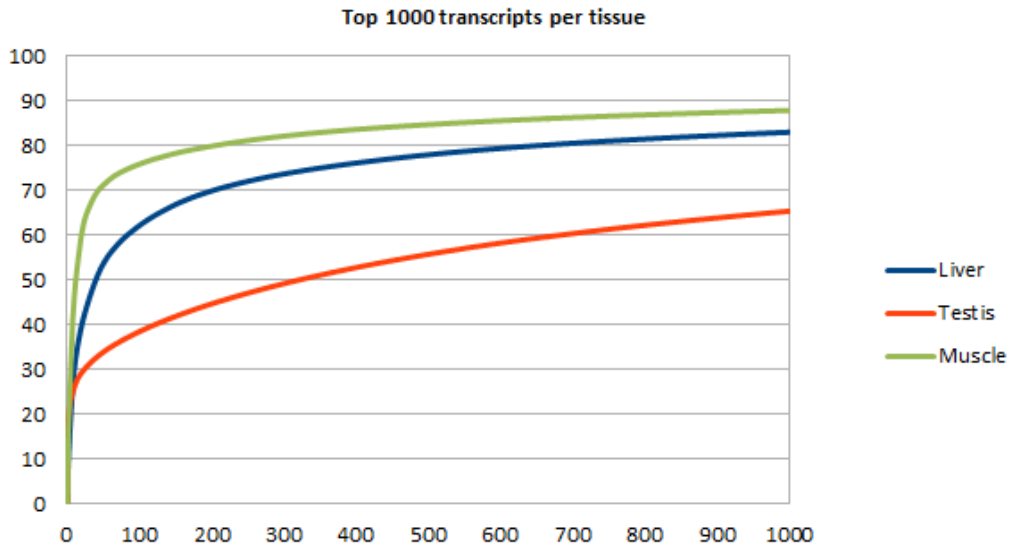


Supplementary Figure 5. TE expansion history in the coelacanth genome. The X-axis indicates a specific TE family at a given divergence from the repeat consensus and Y-axis indicates its fraction of the genome. Arrows indicate the four waves of TE burst.

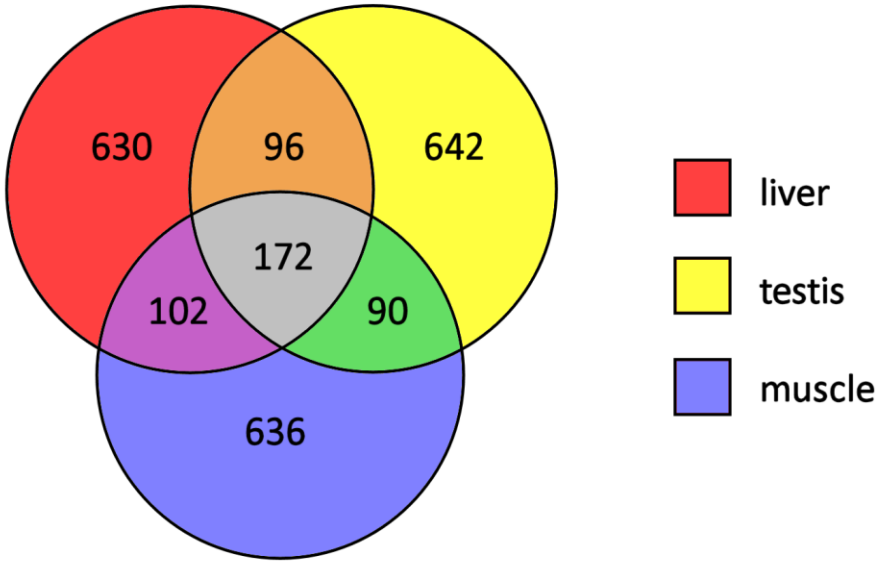


Supplementary Figure 6. Plot of the distribution of linkages between chicken chromosomal homologs among coelacanth scaffolds. Dots represent pairs of chicken chromosomes that align to the same coelacanth scaffold. The size of each dot corresponds to the number bases supporting linkage of homologous sequences in the coelacanth genome, summed across scaffolds. The coloration of dots indicated conservation of linkage in other tetrapod lineages: amphibians (*Ambystoma* and *Xenopus*)[157] and lizard (*Anolis*)[210]. Microchromosomes 29 and 30 correspond to linkage groups E22C19W28_E50C23and E64, respectively. An additional 8 microchromosomes have no assigned sequence in the current chicken genome assembly (galGal3). Coloured cells and those filled completely by alignments correspond to a p-value of <0.01, red cells are <1e-20. Note: chr1/2 and chr1/3 associations have corresponding p-values >0.01.

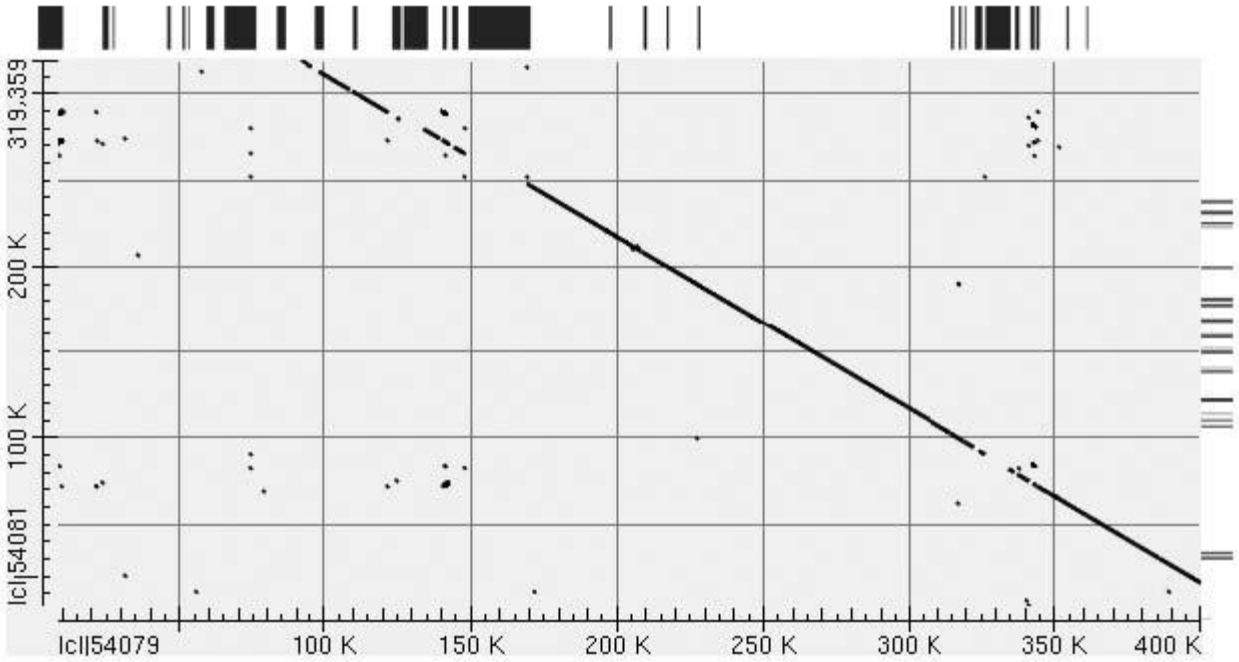
A)



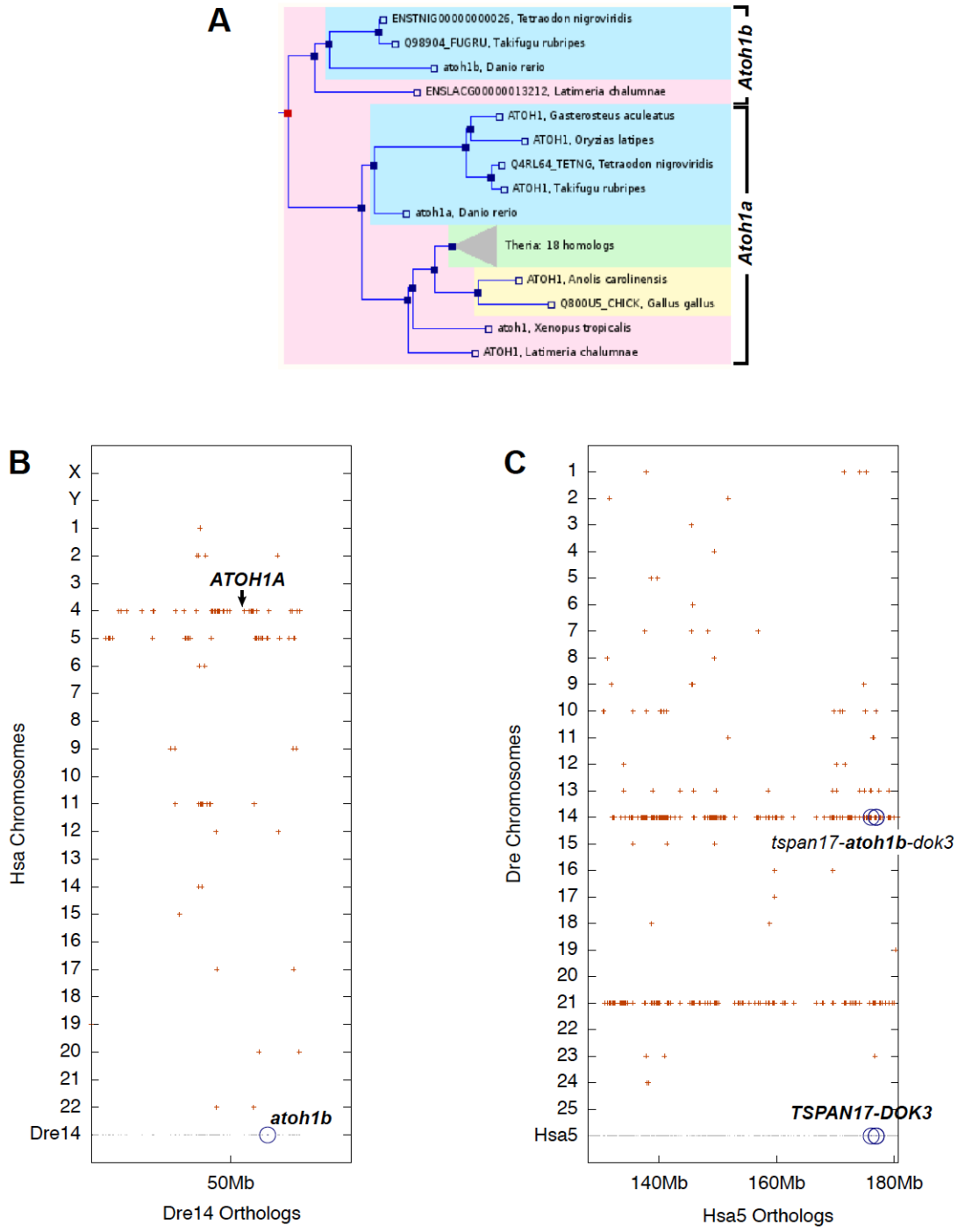
B)



Supplementary Figure 7. *Latimeria* transcriptome comparison: a) Transcriptome richness comparison between coelacanth liver, testis and muscle tissues. Liver and testis expression data were obtained from *L. menadoensis*, whereas muscle expression data were obtained from *L. chalumnae*. The graph represents the cumulative contribution to the total transcription (indicated as the % of the total expression observed in each tissue, on the Y axis) of the 1,000 most expressed transcripts per tissue. b) Venn diagram depicting the overlap between the coelacanth liver, testis and muscle transcriptomes, inferred by the comparison of the 1,000 most expressed transcripts of each tissue. A common set of 172 genes expressed at high levels can be identified in the three tissues. More than 60% of the most highly expressed genes in each of the three tissues analyzed appear to be tissue-specific.



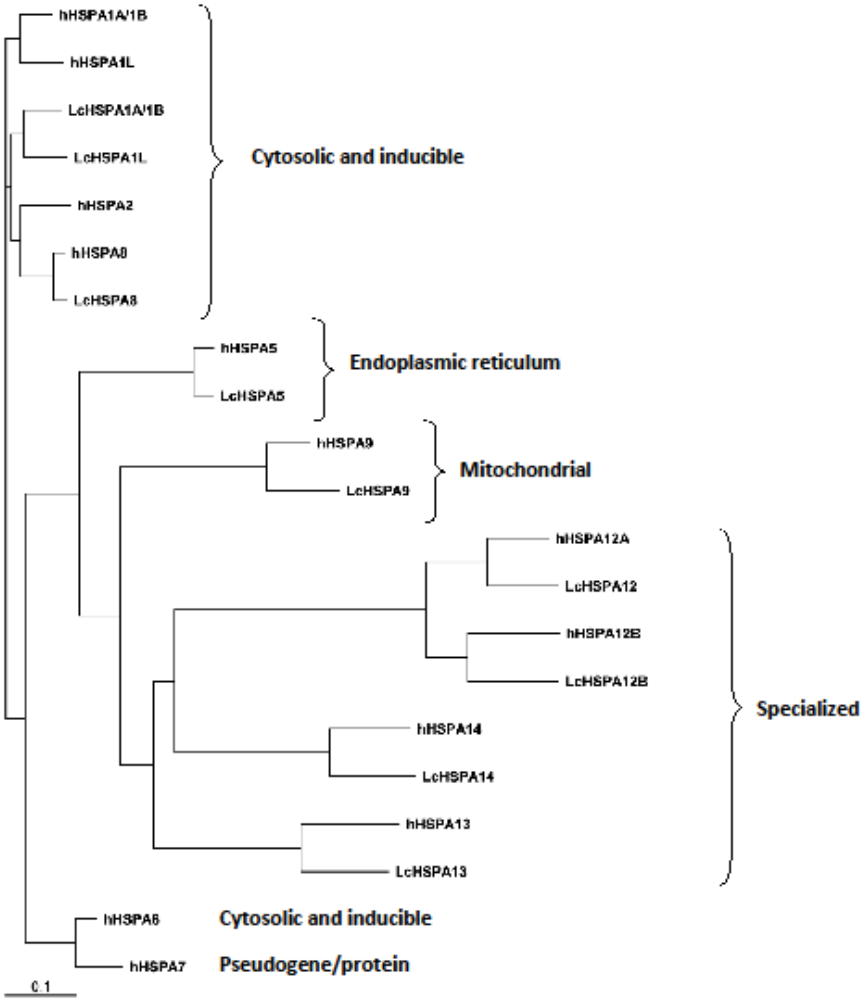
Supplementary Figure 8. Representative dot plot of Megablast alignment between orthologous *L. chalumnae* and *L. menadoensis* genomic regions. Horizontal axis is Lc scaffold 00150 and vertical axis is the Lm HOX-A region derived from overlapping BAC clones (GenBank FJ497005.1, Amemiya et al., 2010). The diagonal line represents the aligned regions between Lc and Lm. Boxes on the right side of the plot represent exons in the HOX-A cluster, with the bottom-most boxes showing the location of *Evx-1* and the top-most boxes showing the location of the anterior end of the cluster. The regions that are not aligned (gaps) are generally accounted for by runs of Ns in the Lc scaffold that are depicted above the plot by boxes.



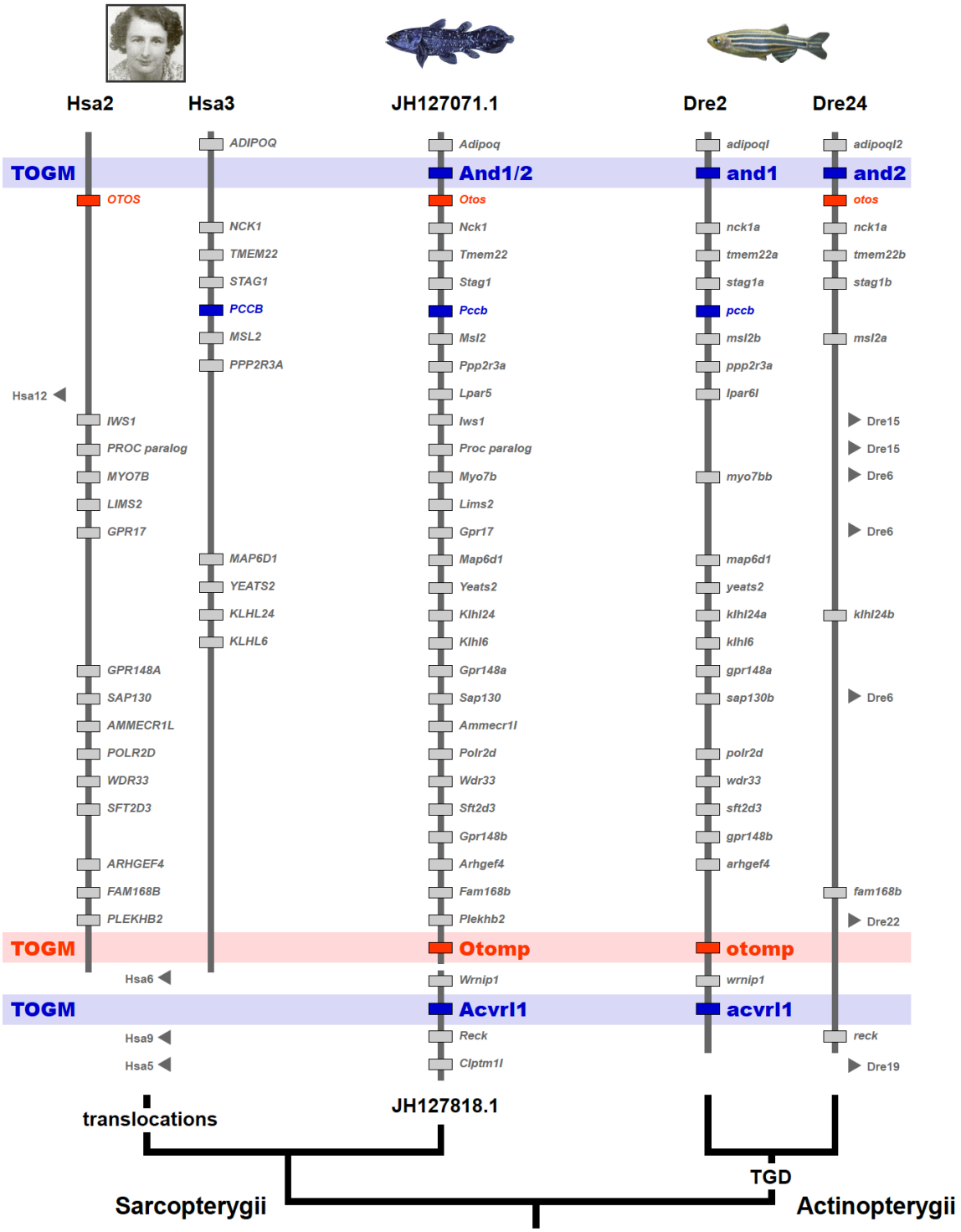
Supplementary Figure 9. Identification of a gene lost in tetrapods, the *Atonal homolog 1b* (*Atoh1b*) gene. A) In EnsemblCompara GeneTree **ENSGT00630000089619**, two *Atoh1* gene clades are apparent: *Atoh1a*, present in teleosts, *Latimeria*, and tetrapods, and *Atoh1b*, present in teleosts and *Latimeria* only. B) and C) Dotplots of zebrafish (Dre) vs. human (Hsa) chromosomes from the Synteny Database. B)

The zebrafish *atoh1b* gene region on Dre14 shares conserved syntenies with human chromsomes Hsa4 (containing *ATOH1A*) and Hsa5 (no *ATOH1* gene present). Zebrafish *atoh1b* is found on Dre8 (not shown). Human chromosomes Hsa4 and Hsa5 are derived from the ancestral vertebrate chromosome *C.* C) Orthologs of the genes flanking *atoh1b*, *tspan17* and *dok3*, are found on Hsa5, which shows double conserved synteny with Dre14 (containing *atoh1b*) and Dre21, as result of the teleost genome duplication.

The combination of phylogenetic (A) and syntenic (B, C) data provides evidence that an *Atoh1* gene on the ancestral vertebrate chromosome *C* was duplicated in the course of the two rounds of vertebrate genome duplication. The *Atoh1a* paralog (ohnolog) was retained in all bony vertebrate lineages (ray-finned and lobe-finned fish, including tetrapods), while *Atoh1b* was lost in tetrapods from a region located on Hsa5 in the human genome.

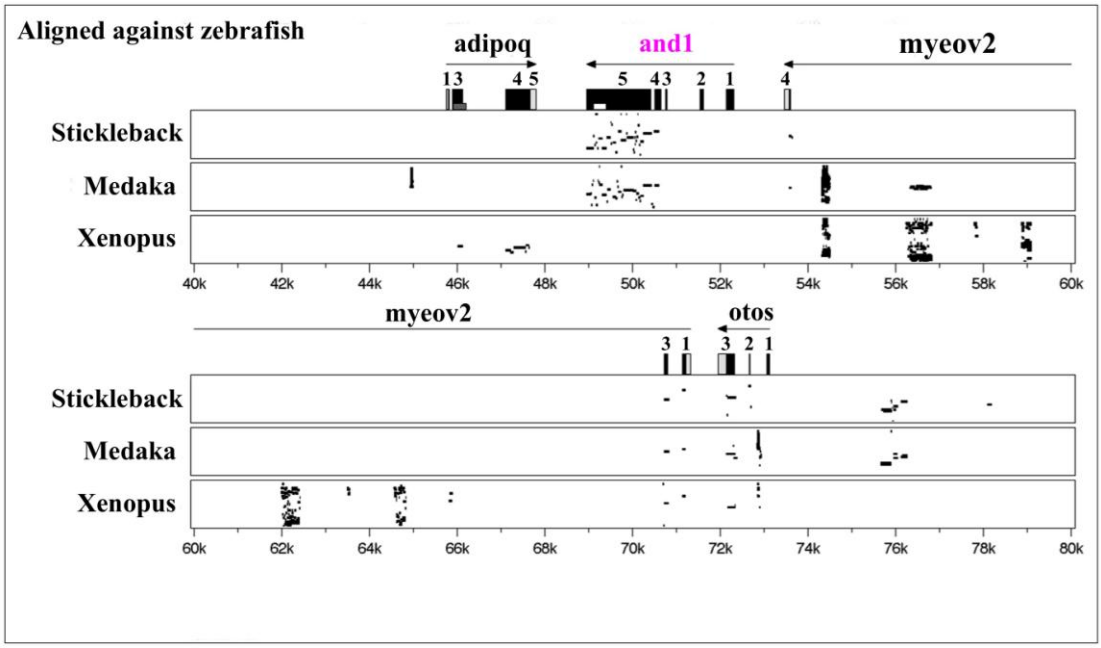


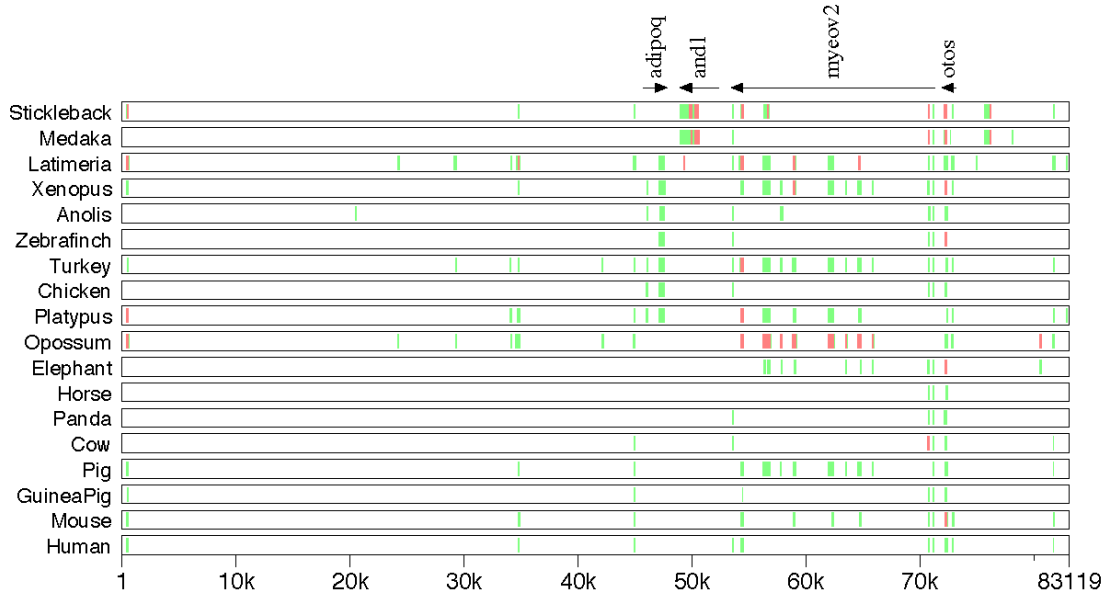
Supplementary Figure 10 - Phylogenetic tree of coelacanth Hsp70s. The tree was calculated using Treeview (version 1.6.6) with the input file being a ClustalW alignment output file generated using the human and coelacanth amino acid sequences in Supplementary Table 17 .



Supplementary Figure 11 - Evolution of the *And1/2 – Otomp – Acvrl1* region in bony vertebrates. Orthologs of genes on coelacanth scaffolds JH126651.1 and JH127818.1 are distributed across chromosomes Hsa2 and Hsa3 in the human (e.g. M. Courtenay-Latimer) genome, indicating translocations on the tetrapod branch leading to human, while teleost (co-) orthologs of these *Latimeria* genes are distributed among two zebrafish chromsomes, Dre2 and Dre24, which contain paralogons

from the teleost genome duplication (TGD). Note that gene order in human and zebrafish is presented according to the coelacanth gene order. The region contains several genes involved in fin (blue) and ear development (red), among them three genes lost in tetrapods: Zebrafish *actinodin 1* (*and1*) and *actinodin2* (*and2*) genes encode structural proteins of the actinotrichia, the skeletal elements that stiffen fin folds, and their loss in tetrapods has been suggested to have contributed to the fin-to-limb transition[180]. Loss of *acvrl1*, encoding a BMP receptor, leads in teleosts to the malformation of the ventral tail fin (*lost-a-fin* mutant)[211-212]. The *otolith matrix protein* (*otomp*) gene is essential for otolith formation in the zebrafish ear[213], the tetrapod homolog of which evolved adaptations for signal detection in air.

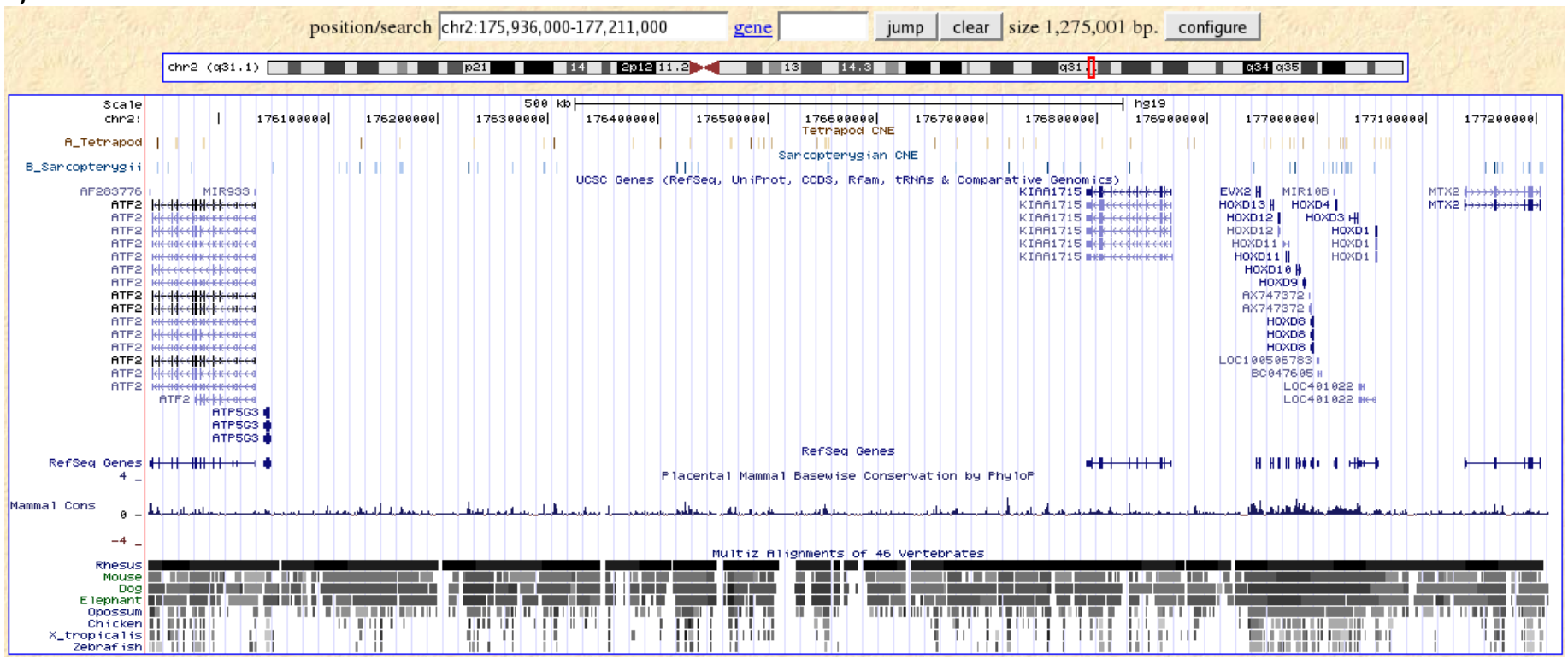




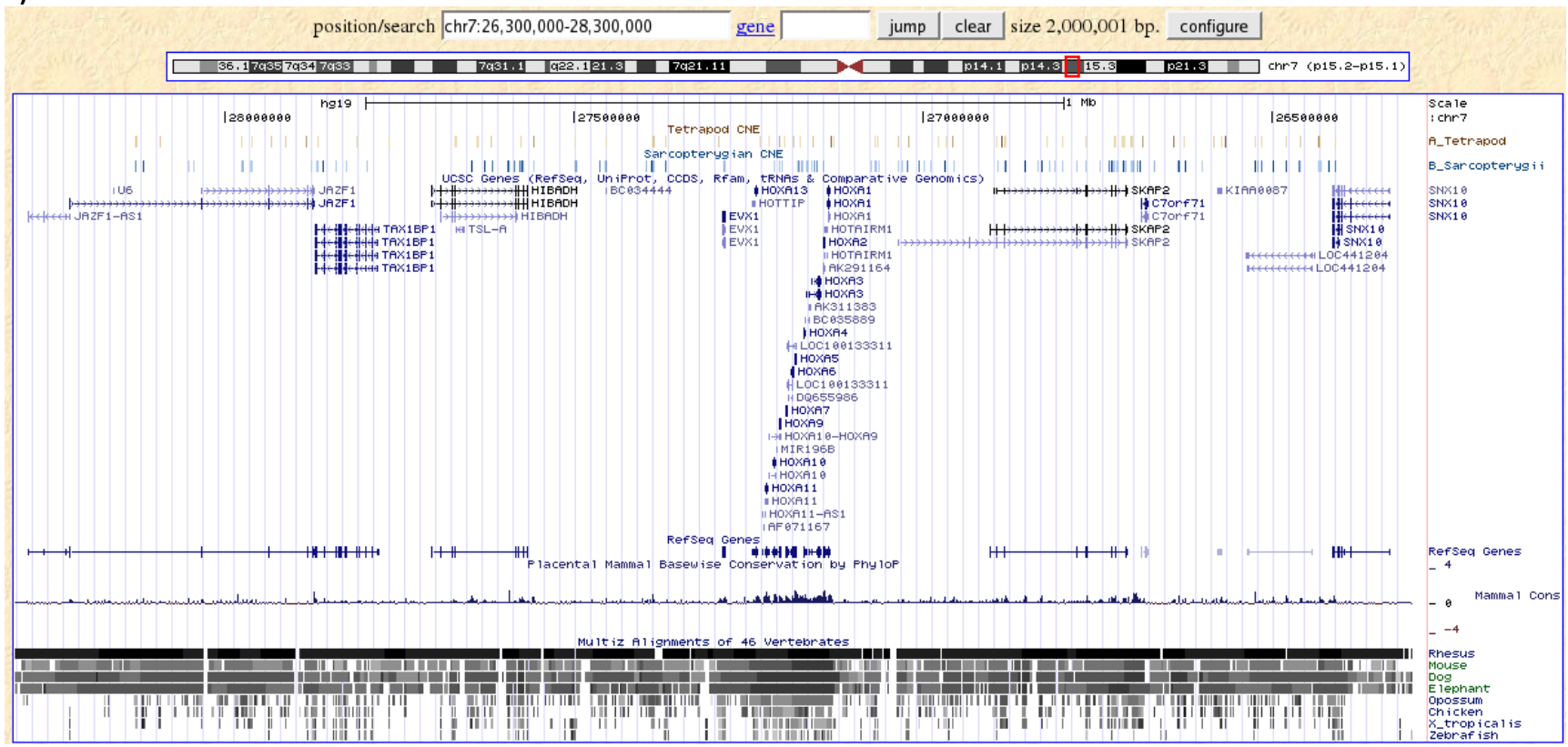
Supplementary Figure 12. Actinodin alignments across vertebrates. a) A MultiPipMaker global alignment of zebrafish *and1* (actinodin) syntenic regions (around 40k bases) among three fishes, zebrafish, Medaka and stickleback, and one amphibian *Xenopus tropicalis*. The comparison was made against the zebrafish sequence. All sequences were extracted from Ensembl genomic databases available at the Welcome Trust Sanger Institute, Genome Research Limited. In *Xenopus*, the comparable conserved elements to fish *and1* were not observed in the sequence between *adipoq* and *myeov2*. *adipoq*: adiponectin; *myeov2*: myeloma overexpressed2; *otos*: otospiralin. The annotation of the zebrafish *and1* was made using NCBI GenBank accession NM_00119725 and Zhang et al. (2010). MultiPipMaker: http://pipmaker.bx.psu.edu/pipmaker/ (Schwartz et al., 2000). Note that the teleost fish Atlantic cod (*Gadus morhua*) does have *adipoq*, *and1*, *myeov2,* and *otos* but their syntenic relationships

are not yet determined (Ensembl genomic databases at the Welcome Trust Sanger Institute, Genome Research Limited, updated on January 17, 2012). b) A MultiPipMaker alignment of zebrafish and *Latimeria chalumnae and1* (actinodin) syntenic regions among different vertebrate animals. The comparison was made against the zebrafish. The conserved *and1* element was only found in medaka, stickleback and *L. chalumnae* when the zebrafish *and1* syntenic region was compared against the regions of the other vertebrates. *adipoq*: adiponectin; *myeov2*: myeloma overexpressed2; *otos*: otospiralin.

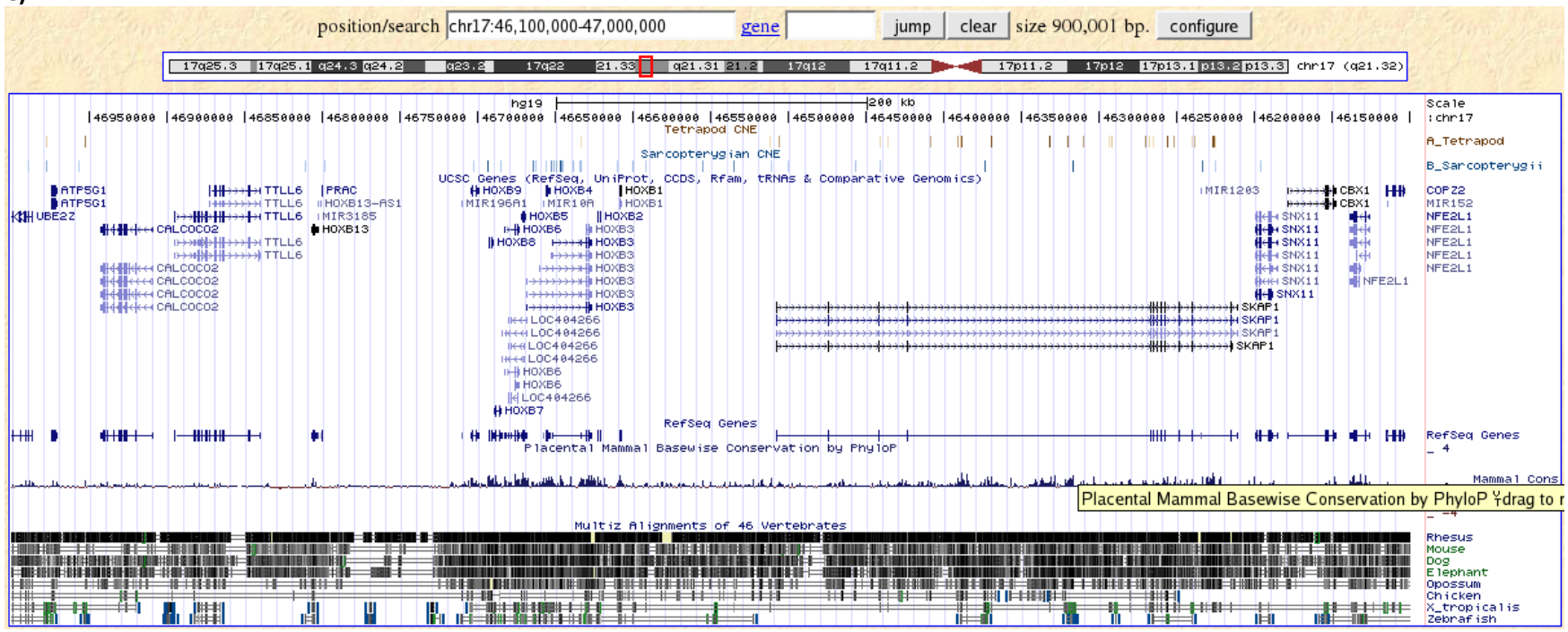
A)


position/search
chr2:175,936,000-177,211,000
gene
jump
clear
size 1,275,001 bp.
configure
chr2 (q31.1)
Scale
chr2:
500 kb
hg19
Tetrapod CNE
A_Tetrapod
Sarcopterygian CNE
B_Sarcopterygii
UCSC Genes (RefSeq, UniProt, CCDS, Rfam, tRNAs & Comparative Genomics)
AF283776
MIR933
ATF2
ATP5G3
KIAA1715
EVX2
MIR10B
HOXD13
HOXD4
HOXD12
HOXD3
HOXD1
HOXD11
HOXD10
HOXD9
AX747372
HOXD8
LOC100506783
BC047605
LOC401022
MTX2
RefSeq Genes
Placental Mammal Basewise Conservation by PhyloP
Mammal Cons
Multiz Alignments of 46 Vertebrates
Rhesus
Mouse
Dog
Elephant
Opossum
Chicken
X_tropicalis
Zebrafish

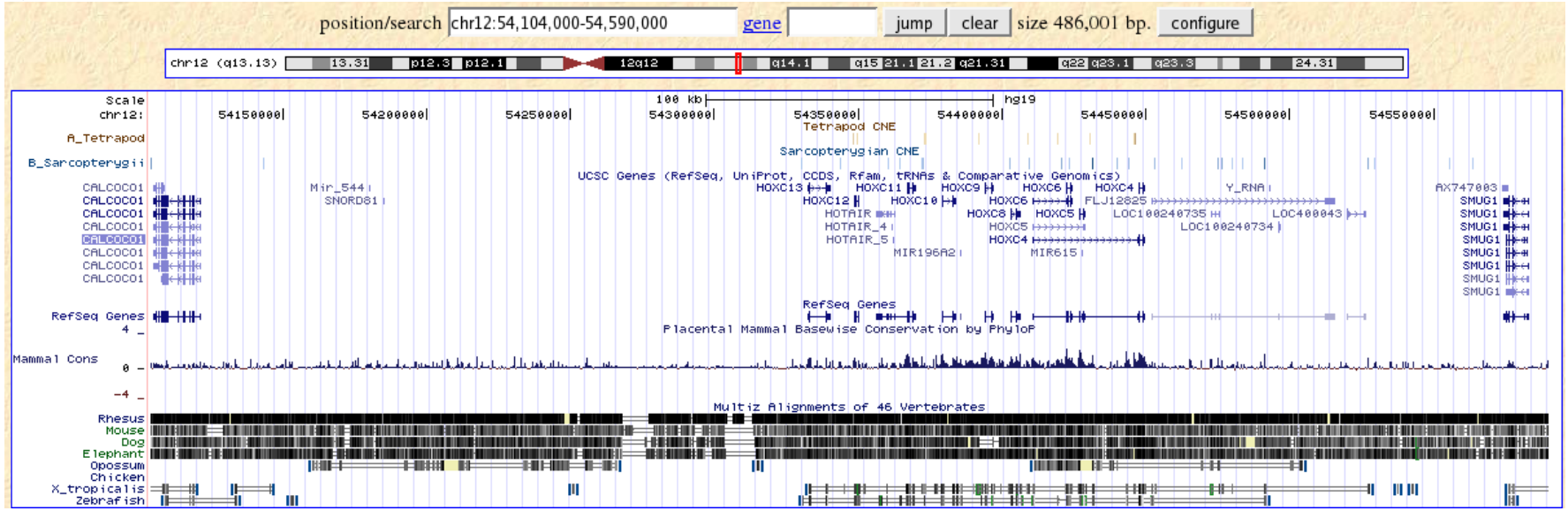
B)


position/search chr7:26,300,000-28,300,000
gene
jump
clear
size 2,000,001 bp.
configure
chr7 (p15.2-p15.1)
hg19
1 Mb
28000000
27500000
27000000
26500000
Tetrapod CNE
Sarcopterygian CNE
UCSC Genes (RefSeq, UniProt, CCDS, Rfam, tRNAs & Comparative Genomics)
JAZF1
HIBADH
BC034444
HOXA13
HOTTIP
HOXA1
HOXA2
HOTAIRM1
AK291164
HOXA3
AK311383
BC035889
HOXA4
LOC100133311
HOXA5
HOXA6
DQ655986
HOXA7
HOXA9
HOXA10-HOXA9
MIR196B
HOXA10
HOXA11
HOXA11-AS1
AF071167
EVX1
JAZF1-AS1
TAX1BP1
TSL-A
SKAP2
C7orf71
KIAA0087
SNX10
LOC441204
RefSeq Genes
Placental Mammal Basewise Conservation by PhyloP
Multiz Alignments of 46 Vertebrates
Scale
chr7
A_Tetrapod
B_Sarcopterygii
Mammal Cons
Rhesus
Mouse
Dog
Elephant
Opossum
Chicken
X_tropicalis
Zebrafish

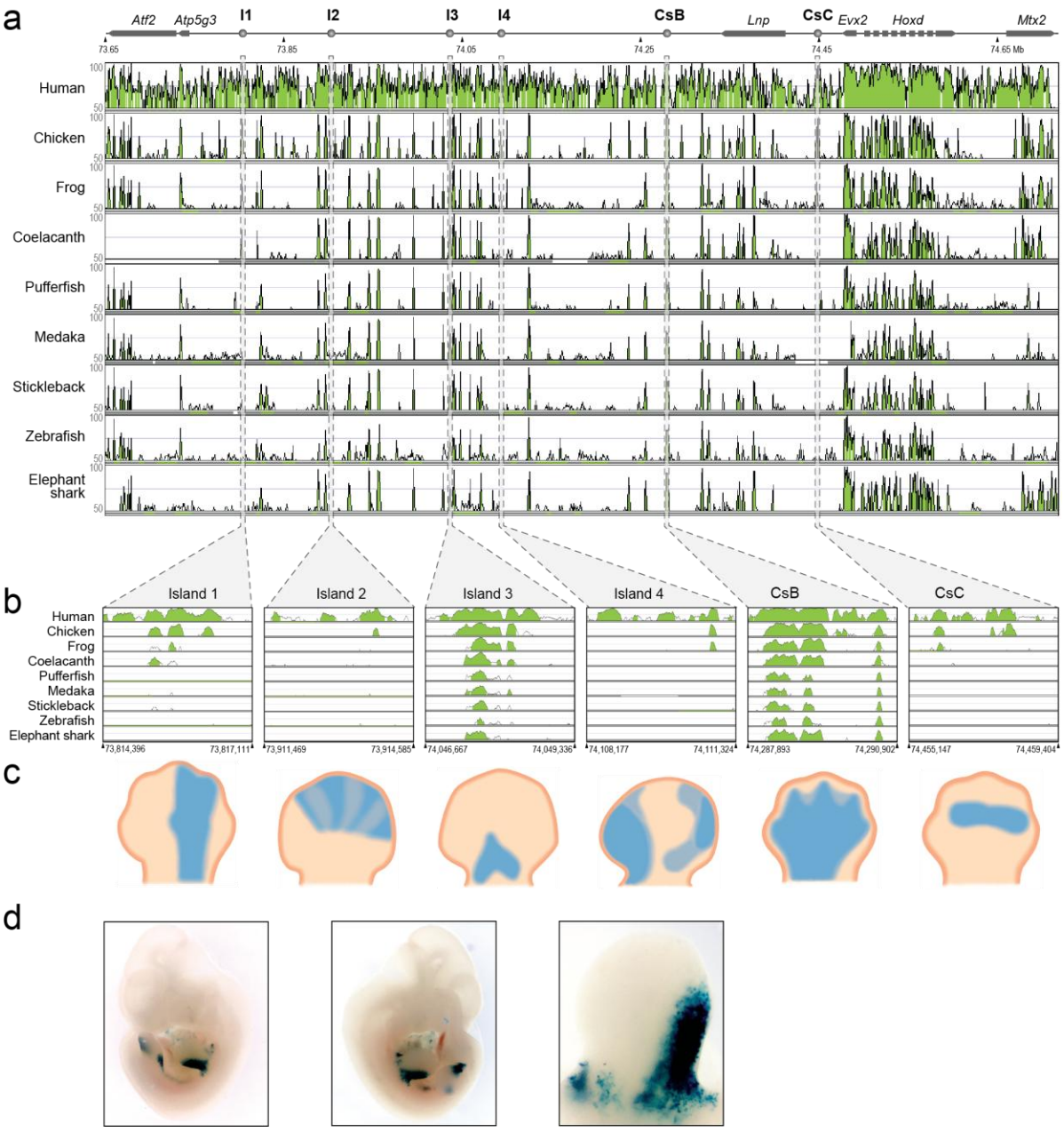
C)


position/search chr17:46,100,000-47,000,000
gene
jump
clear
size 900,001 bp.
configure
hg19
200 kb
Tetrapod CNE
Sarcopterygian CNE
UCSC Genes (RefSeq, UniProt, CCDS, Rfam, tRNAs & Comparative Genomics)
RefSeq Genes
Placental Mammal Basewise Conservation by PhyloP
Placental Mammal Basewise Conservation by PhyloP
Multiz Alignments of 46 Vertebrates
Rhesus
Mouse
Dog
Elephant
Opossum
Chicken
X_tropicalis
Zebrafish

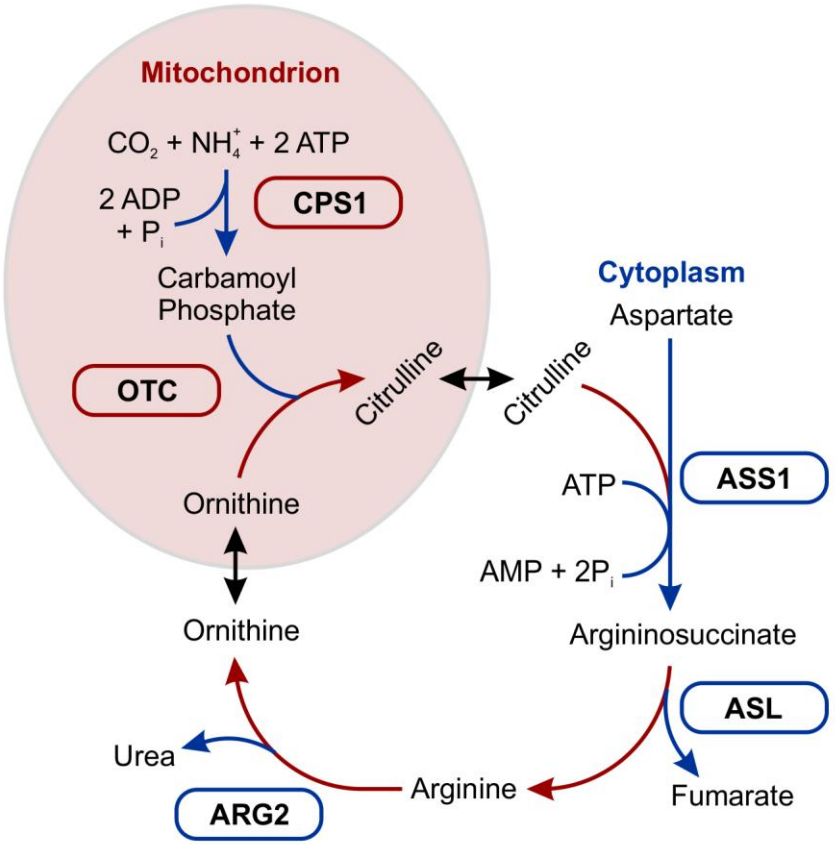
D)



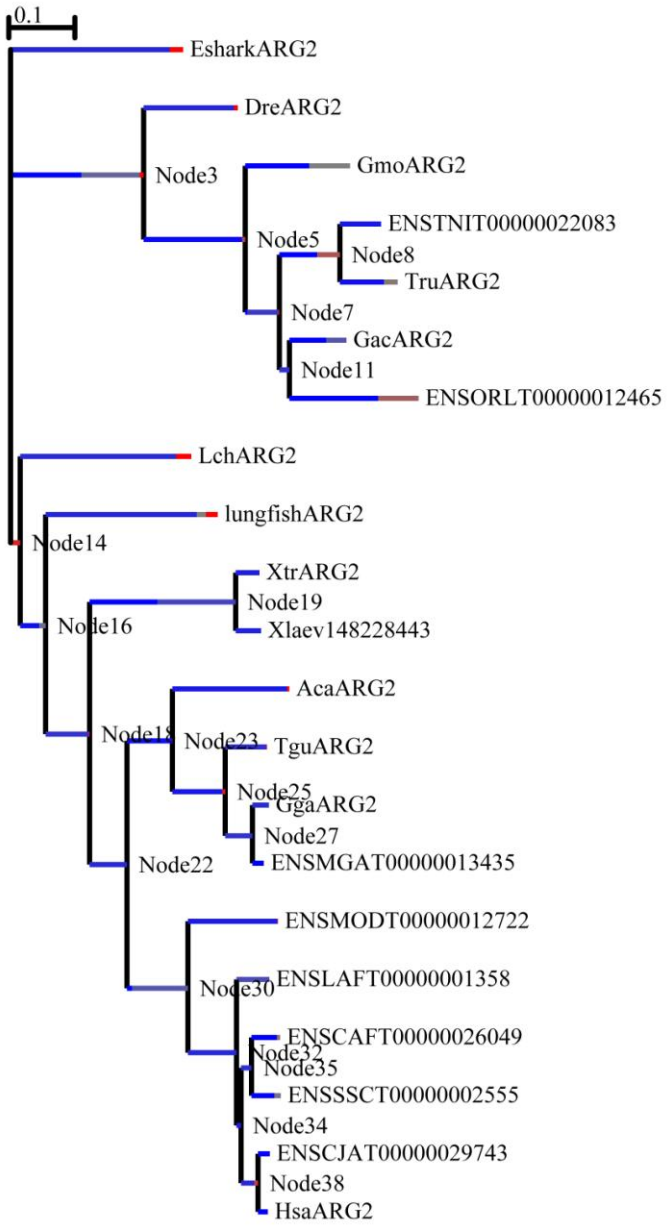
Supplementary Figure 13. Tetrapod and sarcopterygian CNEs in human Hox regions a) HoxD b) HoxA c) Hox B d) Hox C



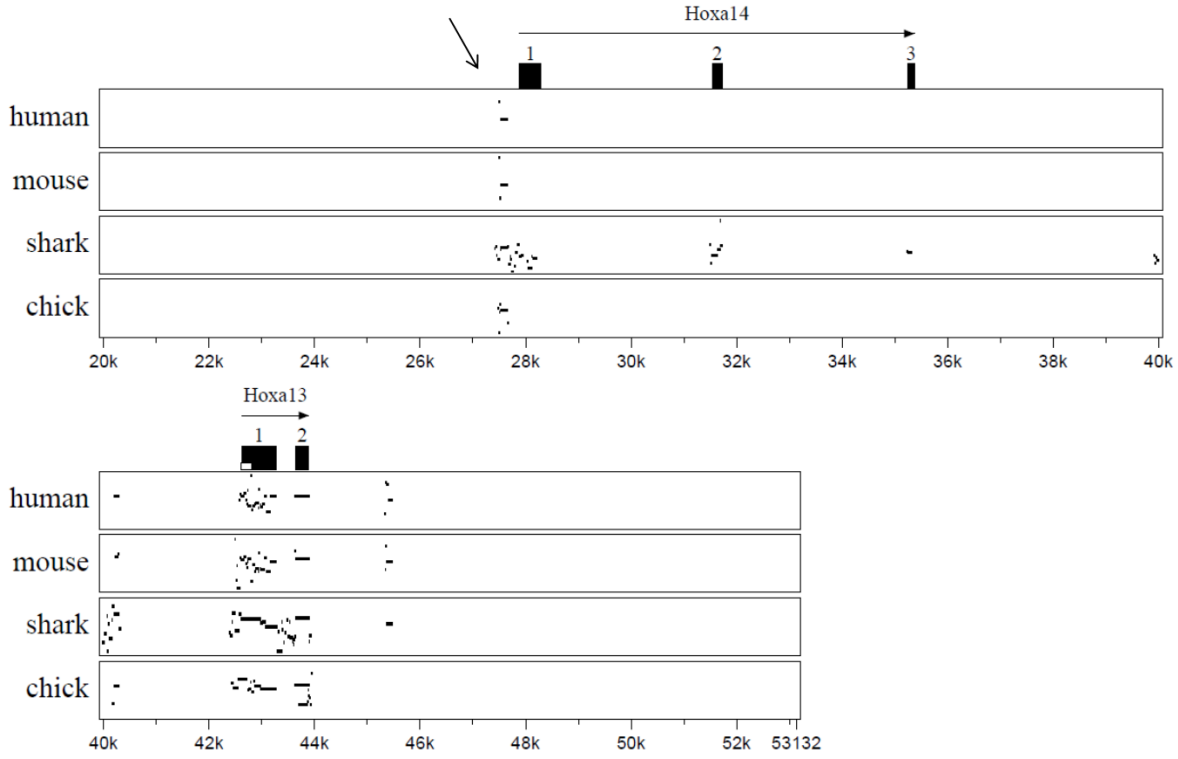
Supplementary Figure 14. Alignment of the HoxD locus and upstream gene desert identifies conserved limb enhancers. (a) Organization of the mouse HoxD locus and centromeric gene desert, flanked by the ATF2 and MTX2 genes. Limb regulatory sequences (I1, I2, I3, I4, CsB and CsC) are noted. Using the mouse locus as a reference (NCBI37/mm9 assembly), corresponding sequences from human, chicken, frog, coelacanth, pufferfish, medaka, stickleback, zebrafish and elephant shark were aligned. Alignment (mVISTA program, homology threshold 70%) shows regions of homology between tetrapod, coelacanth and ray-finned fishes. (b) Alignment of vertebrate cis-regulatory elements I1, I2, I3, I4, CsB and CsC. (c) Expression patterns driven by each regulatory element assayed via mouse transgenesis. (d) Expression patterns of coelacanth Island I in a transgenic mouse. Limb buds indicated by arrowheads in the first two panels. The third panel shows a close-up of a limb bud.



Supplementary Figure 15 - Schematic representation of the hepatic urea cycle. In the mitochondrion the toxic ammonium (NH_4^+) is coupled with carbondioxide (CO_2)and phosphate from ATP to produce carbamoyl phosphate. This is the rate limiting step of the cycle and is catalyzed by the enzyme carbamoyl phosphate synthase 1 (CPS1). The carbamoylphosphate is then transferred to ornithine by ornithine-carbamoylphosphate transferase, leaves the mitochondrion and is further metabolised in two steps by argininosuccinate synthase 1 (ASS1) and argininosuccinatelyse (ASL) to finally generate the aminoacid arginine. By arginase 2 (ARG2) urea is released and ornithine is recovered, which then enters the mitochondrion to initiate a new round of the cycle.



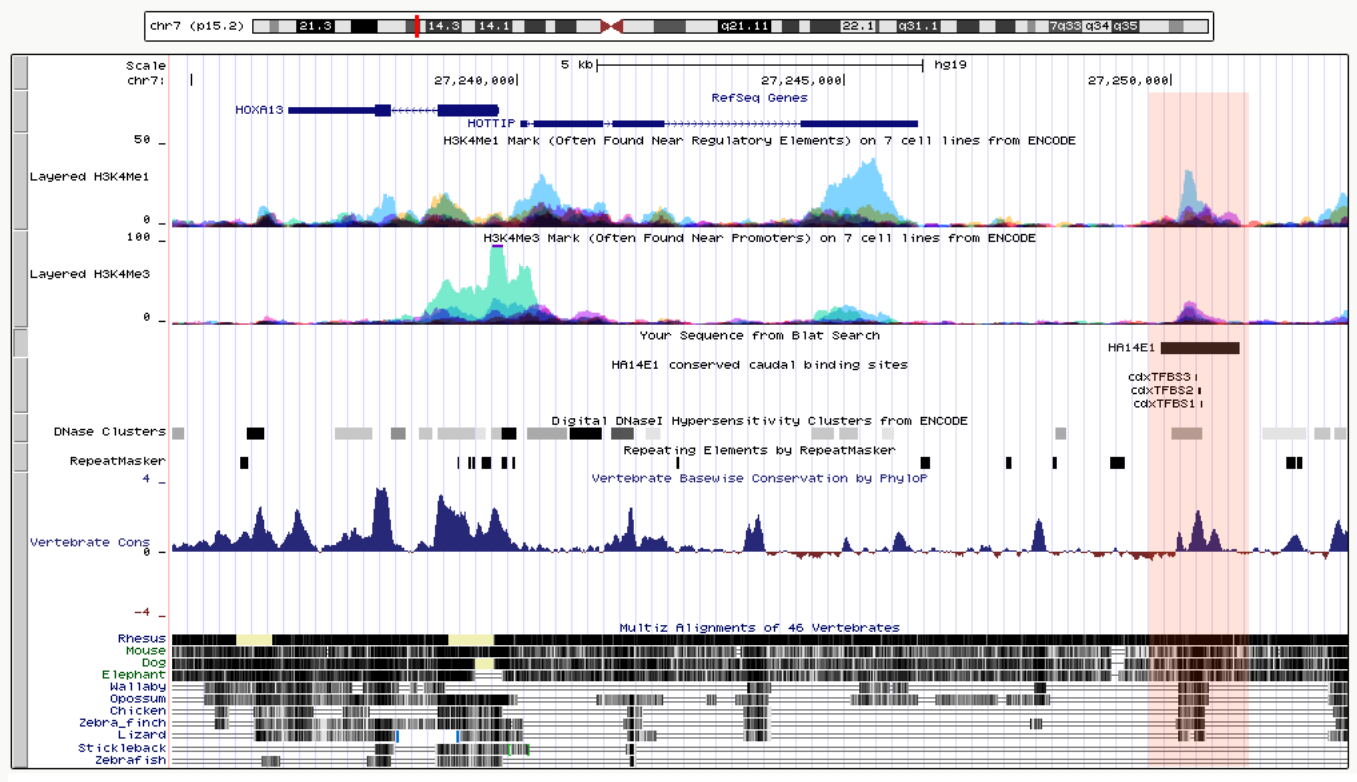
Supplementary Figure 16 - Test for episodic positive selection on ARG2 coding sequences. Branch lengths are scaled to the expected number of substitutions/nucleotide and Branch colour indicates the type of selection (dN/dS or ω) with red corresponding to positive or diversifying selection (ω > 5), blue to purifying selection (ω = 0), and grey to neutral evolution (ω = 1). The proportion of each colour on a branch represents the fraction of the sequence undergoing the corresponding class of selection. Thick branches would indicate statistical support for positive selection. Note that there is no evidence for selection in ARG2 within the vertebrate tree.



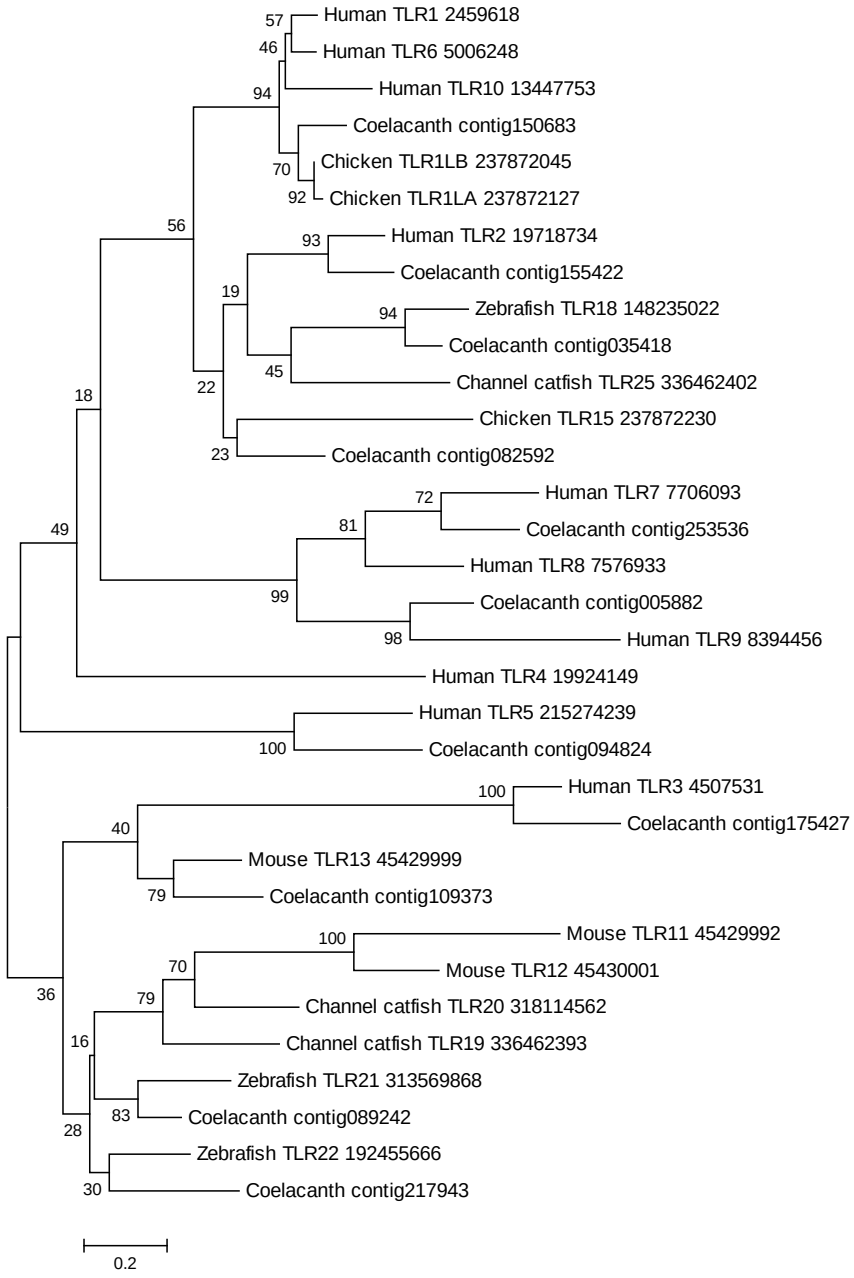
Supplementary Figure 17a. Percent identity plot of the 5' end of coelacanth HOX-A with orthologous regions from the human, mouse, horn shark and chicken. Coelacanth was used as the reference sequence for comparisons. The horn shark (*Heterodontus francisci*) *Hoxa14* gene is a degenerated pseudogene. Black boxes above the plots indicate exon positions; notable nucleotide identities are given by dots in the plot. The region upstream of the 1st exon of *Hoxa14* (arrow) is a highly conserved *cis*-regulatory element called H14E1. Based on functional experiments in chick and mouse, we surmise that this promoter-enhancer element may be involved in development of extraembryonic structures such blood islands/vasculature in the chick as well as placental labyrinth vasculature in the mouse.

```
humanA          GTCGGAGGAAACGCTTTTACCACCTGGGCGACCTTGACTGCAGCCGATTAAAGTTTAATC
60

mouseA          -----AGGAAACGCTTTTACCACCTGGACGACCTTGACTGCAGCCGATTAAAGTTTAATC
55

chickenA        -TCAAAGTGAAGTCATTTACCACCTGGACGTCCTTGACTTCGGATGATTAAAGTTTAATC
59

                     **  **  * *********** ** ******** * *  ***************


humanA          CGAGGTGTGTGCTCAGACTTGCCATGTTATTTAAACACATCAAAGGTCATAAAAAGATTC
120

mouseA          CGAGGTGTGTGCTCAGCCTTGCCATGTTATTTAAACACATCAAAGGTCATAAAAAGATTC
115

chickenA        CGAGGTGTGTGCACAGCTTTACCGTGTTATTTAAACACATCAAAGGTCATAAAGGGATTC
119

                ************ ***  ** ** ****************************  *****


humanA          CAATGGCGT 129

mouseA          CAAAGGCGT 124

chickenA        CAACAGC-- 126

                ***  **
```

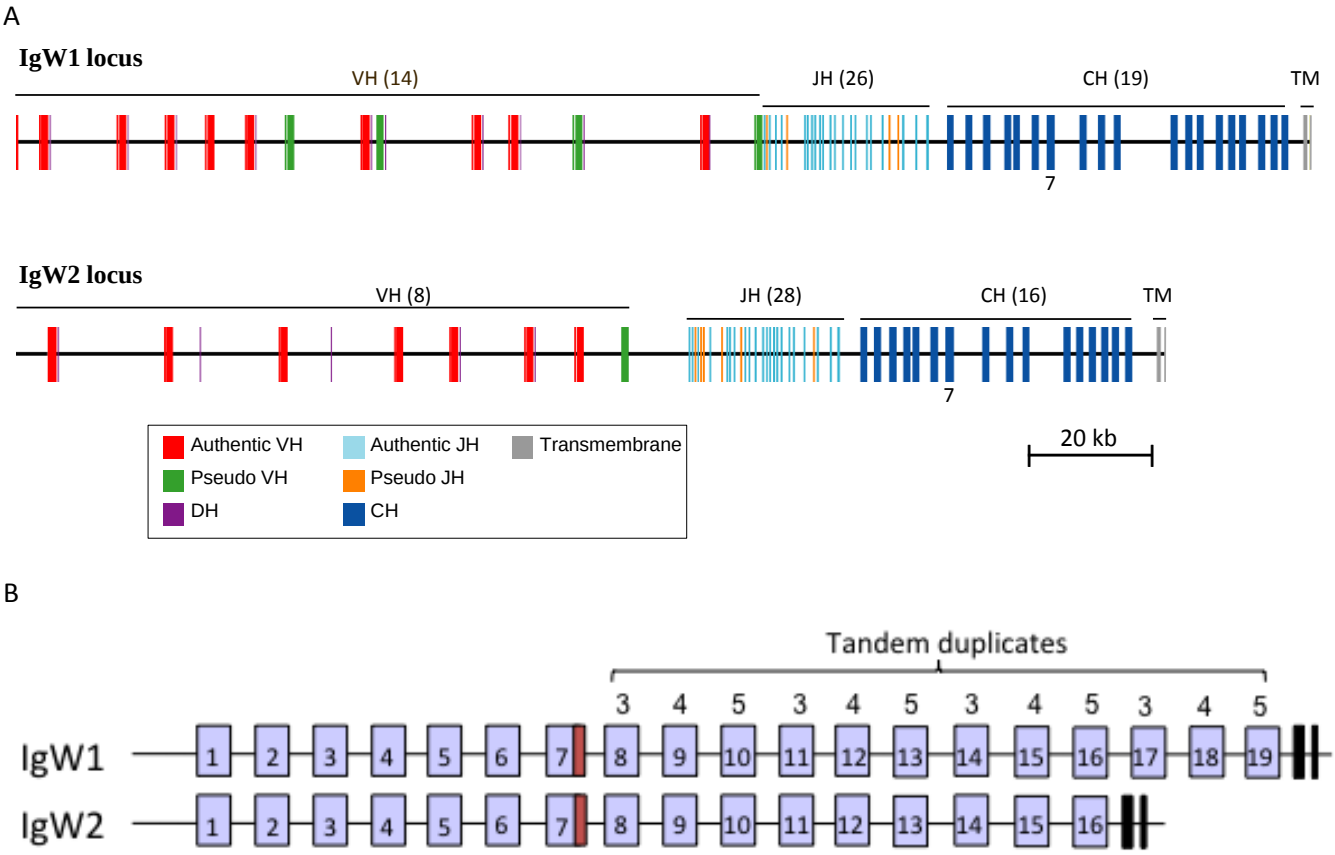
Supplementary Figure 17b. Alignment of core region of HA14E1 from human, mouse and chicken. The shaded areas are *bona fide* caudal binding sites as identified by searches of the TransFac database. Caudal is a homeodomain-containing transcription factor and known regulator of *Hox* genes[214-216].



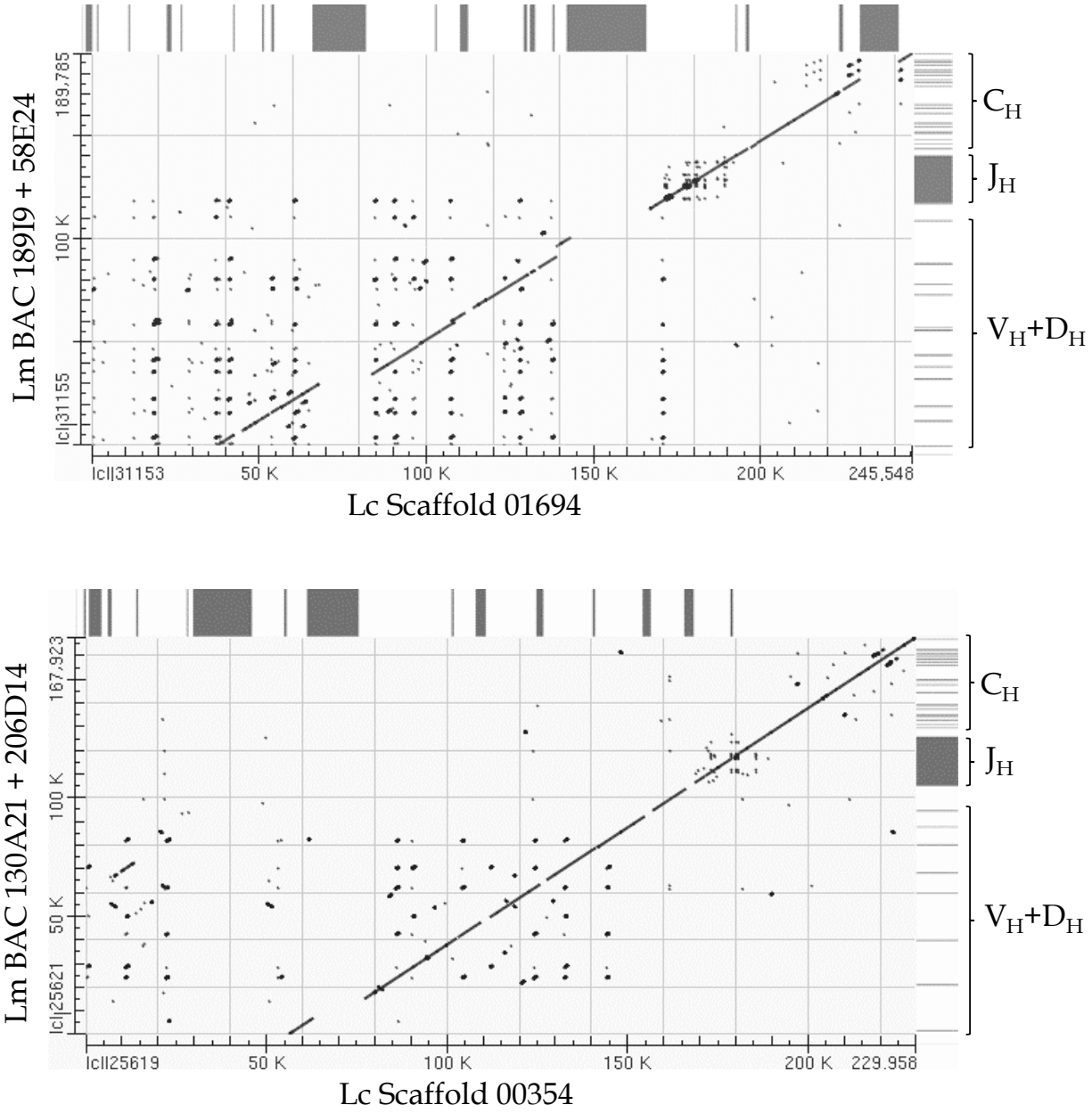
Supplementary Figure 17c. Transcriptional landscape of the HA14E1 and its immediate vicinity. This figure was generated via the UCSC Browser and by incorporating custom tracks. The region is centered using the human assembly (NCBI36/hg18) from March 2006. The pink highlighted region represents the 1.5 kb HA14E1 sequence upstream of *Hoxa13* that was previously used for a mouse transgenic experiment (http://enhancer.lbl.gov/cgi-bin/imagedb3.pl?form=presentation&show=1&experiment_id=501&organism_id=1) but which showed no reporter activity along the AP axis at E11.5. Profiles of chromatin methylation marks H3K4Me1 and H3K4Me3 as well as DNase hypersensitivity clusters suggest that the HA14E1 region has promoter-enhancer activity, at least *in vitro*. Strong vertebrate conservation is seen in the composite vertebrate conservation plot (dark blue) and the MultiZ alignments on the bottom of the figure. There is no apparent conservation to human HA14E1 with the two teleost fishes (stickleback and zebrafish), whereas the conservation in marsupials (wallaby, opossum), birds (chicken, zebra finch) and lizard is more restricted to the core region of the HA14E1 element. This core region contains the three caudal binding sites as shown in Supplementary Figure 16b above.



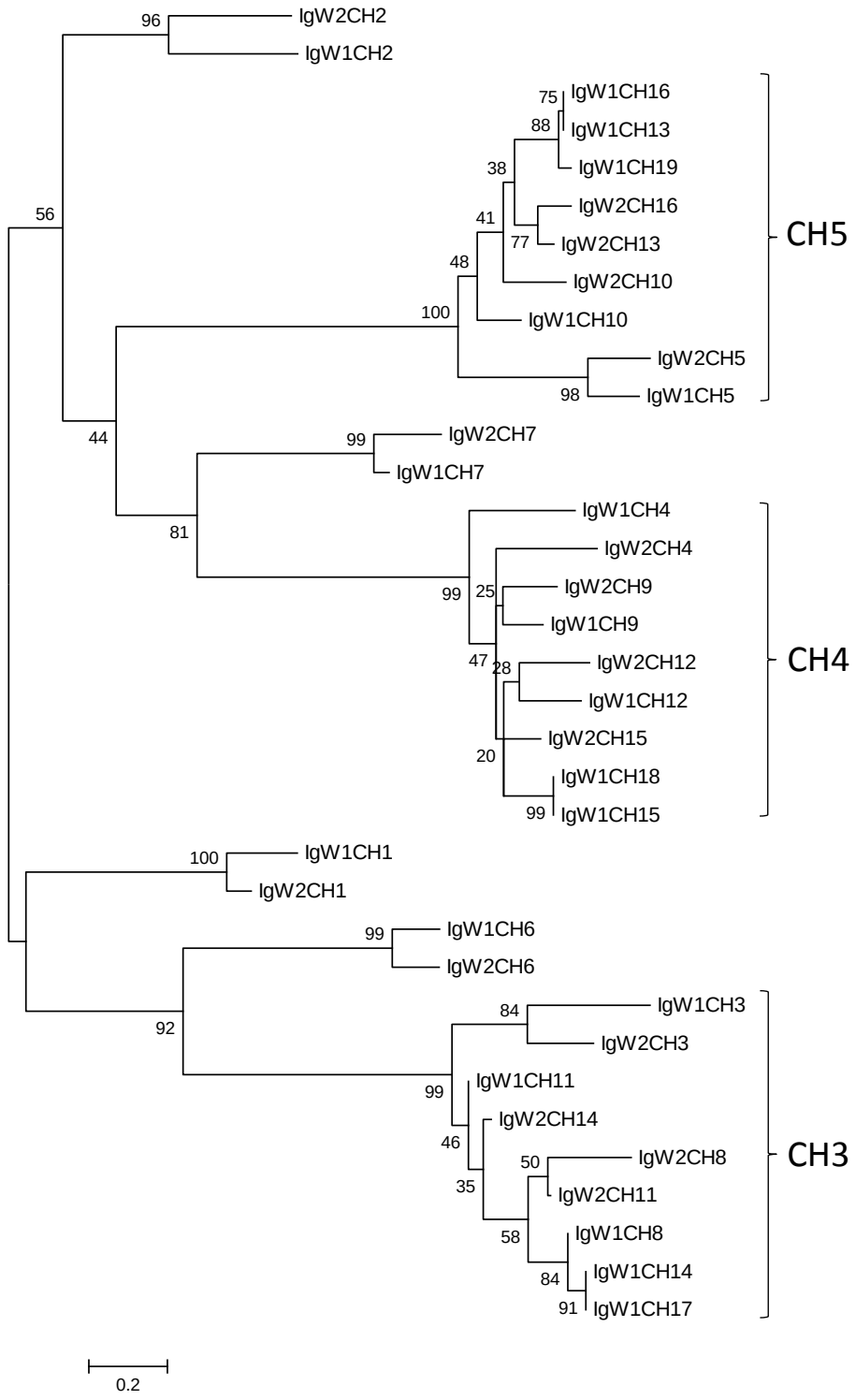
Supplementary Figure 18 – Toll-Like Receptor Phylogeny.
The evolutionary history was inferred by using the Maximum Likelihood method based on the JTT matrix-based model for TIR domain of Toll-like receptors. The tree with the highest log likelihood (-4795.6723) is shown. The percentage of trees in which the associated taxa clustered together is shown next to the branches. The tree is drawn to scale, with branch lengths measured in the number of substitutions per site. All positions containing gaps and missing data were eliminated and a total of 102 positions were used. Evolutionary analyses were conducted in MEGA5.



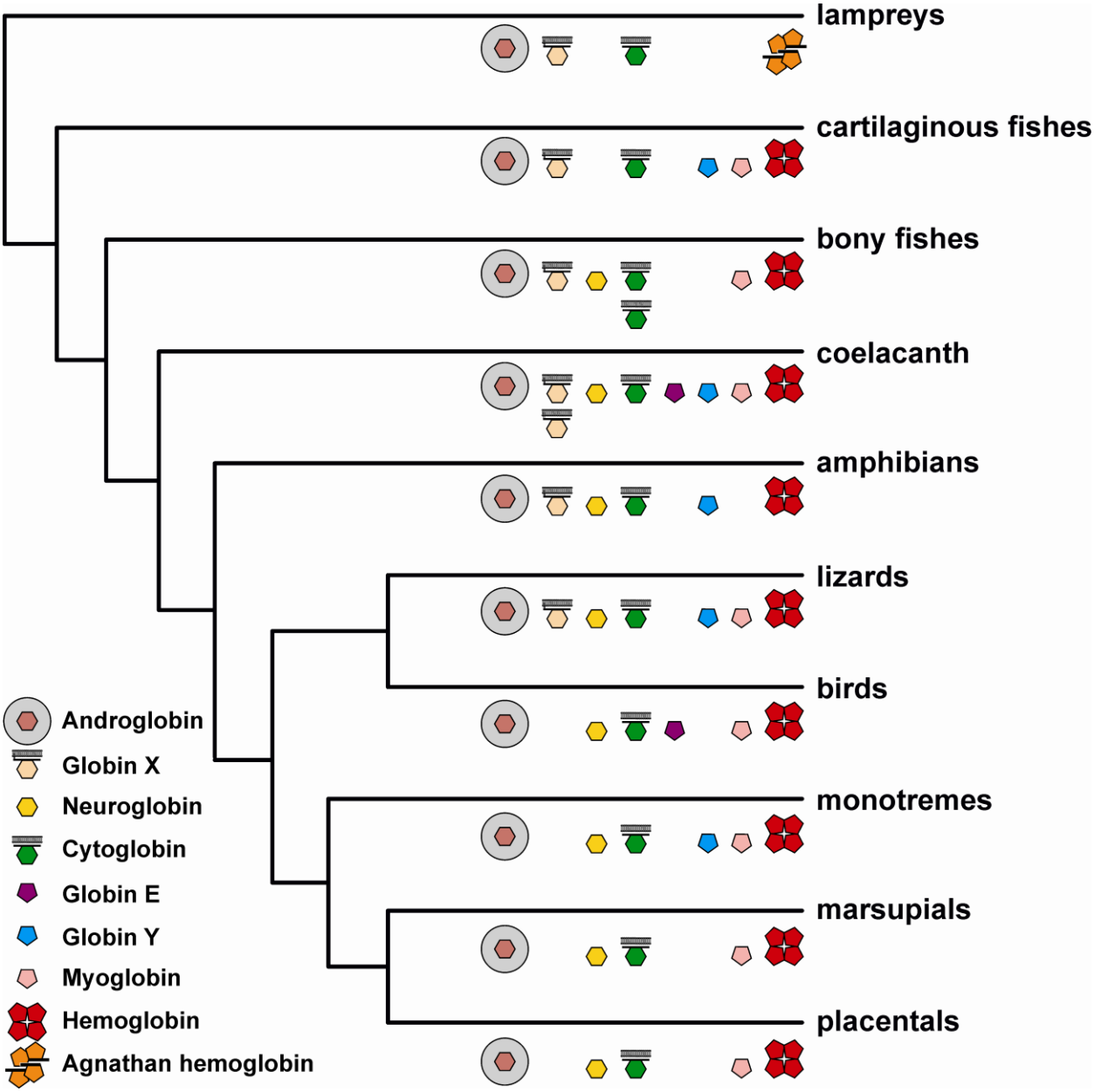
Supplementary Figure 19 - Immunoglobulin heavy chain genome organization in the coelacanth. Overlapping BAC clones encompassing the heavy chain loci of *L. menadoensis* were isolated and sequenced and shown to encompass two discrete IgH loci. The two loci are shown in (A) and were both found to encode IgW molecules similar to that found in the African lungfish and cartilaginous fishes. A locus encoding a heavy chain recognizable as IgM could not be identified through bioinformatics searches or via direct hybridization or degenerate primer PCR screening strategies. (B) Illustration of the exon structure encoding IgW constant domains (not to scale). Blue boxes represent the CH exons whereas the red boxes represent regions encoding secretory domains, and the black boxes represent regions encoding transmembrane domains. For both IgW1 and IgW2 the exons following exon 7 possess four and three tandem duplications, respectively, of exons 3-5 (Supplementary Figure 20). The usage of these other exons is, as yet, unclear.



Supplementary Figure 20 - Concordance of *L. chalumnae* (Lc) and *L. menadoensis* (Lm) genomic sequences encoding the two IgW loci (top – IgW1, bottom – IgW2). The dot-plots are graphical depictions of Megablast alignments of orthologous IgW genomic regions from the two coelacanth species. Horizontal axes represent Lc scaffolds and vertical axes represent Lm assemblies based on overlapping BAC clones. Boxes on the right side of the plots denote relative positions of coding sequences in the respective IgW loci within the BAC assemblies. The diagonal lines in the plots indicate strong concordance between the IgW loci of both species. The regions that are not aligned (gaps in the diagonal line) are largely accounted for by runs of N's in the Lc scaffold that are depicted above the plot by black boxes.



Supplementary Figure 21 - Neighbor-Joining tree of amino acid sequences of all *Latimeria menadoensis* CH exons. The tree was constructed using MEGA 5 and 1000 bootstrap replications. Brackets on the right denote the clustering of the respective tandemly duplicated exons as diagrammed in Supplementary Figure 18.



Supplementary Figure 22 - Distribution of globin genes in vertebrates. The hexagon indicates hexacoordinate globins, the pentagon pentacoordinate globins. N- and C-terminal extensions are indicated by bars, the acylation of the N-terminus of GbX is shown. Note the duplicated *GbX* genes in *L. chalumnae* and the duplicated *Cygb* genes in the teleosts. Globin sequences were identified in representative vertebrate genomes employing the BLAST algorithm. The genomes of man (*Homo sapiens*, build 37.3), mouse (*Mus musculus* build 37.2), opossum (*Monodelphis domestica*, build 2.2), chicken (*Gallus gallus*, build 2.1), zebra finch (*Taeniopygia guttata*, build 1.1) and zebrafish (*Danio rerio*, Zv9) were obtained from the NCBI web site at http://www.ncbi.nlm.nih.gov/projects/mapview/. The genome data from the coelacanth (*Latimeria chalumnae*, LatCha1), platypus (*Ornithorhynchus anatinus*, OANA5), anole (*Anolis carolinensis*, AnoCar2.0), clawed frog (*Xenopus tropicalis*, JGI_4.2), pufferfish (*Tetraodon nigroviridis*, TETRAODON7) and lamprey (*Petromyzon marinus*, Pmarinus_7.0) derive from

http://www.ensembl.org. The elephant shark (***Callorhinchus milii***) genome data sequences were obtained from http://esharkgenome.imcb.a-star.edu.sg/. Additional information derives from BLAST searches of the non-redundant nucleotide and ESTs databases.

Supplementary Tables

Supplementary Table 1. Physical coverage of human and coelacanth datasets

Species	Library Name	Mean Fragment Size	Usable Pairs	Physical Coverage
Human	Solexa-30824	2471	131513259	101.6
	Solexa-30807	2987	146165016	136.4
			Jump Total	*238.0*
	Solexa-22993	32615	1254194	12.8
	Solexa-21447	32823	2466369	25.3
			Fosmid Total	*38.1*
Coelacanth	Solexa-31766	1795	26983978	16.7
	Solexa-31767	2322	26588398	21.3
	Solexa-31768	2100	26872930	19.5
	Solexa-35322	2000	16805781	11.6
	Solexa-35350	2524	13452116	11.7
	Solexa-35377	2815	6271657	6.1
			Jump Total	*86.8*
	Solexa-63288	36574	1338279	16.9
	Solexa-64284	36866	323081	4.1
			Fosmid Total	*21.0*

Supplementary Table 2. Recalcitrant sequences in the human and coelacanth genomes

Motif	Recalcitrant sequences (per Mb)	
	Coelacanth	Human
A^20 or T^20	71.0	29.2
(AT)^10	19.9	2.7
G^10 or C^10	205.4	2.0

Supplementary Table 3: Number of protein-coding genes in human, mouse and zebrafish with 1:1 orthologues in coelacanth and flanked by a lncRNA

	lncRNA	Genes with lncRNA in coelacanth and 1:1 ortholog	Proteins with 1:1 ortholog in coelacanth	Protein-coding genes flanking lncRNA	Protein-coding genes flanking lncRNA with ortholog	Proteins with 1:1 ortholog and flanked by lncRNA in both species
human	9,794	1,121	12,317	7,422	4,769	520
mouse	2,376	1,108	12,227	3,427	2,292	242
zebrafish	1,656	855	9,677	2,447	1,014	102

Supplementary Table 4: Comparison of the automatic and manual annotation for the small ncRNAs in *L. chalumnae*.

Class	Manual	Ensembl
miRNA	249	515
snoRNA	246	182
tRNA	676	NA
snRNA	147	139
YRNA	3	4
RnaseP	3	3
RnaseMRP	1	1
Vault	1	1
Conserved Structured Elements (RNAz)	24,045	NA

Supplementary Table 5 : Statistically enriched GO Terms, Domains and P-values from a Hypergeometric test (Bonferroni corrected) for 111 *L. chalumnae* gene pairs.

GO Accession	GO Domain	GO Term Name	P-Value
GO:0004984	Molecular Function	olfactory receptor activity	2.38E-20
GO:0004842	Molecular Function	ubiquitin-protein ligase activity	1.76E-12
GO:0004866	Molecular Function	endopeptidase inhibitor activity	7.58E-07
GO:0004930	Molecular Function	G-protein coupled receptor activity	1.07E-42
GO:0030246	Molecular Function	carbohydrate binding	1.00E-08
GO:0019882	Biological Process	antigen processing and presentation	9.37E-06
GO:0006955	Biological Process	immune response	1.14E-06
GO:0050909	Biological Process	sensory perception of taste	1.48E-27
GO:0016567	Biological Process	protein ubiquitination	1.72E-13
GO:0007186	Biological Process	G-protein coupled receptor signaling pathway	1.98E-41
GO:0015074	Biological Process	DNA integration	4.32E-04
GO:0016021	Cellular Component	integral to membrane	1.32E-18
GO:0000151	Cellular Component	ubiquitin ligase complex	1.92E-14
GO:0042613	Cellular Component	MHC class II protein complex	8.87E-03
GO:0005615	Cellular Component	extracellular space	1.87E-02

Supplementary Table 6: Statistically enriched InterPro terms, domain annotations and P-values from a Hypergeometric test (Bonferroni corrected) for 166 *L. chalumnae* gene pairs

InterPro ID	InterPro Domain Description	Bonferroni Adjusted P-value
IPR007960	TAS2_rcpt	5.05E-29
IPR017452	GPCR_Rhodpsn_supfam	8.04E-28
IPR000276	7TM_GPCR_Rhodpsn	1.59E-23
IPR011500	GPCR_3_9-Cys_dom	2.49E-22
IPR000337	GPCR_3	3.59E-22
IPR004073	GPCR_3_vmron_rcpt_2	6.20E-22
IPR000725	Olfact_rcpt	2.28E-21
IPR007110	Ig-like	7.93E-20
IPR003879	Butyrophylin	2.95E-19
IPR001828	ANF_lig-bd_rcpt	4.17E-19
IPR017978	GPCR_3_C	4.79E-19
IPR000315	Znf_B-box	9.44E-19
IPR013106	Ig_V-set	1.96E-16
IPR003877	SPRY_rcpt	5.37E-16
IPR001870	B30.2/SPRY	8.97E-16
IPR003613	Ubox_domain	2.94E-15
IPR001304	C-type_lectin	3.11E-12
IPR003597	Ig_C1-set	1.99E-10
IPR018957	Znf_C3HC4_RING-type	2.51E-10
IPR002353	AntifreezeII	3.54E-09
IPR011625	A2M_N_2	1.68E-07
IPR007111	NACHT_NTPase	7.52E-07

IPR001599	Macroglobln_a2	3.35E-06
IPR001841	Znf_RING	3.87E-06
IPR009048	A-macroglobulin_rcpt-bd	8.93E-06
IPR011626	A2M_comp	8.93E-06
IPR019424	7TM_GPCR_serpentine_rcpt_Srsx	4.94E-05
IPR019565	MacroglobInA2_thiol-ester-bond	6.43E-05
IPR001584	Integrase_cat-core	8.85E-05
IPR013151	Immunoglobulin	1.77E-04
IPR013162	CD80_C2-set	2.52E-04
IPR004020	Pyrin	4.54E-04
IPR007990	SV_autoAg	6.27E-04
IPR001604	DNA/RNA_non-sp_Endonuclease	6.43E-04
IPR002890	A2M_N	2.56E-03
IPR006612	Znf_C2CH	1.47E-02
IPR001613	Flavin_amine_oxidase	2.01E-02
IPR008906	HATC	4.06E-02

Supplementary Table 7 – Repeat content of selected vertebrate genomes

Species name	Genome size (Gb)	Interspersed repeat content (%)	Most abundant TE in each genome	Reference
Human	2.9	45	L1/LINE	[217]
Mouse	2.5	38	L1/LINE	[218]
Chicken	1.1	8.6	CR1/LINE	[219]
Dog	2.4	34	L1/LINE	[220]
Lizard	1.8	30	DNA transposon	[151]
Frog	3.1	35	hAT/DNA transposon	[221]
Stickleback	0.46	25	NA	[222]
Fugu	0.33	2.7	Maui/LINE	[152]
Medaka	0.70	16	DNA transposon	[223]
Cod	0.75	18	DNA transposon	[10]

Supplementary Table 8 - Frequencies and fractions of repeats in the genome of *L. chalumnae*.

Transposable element families	Number of copies	Total length	Percentage
DNA Transposon	**19,437**	**5,219,790**	**1.65**
other	442	36,188	0.00
TcMar	4,427	1,129,275	0.04
P	74	21,508	0.00
Sola	2,234	597,025	0.02
Harbinger LatiHarb1	1,793	427,243	0.02 1.45
Kolobok-T2	191	86,298	0.00
hAT	10,276	2,922,253	0.11
LTR Retrotransposon	**74,906**	**22,963,535**	**0.86**
Other	16	2,265	0.00
ERVK	284	200,271	0.01
ERV1	41,559	4,969,865	0.19
Gypsy	5,985	2,175,944	0.08
DIRS	27,062	15,615,190	0.58
Non-LTR Retrotransposon	**597,715**	**169,291,067**	**6.34**
Penelope	14,292	4,116,402	0.15
Tx1	791	429,955	0.02
L2	96,638	34,565,553	1.30
L1	28,830	13,734,834	0.51
RTE	34,890	10,647,255	0.40

R4	265	108,596	0.00
CR1	192,193	55,792,081	2.09
Deu	215,151	48,445,166	1.82
SINE	14,531	1,432,326	0.05
MIR	134	18,899	0.00
rRNA	**13,207**	**2,281,385**	**0.09**
rRNA	13,207	2,281,385	0.09
tRNA	**19,865**	**3,558,067**	**0.13**
tRNA	19,865	3,558,067	0.13
Simple_repeat	**297,493**	**28,972,629**	**1.09**
Simple_repeat	297,493	28,972,629	1.09
Low_complexity	**730,965**	**105,282,913**	**3.95**
other	730,965	105,282,913	3.95
Unclassified	**362,241,939**	**362,241,939**	**13.6**
Total	**363,995,527**	**699,811,325**	**27.71**

Supplementary Table 9 - Copy number of repeats and TE families in the genome of *L. chalumnae*. Pseudogenes are non-coding RNA and Unknown are unclassified repeat sequences. The reference size is the size of each element in the library. The number of copies for each element that make 30%, 50% and 80% of the reference size were counted.

	Total copy number	Copy number (30% of the ref size)	Copy number (50% of the ref size)	Copy number (80% of the ref size)
Pseudogenes	27,451	25,102	18,381	14,358
Unkwown	1,852,161	1,666,547	1,448,251	1,045,996
SINE	593,052	483,009	392,853	272,564
LINE				
LINE1	27,790	23,288	19,839	14,075
Tx1	6,672	5,289	3,904	2,280
Jockey	1,873	1,041	699	377
CR1	259,925	142,597	102,112	51,893
RTE	33,825	18,830	12,583	7,159
R2	46	6	3	2
R4	422	354	297	195
Penelope	13,323	7,559	5,149	2,659
LTR				
Gypsy	44,452	5,857	5,434	4,516
DIRS	32,309	18,479	14,822	9,822
DNA				
TcMar	7,771	6,367	5,239	4,136
hAT	27,748	14,025	10,901	8,240

Harbinger	2,483	2,321	1,932	1,501
Kolobok	171	160	144	107
MITE	66,851	66,848	62,731	41,419
Helitron	1,062	445	268	115
Polinton	78,977	3	1	1

Supplementary Table 10. Number of active TE families and sequences in the coelacanth genome.

	Superfamily name	No. Family based on sequence similarity	No. Sequences based on sequence similarity	No. Family based on RNA-seq	No. Sequence in RNA-seq
Non-LTR retrotransposon	Penelope	9	199	9	64
	L1	1	1	2	3
	RTE	40	1885	31	123
	Tx1	1	4	1	11
	SINE	1	2	1	20
	R4	3	19	2	3
	CR1	24	1278	15	53
LTR retrotransposon	DIRS	1	3	2	34
	Gypsy	5	5	5	7
DNA transposon	Sola	1	3	1	33
	Helitron	1	173	1	1
	hAT	35	314	19	43
	Harbinger	0	0	1	3
	P	1	2	0	0
	TcMar	14	106	5	15

Supplementary Table 11 - Summary of Megablast alignments of available *Latimeria menadoensis* BAC sequences to orthologous *Latimeria chalumnae* scaffolds. The analysis revealed that an average of 72% of sequences were aligned with 96%-99% nucleotide identity in the aligned regions, with the average across all alignments being 98.7%. Unaligned regions were primarily accounted for by runs of Ns in the *L. chalumnae* scaffolds, or the alignment running off the end of a scaffold.

BAC ID	Accession Number	Lm BAC Insert Length (bp)	Lc Scaffold ID	Scaffold start	Scaffold end	N count (bp)	Excluded Region (bp)	Aligned sequence (bp)	Identical match (bp)
IgW-BAC1		189,785	01694	350,000	246,000	58,245	0	137,580 (72%)	136,271 (99%)
IgW-BAC2		168,849	00354	55,000	230,000	24,421	0	133,184 (79%)	131,908 (99%)
HOX-A1	FJ497005.1	319,360	00150	80,000	400,000	83,978	0	204,223 (64%)	203,151 (99%)
HOX-B1	FJ497006.1	373,046	00056	60,000	440,000	73,437	15,000	270,452 (72%)	268,050 (99%)
HOX-C1	FJ497007.1	403,307	00268	1,450,000	1,760,000	7,525	120,000	239,860 (59%)	235,719 (98%)
HOX-D1	FJ497008.1	517,039	00623	200,000	750,000	61,275	0	39,846 (76%)	389,352 (98%)
VMRC4-40C18	AC150310.1	170,774	01558	240,000	420,000	22,414	0	137,328 (80%)	136,868 (99%)
VMRC4-24C12	AC150308.1	150,025	00254	100,000	300,000	44,131	0	125,817 (84%)	121,344 (96%)
VMRC4-188C23	AC150283.1	171,414	01558	390,000	570,000	36,581	0	130,381 (76%)	128,874 (98%)
VMRC4-39G19	AC150309.1	196,157	00254	0	200,000	15,036	0	160,381 (82%)	157,842 (98%)
VMRC4-44H8	AC150284.1	187,607	01558	410,000	570,000	35,101	35,000	116,270 (62%)	114,446 (98%)
VMRC4-121C4	AC215495.4	166,048	00118	2,025,000	2,250,000	37,131	0	117,044 (70%)	115,164 (98%)
VMRC4-97N21	AC216641.6	163,500	01950	20,000	190,000	68,696	0	100,766 (61%)	99,857 (99%)
VMRC4-73N21	AC215493.2	159,151	00606	920,000	1,075,000	4,641	0	143,914 (90%)	142,424 (98%)
VMRC4-37P20	AC216956.3	176,031	01681	110,000	300,000	40,056	0	126,967 (72%)	123,557 (97%)
VMRC4-133N21	AC217916.7	183,432	01303	360,000	560,000	43,962	0	150,081 (82%)	147,699 (98%)
VMRC4-121N21	AC216642.5	163,289	01377	350,000	540,000	64,810	0	116,648 (71%)	114,509 (98%)
VMRC4-109C4	AC215494.8	171,002	00744	420,000	620,000	28,399	0	135,509 (79%)	135,088 (99%)
VMRC4-85N21	AC218927.3	158,654	00705	700,000	900,000	97,373	0	87,883 (55%)	86,698 (98%)
VMRC4-61P20	AC218030.4	159,380	00155	50,000	225,000	13,933	0	135,732 (85%)	13,444 (99%)
VMRC4-13P20	AC215902.3	176,114	00568	500,000	675,000	30,949	0	132,517 (75%)	131,427 (99%)
VMRC4-49D2	AC215983.8	162,516	00739	25,000	190,000	41,609	0	106,908 (66%)	105,970 (99%)
VMRC4-25P20	AC218926.1	163,030	01893	350,000	455,000	45,574	50,000	57,598 (35%)	56,486 (98%)
VMRC4-25C4	AC215903.2	167,414	01958	300,000	432,348	30,103	40,000	109,852 (66%)	106,313 (96%)
VMRC4-25N21	AC215904.2	160,991	00130	375,000	550,000	26,941	0	140,465 (87%)	139,549 (99%)
VMRC4-66G11$_2$	EU284132.1	166,067	01284	420,000	600,000	35,209	0	139,582 (84%)	137,876 (98%)
Total		5,343,982				1,071,530	260,000	3,850,788 (72%)	3,800,888 (99%)

Supplementary Table 12 – Homeodomain genes in vertebrate species

homeodomain gene class	human	coelacanth	Amphioxus	zebrafish
ANTP class	100	106	60	132
PRD class	51	39	29	49
LIM class	12	14	7	20
POU class	16	17	7	20
HNF class	3	3	3	6
SINE class	6	7	3	13
TALE class	16	15	8	29
CUT class	5	6	4	7
PROS class	2	2	1	3
ZF class	14	15	5	17
CERS class	5	4	1	3
total:	230	228	128	299

Supplementary Table 13: Coelacanth Hsp70 and Hsp40 proteins

Chaperone Category	**Genomic location**	**Putative Type in humans** (human gene ID; protein accession number); common name; localization/expression	**Comments and predictions on localization and function**
Hsp70			
	ENSLACG00000012423 ENSLACT00000014214 ENSLACP00000014115 JH127214.1:715287-717212 1	HSPA1A (3303 NP_005336.3); Hsp72 or HSPA1B (3304); Hsp70-2; Cytosolic; inducible;	Identical (100%) to published LmHsp70 (EU016555; Modisakeng et al., 2009); and almost identical (99%) to published LcHsp70 (AY929184; Modisakeng et al., 2004); C-terminal EEVD for Hop interaction; cytosolic?
	ENSLACG00000014762 ENSLACT00000016873 ENSLACP00000016755 JH126993.1:1094431-1096356 -1	HSPA1L (3305; NP_005518.3); hum70t; Cytosolic; inducible; highly expressed in testis	C-terminal EEVD for Hop interaction; cytosolic; inducible
	?	HSPA2 (3306; NP_068814.2); Hsp70B'; Cytosolic; inducible; highly expressed in testis	
	ENSLACG00000018251 ENSLACT00000020915 ENSLACP00000020775 JH126572.1:2942341-2948653 1	HSPA5 (3309; NP_005338.1); BiP; ER	C-terminal ER retention signal (KDEL); ER
	?	HSPA6 (3310; NP_002146.2); cytosolic;	

		inducible	
	?	HSPA7 (3311; P48741.2); pseudogene	
	ENSLACG00000001177 ENSLACT00000001324 ENSLACP00000001312 JH129310.1:15420-22841 -1	HSPA8 (3312; NP_006588.1); Hsc70; cytosolic; constitutive	C-terminal EEVD for Hop interaction; cytosolic; constitutive
	ENSLACG00000009096 ENSLACT00000010406 ENSLACP00000010328 JH127522.1:397401-419559 -1	HSPA9 (3313; NP_004125.3); Grp75; mitochondrial	mitochondrial
	ENSLACG00000005550 ENSLACT00000006309 ENSLACP00000006257 JH126834.1:173038-256308 -1	HSPA12A (259217; NP_079291.2); expressed specifically in the brain	
	ENSLACG00000006455 ENSLACT00000007337 ENSLACP00000007277 JH128823.1:219127-314124 1	HSPA12B (116835; NP_443202.3)	
	ENSLACG00000008053 ENSLACT00000009193 ENSLACP00000009124 JH128206.1:317368-326125 1	HSPA13 (6782; NP_008879.3); microsomal associated	Microsomal associated
	ENSLACG00000010343 ENSLACT00000011840 ENSLACP00000011750 JH128184.1:494599-516681 1	HSPA14 (51182; NP_057383.2)	
Hsp40/DnaJ			
Type I/ DnaJA	HUMHOMP00000369127_1	DnaJA1	No CxxCxGxG repeats
	HUMHOMP00000314030_1	DnaJA2	
	HUMHOMP00000262375_1	DnaJA3	Mitochondrial

	HUMHOMP00000378324_1	DnaJA4	
Type II/	HUMHOMP00000254322_1	DnaJB1	
DnaJB	HUMHOMP00000338019_1	DnaJB2	
	DARHOMP00000056895_1	DnaJB4	
	HUMHOMP00000413684_1	DnaJB5	
	Scaffold JH131963.1: 45284-45391	DnaJB6	J domain only
	GACHOMP00000005437_1	DnaJB9	ER; ERdj4 homologue
	HUMHOMP00000414398_1	DnaJB11	ER; ERdj3 homologue
	HUMHOMP00000345575_1	DnaJB12	
		DnaJB13	
	HUMHOMP00000404381_1	DnaJB14	
Type III/	HUMHOMP00000366179_1	DnaJC1	ER; ERdj1 homologue
DnaJC		DnaJC2	Zoutin homologue
	HUMHOMP00000365991_1	DnaJC3	ER; TPR domains; $p58^{IPK}$ homologue
	HUMHOMP00000354111_1	DnaJC5	
		DnaJC5b	
		DnaJC7	
		DnaJC8	
	GACHOMP00000013582_1	DnaJC9	
	HUMHOMP00000264065_1	DnaJC10	ER; ERdj5 homologue
	DARHOMP00000026754_1	DnaJC11	
		DnaJC12	

	HUMHOMP00000344431_1	DnaJC13	
	GACHOMP00000006335_1	DnaJC13	
		DnaJC14	
	HUMHOMP00000447000_1	DnaJC14	
		DnaJC15	
	DARHOMP00000080290_1	DnaJC16	TRX domains
	GACHOMP00000027323_1	DnaJC18	
	HUMHOMP00000371451_1	DnaJC21	
	HUMHOMP00000446830_1	DnaJC22	
	Scaffold JH126688.1: 1042344-1134726 -1	DnaJC23	ER; ERdj2/Sec63 homologue
	DARHOMP00000037632_1	DnaJC24	J domain only
	HUMHOMP00000320650_1	DnaJC25	
	HUMHOMP00000264711_1	DnaJC27	
	HUMHOMP00000371373_1	DnaJC28	
	HUMHOMP00000378605_1	DnaJC30	

Supplementary Table 14 - bHLH-PAS Gene Family

Subfamily		*Coelacanth*	*Zebrafish*	*Human*
AHR	AHR1	2	2	1
	AHR2	1	1	0
	AHRx	1	0	0
	AHRR	1	2	1
HIF	HIF1	1	3	1
	HIF2	2	2	1
	HIF3	1	1	1
SIM	SIM1	1	2	1
	SIM2	1	1	1
ARNT	ARNT1	1	1	1
	ARNT2	1	1	1
BMAL	BMAL1	1	2	1
	BMAL2	1	1	1
PER	PER1	1	2	1
	PER2	1	1	1
	PER3	1	1	1
CLOCK	CLOCK1	1	1	1
	CLOCK2	2	1	1
	CLOCK3	0	1	0
	PASD1	0	0	1
NPAS	NPAS1	1	1	1

	NPAS3	1	1	1
NCOA	NCOA1	1	1	1
	NCOA2	1	1	1
	NCOA3	1	1	1
NXF	NXF	2	3	1

Numbers of genes in each subfamily corresponding to (putative orthologs of) human and zebrafish PAS genes.

Supplementary Table 15 – Tetrapod and sarcopterygian-specific proteins

Ensembl gene ID	CDS (aa)	phylogenetic destribution	coelacanth gene ID	Lungfish transcript ID	description
ENSACAG00000003944	289	tetrapoda	-	-	-
ENSGALG00000001451	173	tetrapoda + coalacanth	ENSLACG00000016679	-	uncharacterized protein
ENSGALG00000009622	306	tetrapoda	-	-	uncharacterized protein
ENSMUSG00000074300	71	tetrapoda + lungfish	-	comp24360_c0_seq1 len=604 path=[0:0-603]	cDNA sequence BC030870
ENSOANG00000005335	314	tetrapoda + lungfish + coelacanth	ENSLACG00000000680	comp3474_c0_seq2 len=2072 path=[2185:0-144 200:145-868 924:869-871 927:872-1973 2029:1974-2071]	uncharacterized protein
ENSG00000111644	544	tetrapoda	-	-	acrosin binding protein
ENSG00000121314	310	tetrapoda + coelacanth	contig217410	-	taste receptor, type 2, member 8
ENSG00000139971	774	tetrapoda + lungfish	-	comp16060_c0_seq1 len=528 path=[3:0-527]	chromosome 14 open reading frame 37
ENSG00000146857	330	tetrapoda	-	-	stimulated by retinoic acid gene 8

					homolog (mouse)
ENSG00000154768	174	tetrapoda	-	-	chromosome 17 open reading frame 50
ENSG00000163519	187	tetrapoda	-	-	T cell receptor associated transmembrane adaptor 1
ENSG00000164106	99	tetrapoda	-	-	stimulator of chondrogenesis 1
ENSG00000178821	210	tetrapoda	-	-	transmembrane protein 52
ENSG00000166012	279	tetrapoda	-	-	TATA box binding protein (TBP)-associated factor, RNA polymerase I, D, 41kDa
ENSG00000186329	195	tetrapoda	-	-	transmembrane protein 212
ENSG00000188133	236	tetrapoda + coelacanth	ENSLACG00000007107	-	transmembrane protein 215
ENSG00000125531	319	tetrapoda + coelacanth	ENSLACG00000007305	-	chromosome 20 open reading frame 195
ENSG00000064787	615	tetrapoda	-	-	breast carcinoma amplified sequence 1
ENSG00000205208	114	tetrapoda	-	-	chromosome 4 open reading frame 46
ENSG00000214097	215	tetrapoda	-	-	chromosome 3 open reading frame 43

ENSG00000181143	14508	tetrapoda + coelacanth	ENSLACG00000000418	-	mucin 16, cell surface associated
ENSG00000214688	168	tetrapoda + lungfish	-	comp11638_c0_seq1 len=1636 path=[0:0-55 56:56-1635]	chromosome 10 open reading frame 105
ENSG00000214128	158	tetrapoda + lungfish	-	comp1016_c0_seq1 len=1613 path=[0:0-1446 1447:1447-1612]	transmembrane protein 213
ENSG00000157111	325	tetrapoda + coelacanth	ENSLACG00000009377	-	transmembrane protein 171
ENSG00000188817	166	tetrapoda	-	-	sentan, cilia apical structure protein
ENSG00000163705	178	tetrapoda + coelacanth	ENSLACG00000018101	-	chromosome 3 open reading frame 24
ENSG00000142698	599	tetrapoda + coelacanth	ENSLACG00000003013	-	chromosome 1 open reading frame 94
ENSG00000246922	244	tetrapoda	-	-	ubiquitin associated protein 1-like
ENSTGUG00000003796	114	tetrapoda	-	-	uncharacterized protein
ENSTGUG00000006057	194	tetrapoda	-	-	uncharacterized protein

The 30 gene IDs which correspond to the identified tetrapod- and sarcopterygian-specific peptides are listed together with amino acid length (CDS) of the deduced proteins, phylogenetic distribution, gene IDs of coelacanth and lungfish orthologs and description (derived from Ensembl) of the gene provided by Ensembl are shown

Supplementary Table 16 – Tetrapod and sarcopterygian gene loss

Ensembl gene ID	CDS (aa)	phylogenetic destribution	coelacanth gene ID	lungfish transcript ID	description	bit score
ENSDARG00000021849	693	non-sarcopterygians	-	-	peptide-n4-n-acetyl-beta-d-glucosaminylasparagine amidase f precursor	40.8
ENSDARG00000044048	607	non-sarcopterygians	-	-	prion protein 1	41.6
ENSDARG00000056650	513	non-sarcopterygians	-	-	fad-dependent pyridine nucleotide-disulphide oxidoreductase	36.2
ENSDARG00000070800	769	non-sarcopterygians	-	-		35.4
ENSDARG00000058248	416	non-sarcopterygians	-	-	im:6912380 partial	48.9
ENSDARG00000068098	410	non-sarcopterygians	-	-	si:ch211- protein	37.7
ENSDARG00000070604	693	non-tetrapods	ENSLACG00000015344	comp13486_c0_seq1 len=2437 path=[0:0-99 2431:100-103 100:104-2436]	zgc:162509 protein	32.0
ENSDARG00000079532	211	non-sarcopterygians	-	-	methyltransferase type 11	39.7
ENSDARG00000090688	523	non-sarcopterygians + coelacanth	ENSLACG00000008449	-	notochord-related protein	34.7

ENSDARG00000091049	341	non-sarcopterygians	-	-	PREDICTED: hypothetical protein LOC100534841 [Danio rerio]	42.4
ENSDARG00000086181	938	non-sarcopterygians	-	-	zinc finger -60	39.7
ENSDARG00000088444	301	non-sarcopterygians + lungfish	-	comp27163_c0_seq1 len=1643 path=[0:0-1642]	selenoprotein l	42.0
ENSDARG00000093821	233	non-tetrapods	ENSLACG00000018081	comp25044_c0_seq1 len=620 path=[0:0-619]	novel protein	33.5
ENSDARG00000077519	340	non-sarcopterygians + lungfish	-	comp10729_c0_seq1 len=1166 path=[0:0-533 534:534-1165]	uncharacterized transposase-like protein	35.4
ENSDARG00000093126	1026	non-sarcopterygians	-	-	novel protein	35.8
ENSDARG00000095580	574	non-sarcopterygians + coelacanth	ENSLACG00000017893	-	protein	37.4
ENSDARG00000093998	661	non-sarcopterygians	-	-	novel protein	38.9
ENSDARG00000093042	185	non-sarcopterygians	-	-	novel protein	31.6
ENSDARG00000095821	385	non-sarcopterygians	-	-	alpha beta hydrolase fold protein	39.7
ENSGACG00000002069	327	non-sarcopterygians + coelacanth	ENSLACG00000007512	-	crystallin j1a	35.8
ENSGACG00000003509	371	non-sarcopterygians	ENSLACG0000000	-	malate dehydrogenase	35.0

		+ coelacanth	8801			
ENSGACG00000004518	440	non-sarcopterygians	-	-	xylose isomerase	32.3
ENSGACG00000006012	212	non-sarcopterygians	-	-	unnamed protein product [Tetraodon nigroviridis]	33.5
ENSGACG00000008053	302	non-sarcopterygians	-	-	diadenosine tetraphosphate hydrolase	39.3
ENSGACG00000009610	334	non-sarcopterygians	-	-	laminin alpha 5 chain	39.7
ENSGACG00000014241	299	non-sarcopterygians + lungfish	-	comp21205_c0_seq1 len=945 path=[0:0-944]	pyroglutamyl-peptidase 1	33.1
ENSGACG00000015054	183	non-sarcopterygians	-	-	conserved hypothetical protein [Pediculus humanus corporis]	33.9
ENSGACG00000016664	276	non-sarcopterygians	-	-	collagen alpha-1 chain-like	39.7
ENSGACG00000019286	249	non-sarcopterygians + lungfish	-	comp4497_c0_seq1 len=3584 path=[0:0-248 249:249-3583]	PREDICTED: hypothetical protein LOC569091 [Danio rerio]	43.1
ENSORLG00000000443	128	non-sarcopterygians + lungfish	-	comp7405_c0_seq1 len=894 path=[0:0-725 726:726-893]	glutathione s-transferase	41.7
ENSORLG00000001152	182	non-sarcopterygians	-	-	adp-ribosylation crystallin j1	34.7
ENSORLG00000001388	100	non-sarcopterygians	-	-	zinc dhhc-type containing 3-like	41.6
ENSORLG00000002653	185	non-sarcopterygians	-	-	PREDICTED: hypothetical protein	32.7

					LOC100537439 [Danio rerio]	
ENSORLG00000003467	109	non-sarcopterygians + lungfish	-	comp29913_c0_seq1 len=693 path=[0:0-692]	aldo keto reductase family protein	34.7
ENSORLG00000004708	219	non-sarcopterygians	-	-	conserved hypothetical protein [Pediculus humanus corporis]	34.3
ENSORLG00000009572	217	non-sarcopterygians	-	-	Uncharacterized protein [Dicentrarchus labrax]	32.0
ENSORLG00000014401	202	non-sarcopterygians	-	-	hypothetical protein BRAFLDRAFT_79043 [Branchiostoma floridae]	36.6
ENSORLG00000017706	161	non-sarcopterygians	-	-	unnamed protein product [Tetraodon nigroviridis]	29.6
ENSTNIG00000000374	261	non-sarcopterygians	-	-	conserved hypothetical protein [Pediculus humanus corporis]	32.0
ENSTRUG00000000905	189	non-sarcopterygians	-	-	pro-pol polyprotein	46.2

The 40 gene IDs which correspond to the identified genes absent from tetrapod genomes are listed together with amino acid length (CDS) of the deduced proteins, phylogenetic distribution, gene IDs of coelacanth and lungfish orthologs, description (derived from blast2go) of the gene provided by Ensembl are shown and bit scores of blastp searches against sarcopterygians (taxid:8287).

Supplementary Table 17. Numbers of CNEs that originated in different lineages.

Ancestor	Number of CNEs	Total length of CNEs (bp)
Osteichthyes	29,268	3,998,819
Sarcopterygii	53,985	8,148,181
Tetrapods	44,200	6,161,506
Amniotes	92,263	10,253,282
Theria	254,019	25,820,535
Eutheria	225,958	22,755,844
Boreoeutheria	29,755	2,965,269
Euarchontoglires	10,149	931,843

Supplementary Table 18. Functional enrichment of tetrapod CNEs using GREAT binomial test over genomic regions – GO "biological process"

No.	Term ID	Description	No. of observed regions	Fold Enrichment	Bonferroni P-Value
1	GO:0050911	detection of chemical stimulus involved in sensory perception of smell	1,527	4.89	<1E-300
2	GO:0007608	sensory perception of smell	1,579	4.00	<1E-300
3	GO:0009593	detection of chemical stimulus	1,657	3.67	<1E-300
4	GO:0050907	detection of chemical stimulus involved in sensory perception	1,545	4.52	<1E-300
5	GO:0050906	detection of stimulus involved in sensory perception	1,617	3.69	<1E-300
6	GO:0007606	sensory perception of chemical stimulus	1,624	3.57	<1E-300
7	GO:0051606	detection of stimulus	1,904	2.65	2.62E-296
8	GO:0060850	regulation of transcription involved in cell fate commitment	250	3.79	3.43E-63
9	GO:0009956	radial pattern formation	137	4.84	7.38E-45
10	GO:0035137	hindlimb morphogenesis	419	2.20	1.73E-42
11	GO:0001945	lymph vessel development	224	3.08	5.57E-42
12	GO:0060839	endothelial cell fate commitment	155	4.05	9.72E-42
13	GO:0060849	regulation of transcription involved in lymphatic endothelial cell fate commitment	154	4.08	1.02E-41
14	GO:0060836	lymphatic endothelial cell differentiation	155	4.04	1.51E-41

15	GO:0072148	epithelial cell fate commitment	159	3.85	3.79E-40
16	GO:0021517	ventral spinal cord development	353	2.30	1.42E-39
17	GO:0060993	kidney morphogenesis	444	2.04	1.29E-37
18	GO:0048644	muscle organ morphogenesis	385	2.14	1.91E-36
19	GO:0035136	forelimb morphogenesis	392	2.10	1.32E-35
20	GO:0048745	smooth muscle tissue development	253	2.44	1.85E-31
21	GO:0021515	cell differentiation in spinal cord	348	2.10	3.88E-31
22	GO:0061005	cell differentiation involved in kidney development	292	2.20	5.56E-29
23	GO:0072189	ureter development	184	2.75	3.69E-28
24	GO:0048665	neuron fate specification	304	2.13	4.58E-28
25	GO:0045662	negative regulation of myoblast differentiation	192	2.65	1.53E-27
26	GO:0060415	muscle tissue morphogenesis	333	2.01	1.54E-26
27	GO:0051148	negative regulation of muscle cell differentiation	252	2.25	3.70E-26
28	GO:0021889	olfactory bulb interneuron differentiation	303	2.08	3.73E-26
29	GO:0072028	nephron morphogenesis	299	2.09	3.74E-26
30	GO:0048880	sensory system development	193	2.56	7.71E-26
31	GO:0072088	nephron epithelium morphogenesis	292	2.09	1.05E-25
32	GO:0072105	ureteric peristalsis	76	4.99	1.08E-24
33	GO:0072195	kidney smooth muscle cell differentiation	76	4.99	1.08E-24
34	GO:0050975	sensory perception of touch	76	4.85	5.74E-24

35	GO:0021522	spinal cord motor neuron differentiation	237	2.21	1.55E-23
36	GO:0072193	ureter smooth muscle cell differentiation	96	3.86	1.60E-23
37	GO:0021846	cell proliferation in forebrain	248	2.16	2.23E-23
38	GO:0010949	negative regulation of intestinal phytosterol absorption	30	17.49	6.71E-23
39	GO:0045796	negative regulation of intestinal cholesterol absorption	30	17.49	6.71E-23
40	GO:0002087	regulation of respiratory gaseous exchange by neurological system process	135	2.92	2.50E-22
41	GO:0043576	regulation of respiratory gaseous exchange	176	2.47	7.66E-22
42	GO:0045736	negative regulation of cyclin-dependent protein kinase activity	149	2.70	1.24E-21
43	GO:0072194	kidney smooth muscle tissue development	94	3.68	1.52E-21
44	GO:0044065	regulation of respiratory system process	141	2.78	1.90E-21
45	GO:0007386	compartment pattern specification	61	5.43	3.15E-21
46	GO:0021912	regulation of transcription from RNA polymerase II promoter involved in spinal cord motor neuron fate specification	71	4.52	1.55E-20
47	GO:0045879	negative regulation of smoothened signaling pathway	186	2.30	9.71E-20
48	GO:0001937	negative regulation of endothelial cell proliferation	186	2.29	1.25E-19
49	GO:2000794	regulation of epithelial cell proliferation involved in lung	141	2.63	2.11E-19

		morphogenesis			
50	GO:0021913	regulation of transcription from RNA polymerase II promoter involved in ventral spinal cord interneuron specification	77	4.00	2.86E-19
51	GO:0055010	ventricular cardiac muscle tissue morphogenesis	219	2.11	3.33E-19
52	GO:0045109	intermediate filament organization	113	2.99	3.54E-19
53	GO:0060710	chorio-allantoic fusion	97	3.27	9.42E-19
54	GO:0021520	spinal cord motor neuron cell fate specification	111	2.91	6.05E-18
55	GO:0003215	cardiac right ventricle morphogenesis	202	2.12	1.01E-17
56	GO:0003229	ventricular cardiac muscle tissue development	225	2.02	1.96E-17
57	GO:0001823	mesonephros development	201	2.10	3.52E-17
58	GO:0034260	negative regulation of GTPase activity	168	2.27	3.72E-17
59	GO:0051152	positive regulation of smooth muscle cell differentiation	111	2.80	1.13E-16
60	GO:0021871	forebrain regionalization	196	2.10	1.30E-16
61	GO:0003148	outflow tract septum morphogenesis	131	2.53	1.60E-16
62	GO:0045103	intermediate filament-based process	148	2.36	3.78E-16
63	GO:0045104	intermediate filament cytoskeleton organization	145	2.36	8.76E-16
64	GO:0060579	ventral spinal cord interneuron fate commitment	136	2.43	9.54E-16

65	GO:0021521	ventral spinal cord interneuron specification	131	2.45	2.73E-15
66	GO:0021514	ventral spinal cord interneuron differentiation	137	2.39	2.95E-15
67	GO:0060501	positive regulation of epithelial cell proliferation involved in lung morphogenesis	112	2.64	4.72E-15
68	GO:0010092	specification of organ identity	175	2.11	8.19E-15
69	GO:0033152	immunoglobulin V(D)J recombination	73	3.44	1.22E-14
70	GO:0021511	spinal cord patterning	152	2.23	1.52E-14
71	GO:0050881	musculoskeletal movement	131	2.36	4.17E-14
72	GO:0021513	spinal cord dorsal/ventral patterning	139	2.28	5.72E-14
73	GO:0072078	nephron tubule morphogenesis	171	2.08	8.08E-14
74	GO:0002639	positive regulation of immunoglobulin production	65	3.60	9.93E-14
75	GO:0021516	dorsal spinal cord development	130	2.34	1.02E-13
76	GO:0043462	regulation of ATPase activity	133	2.29	2.14E-13
77	GO:0071526	semaphorin-plexin signaling pathway	156	2.13	2.62E-13
78	GO:0051961	negative regulation of nervous system development	121	2.37	4.99E-13
79	GO:0060457	negative regulation of digestive system process	30	7.60	5.66E-13
80	GO:0051964	negative regulation of synapse assembly	118	2.38	1.08E-12
81	GO:0032781	positive regulation of ATPase activity	123	2.32	1.32E-12

82	GO:0051150	regulation of smooth muscle cell differentiation	158	2.07	1.91E-12
83	GO:0060613	fat pad development	36	5.72	3.03E-12
84	GO:0045663	positive regulation of myoblast differentiation	169	2.00	3.26E-12
85	GO:0021979	hypothalamus cell differentiation	64	3.36	4.32E-12
86	GO:0072079	nephron tubule formation	149	2.09	6.05E-12
87	GO:0048672	positive regulation of collateral sprouting	70	3.11	7.58E-12
88	GO:0003139	secondary heart field specification	106	2.41	1.61E-11
89	GO:0021797	forebrain anterior/posterior pattern specification	73	2.94	3.11E-11
90	GO:0060644	mammary gland epithelial cell differentiation	128	2.17	4.07E-11
91	GO:0060413	atrial septum morphogenesis	101	2.43	4.53E-11
92	GO:0022011	myelination in peripheral nervous system	101	2.42	5.25E-11
93	GO:0072077	renal vesicle morphogenesis	149	2.02	6.95E-11
94	GO:0014044	Schwann cell development	101	2.39	1.03E-10
95	GO:0010455	positive regulation of cell fate commitment	49	3.81	1.15E-10
96	GO:0072202	cell differentiation involved in metanephros development	105	2.33	1.67E-10
97	GO:0072207	metanephric epithelium development	143	2.03	1.72E-10
98	GO:0060914	heart formation	112	2.24	2.72E-10
99	GO:0072234	metanephric nephron tubule development	136	2.06	3.27E-10

100	GO:0072070	loop of Henle development	93	2.44	3.52E-10

Supplementary Table 19. Functional enrichment of tetrapod CNEs using GREAT binomial test over genomic regions - GO "molecular function"

No.	Term ID	Description	No. of observed regions	Fold Enrichment	Bonferroni P-Value
1	GO:0004984	olfactory receptor activity	1,527	4.89	<1E-300
2	GO:0000976	transcription regulatory region sequence-specific DNA binding	836	2.08	6.21E-77
3	GO:0003706	ligand-regulated transcription factor activity	251	2.73	3.73E-39
4	GO:0001972	retinoic acid binding	155	3.29	6.66E-32
5	GO:0016524	latrotoxin receptor activity	138	3.41	9.53E-30
6	GO:0000977	RNA polymerase II regulatory region sequence-specific DNA binding	332	2.07	5.91E-29
7	GO:0005501	retinoid binding	175	2.85	7.80E-29
8	GO:0001012	RNA polymerase II regulatory region DNA binding	333	2.05	2.29E-28
9	GO:0019840	isoprenoid binding	179	2.75	1.06E-27
10	GO:0001159	core promoter proximal region DNA binding	233	2.10	9.45E-21
11	GO:0000987	core promoter proximal region sequence-specific DNA binding	228	2.10	2.86E-20
12	GO:0000978	RNA polymerase II core promoter proximal region sequence-specific DNA binding	173	2.35	1.70E-19
13	GO:0001671	ATPase activator activity	90	3.19	6.66E-17
14	GO:0050290	sphingomyelin phosphodiesterase D activity	41	5.64	1.73E-14

15	GO:0003908	methylated-DNA-[protein]-cysteine S-methyltransferase activity	55	3.88	5.58E-13
16	GO:0050682	AF-2 domain binding	81	2.88	1.17E-12
17	GO:0060590	ATPase regulator activity	94	2.58	4.68E-12
18	GO:0052658	inositol-1,4,5-trisphosphate 5-phosphatase activity	30	6.68	6.10E-12
19	GO:0052659	inositol 1,3,4,5-tetrakisphosphate 5-phosphatase activity	30	6.68	6.10E-12
20	GO:0008420	CTD phosphatase activity	31	5.96	4.27E-11
21	GO:0044323	retinoic acid-responsive element binding	53	3.40	2.97E-10
22	GO:0003708	retinoic acid receptor activity	92	2.38	7.91E-10
23	GO:0004000	adenosine deaminase activity	76	2.61	1.25E-09
24	GO:0001102	RNA polymerase II activating transcription factor binding	119	2.09	1.77E-09
25	GO:0002151	G-quadruplex RNA binding	47	3.32	1.67E-08
26	GO:0015108	chloride transmembrane transporter activity	50	3.17	1.69E-08
27	GO:0004447	iodide peroxidase activity	38	3.91	1.75E-08
28	GO:0008172	S-methyltransferase activity	56	2.81	9.39E-08
29	GO:0009008	DNA-methyltransferase activity	55	2.71	5.14E-07
30	GO:0030332	cyclin binding	97	2.03	8.54E-07
31	GO:0004931	extracellular ATP-gated cation channel activity	29	4.18	1.22E-06
32	GO:0043425	bHLH transcription factor binding	97	2.01	1.47E-06
33	GO:0004407	histone deacetylase activity	88	2.04	4.20E-06
34	GO:0015018	galactosylgalactosylxylosylprotein	45	2.82	6.68E-06

		3-beta-glucuronosyltransferase activity			
35	GO:0008384	IkappaB kinase activity	18	6.18	7.72E-06
36	GO:0004999	vasoactive intestinal polypeptide receptor activity	33	3.45	9.03E-06
37	GO:0035478	chylomicron binding	10	14.82	9.93E-06
38	GO:0035198	miRNA binding	81	2.07	1.02E-05
39	GO:0035473	lipase binding	18	5.97	1.32E-05
40	GO:0005152	interleukin-1 receptor antagonist activity	34	3.28	1.71E-05
41	GO:0015377	cation:chloride symporter activity	39	2.98	1.83E-05
42	GO:0052743	inositol tetrakisphosphate phosphatase activity	31	3.37	4.37E-05
43	GO:0008112	nicotinamide N-methyltransferase activity	16	6.09	7.12E-05
44	GO:0005528	FK506 binding	52	2.40	7.79E-05
45	GO:0046922	peptide-O-fucosyltransferase activity	18	5.17	1.13E-04
46	GO:0004445	inositol-polyphosphate 5-phosphatase activity	34	2.99	1.56E-04
47	GO:0048019	receptor antagonist activity	70	2.04	2.13E-04
48	GO:0004452	isopentenyl-diphosphate delta-isomerase activity	23	3.87	2.83E-04
49	GO:0004069	L-aspartate:2-oxoglutarate aminotransferase activity	59	2.15	4.03E-04
50	GO:0008190	eukaryotic initiation factor 4E binding	32	2.97	4.45E-04
51	GO:0080130	L-phenylalanine:2-oxoglutarate aminotransferase activity	59	2.09	9.79E-04

52	GO:0015379	potassium:chloride symporter activity	16	4.84	1.50E-03
53	GO:0070546	L-phenylalanine aminotransferase activity	59	2.06	1.53E-03
54	GO:0003721	telomeric template RNA reverse transcriptase activity	9	9.75	2.00E-03
55	GO:0001011	sequence-specific DNA binding RNA polymerase recruiting transcription factor activity	36	2.57	2.14E-03
56	GO:0001087	TFIIB-class binding transcription factor activity	36	2.57	2.14E-03
57	GO:0001093	TFIIB-class transcription factor binding	36	2.57	2.14E-03
58	GO:0008502	melatonin receptor activity	55	2.06	3.54E-03
59	GO:0004370	glycerol kinase activity	34	2.52	6.48E-03
60	GO:0032051	clathrin light chain binding	6	15.89	9.91E-03

Supplementary Table 20. Functional enrichment of tetrapod CNEs using GREAT binomial test over genomic regions – HGNC gene families

No.	Description	No. of observed regions	Fold Enrichment	Bonferroni P-Value
1	OR	1,477	4.94	<1E-100
2	ZNF, ZFHX	206	3.10	3.17E-40
3	IRX, TALE	178	3.28	1.32E-37
4	TSHZ, ZFHX	105	3.79	1.43E-26
5	PRD	241	2.09	7.13E-22
6	NKL	245	2.05	2.38E-21
7	DNAJ	182	2.29	2.35E-20
8	EFN	122	2.28	4.08E-13
9	ZFHX	47	3.48	4.76E-10
10	CUT	106	2.15	9.26E-10
11	WWC	42	3.39	1.66E-08
12	PHF, FANC	54	2.79	3.94E-08
13	NFAT	46	3.09	4.01E-08
14	LCE	47	2.96	9.25E-08
15	PSM	94	2.04	1.80E-07
16	HOXL	93	2.01	4.48E-07
17	PAX, PRD	59	2.45	6.74E-07
18	TNRC	69	2.22	1.65E-06
19	PAR1, PPP2R	22	4.02	4.15E-05
20	PAR1, PRD	40	2.63	4.42E-05
21	THOC	42	2.39	2.63E-04

22	SKOR	36	2.58	2.86E-04
23	PTP2	12	6.43	3.31E-04
24	CHCHD	39	2.39	6.22E-04
25	FBXW	14	5.04	6.91E-04
26	GK	34	2.52	9.38E-04
27	CTD	32	2.59	1.13E-03
28	ZMYND, AKAP	11	6.27	1.21E-03
29	PTPE	49	2.09	1.31E-03
30	RNF, PCGF	25	2.97	1.41E-03

Supplementary Table 21. Functional enrichment of sarcopterygian CNEs using GREAT binomial test over genomic regions - GO "biological process"

No.	Term ID	Description	No. of observed regions	Fold Enrichment	Bonferroni P-Value
1	GO:0048598	embryonic morphogenesis	4,285	2.08	<1E-300
2	GO:0007389	pattern specification process	3,992	2.23	<1E-300
3	GO:0007420	brain development	5,247	2.04	<1E-300
4	GO:0007423	sensory organ development	4,009	2.17	<1E-300
5	GO:0045944	positive regulation of transcription from RNA polymerase II promoter	5,086	2.01	<1E-300
6	GO:0003002	regionalization	3,057	2.61	<1E-300
7	GO:0045893	positive regulation of transcription, DNA-dependent	7,387	2.01	<1E-300
8	GO:0045165	cell fate commitment	2,979	2.66	<1E-300
9	GO:0030900	forebrain development	3,667	2.39	<1E-300
10	GO:0051253	negative regulation of RNA metabolic process	6,003	2.00	<1E-300
11	GO:0045892	negative regulation of transcription, DNA-dependent	5,977	2.02	<1E-300
12	GO:0000122	negative regulation of transcription from RNA polymerase II promoter	4,064	2.11	<1E-300
13	GO:0001654	eye development	2,627	2.27	<1E-300
14	GO:0009952	anterior/posterior pattern specification	1,970	2.61	4.15E-297
15	GO:0072001	renal system development	2,282	2.38	3.39E-291

16	GO:0060173	limb development	1,924	2.52	1.83E-272
17	GO:0001822	kidney development	2,162	2.37	5.74E-272
18	GO:0001655	urogenital system development	2,504	2.18	2.40E-266
19	GO:0001656	metanephros development	1,443	2.82	3.45E-247
20	GO:0048562	embryonic organ morphogenesis	2,098	2.26	5.17E-238
21	GO:0043010	camera-type eye development	2,164	2.22	2.68E-237
22	GO:0030111	regulation of Wnt receptor signaling pathway	1,808	2.41	1.07E-234
23	GO:0021537	telencephalon development	2,100	2.23	8.56E-233
24	GO:0043583	ear development	1,980	2.28	3.83E-230
25	GO:0035270	endocrine system development	1,838	2.29	2.35E-215
26	GO:0035108	limb morphogenesis	1,652	2.39	9.50E-210
27	GO:0001708	cell fate specification	1,028	3.12	1.62E-205
28	GO:0060485	mesenchyme development	1,726	2.31	3.12E-204
29	GO:0021915	neural tube development	1,452	2.48	2.05E-197
30	GO:0050678	regulation of epithelial cell proliferation	1,952	2.14	2.22E-195
31	GO:0030855	epithelial cell differentiation	1,910	2.15	5.61E-195
32	GO:0048839	inner ear development	1,648	2.28	8.11E-191
33	GO:0048762	mesenchymal cell differentiation	1,526	2.36	3.16E-188
34	GO:0021510	spinal cord development	1,134	2.76	9.83E-187
35	GO:0016331	morphogenesis of embryonic epithelium	1,503	2.36	1.06E-186
36	GO:0003007	heart morphogenesis	1,602	2.27	1.81E-182

37	GO:0035136	forelimb morphogenesis	790	3.47	3.19E-182
38	GO:0031016	pancreas development	1,333	2.43	4.15E-174
39	GO:0030326	embryonic limb morphogenesis	1,274	2.47	2.86E-171
40	GO:0042471	ear morphogenesis	1,349	2.39	1.43E-170
41	GO:0021536	diencephalon development	1,090	2.68	1.47E-170
42	GO:0007517	muscle organ development	1,950	2.02	6.48E-170
43	GO:0001657	ureteric bud development	1,345	2.38	3.35E-168
44	GO:0060537	muscle tissue development	1,910	2.03	3.83E-168
45	GO:0035282	segmentation	1,073	2.67	1.02E-166
46	GO:0001838	embryonic epithelial tube formation	1,302	2.37	1.48E-161
47	GO:0035137	hindlimb morphogenesis	760	3.26	8.22E-161
48	GO:0072175	epithelial tube formation	1,311	2.33	5.76E-158
49	GO:0021983	pituitary gland development	855	2.95	3.23E-156
50	GO:0035148	tube formation	1,369	2.27	9.14E-156
51	GO:0021953	central nervous system neuron differentiation	1,470	2.19	4.81E-155
52	GO:0010001	glial cell differentiation	1,310	2.30	9.89E-153
53	GO:0009953	dorsal/ventral pattern formation	1,029	2.58	2.16E-150
54	GO:0045665	negative regulation of neuron differentiation	919	2.75	1.06E-149
55	GO:0014031	mesenchymal cell development	1,309	2.27	3.95E-149
56	GO:0072009	nephron epithelium development	739	3.14	5.95E-148
57	GO:0042063	gliogenesis	1,369	2.20	1.93E-145

58	GO:0048863	stem cell differentiation	1,023	2.54	9.96E-145
59	GO:0030178	negative regulation of Wnt receptor signaling pathway	1,066	2.48	2.02E-144
60	GO:0072073	kidney epithelium development	803	2.91	1.43E-143
61	GO:0048709	oligodendrocyte differentiation	782	2.90	5.81E-139
62	GO:0048663	neuron fate commitment	838	2.76	7.78E-137
63	GO:0051216	cartilage development	1,275	2.21	4.11E-136
64	GO:0061053	somite development	837	2.75	5.85E-136
65	GO:0014033	neural crest cell differentiation	1,027	2.43	2.90E-133
66	GO:0042472	inner ear morphogenesis	1,114	2.33	7.94E-133
67	GO:0060828	regulation of canonical Wnt receptor signaling pathway	1,122	2.32	1.49E-132
68	GO:0008589	regulation of smoothened signaling pathway	665	3.10	1.37E-130
69	GO:0021889	olfactory bulb interneuron differentiation	592	3.32	6.06E-128
70	GO:0021772	olfactory bulb development	725	2.88	3.85E-127
71	GO:0072080	nephron tubule development	570	3.38	5.63E-126
72	GO:0060993	kidney morphogenesis	745	2.80	1.25E-124
73	GO:0021515	cell differentiation in spinal cord	631	3.11	3.18E-124
74	GO:0072006	nephron development	843	2.59	1.68E-123
75	GO:0001709	cell fate determination	770	2.72	8.27E-123
76	GO:0014032	neural crest cell development	941	2.42	6.90E-121
77	GO:0048592	eye morphogenesis	1,275	2.10	7.11E-121
78	GO:0035051	cardiac cell differentiation	848	2.55	1.95E-120

79	GO:0048593	camera-type eye morphogenesis	1,039	2.29	1.55E-119
80	GO:0050679	positive regulation of epithelial cell proliferation	1,188	2.15	1.37E-118
81	GO:0001756	somitogenesis	666	2.90	6.07E-118
82	GO:0061035	regulation of cartilage development	844	2.53	6.82E-118
83	GO:0061326	renal tubule development	571	3.19	1.62E-116
84	GO:0003231	cardiac ventricle development	892	2.42	1.16E-114
85	GO:0003205	cardiac chamber development	959	2.32	3.74E-113
86	GO:0090090	negative regulation of canonical Wnt receptor signaling pathway	785	2.57	7.78E-113
87	GO:0050680	negative regulation of epithelial cell proliferation	963	2.30	3.06E-111
88	GO:0045666	positive regulation of neuron differentiation	944	2.32	6.94E-111
89	GO:0048864	stem cell development	791	2.53	2.01E-110
90	GO:0048738	cardiac muscle tissue development	991	2.26	3.60E-110
91	GO:0072028	nephron morphogenesis	551	3.15	4.07E-110
92	GO:0021532	neural tube patterning	486	3.39	9.50E-108
93	GO:0060850	regulation of transcription involved in cell fate commitment	354	4.40	1.59E-107
94	GO:0021517	ventral spinal cord development	565	3.02	5.86E-106
95	GO:0061005	cell differentiation involved in kidney development	516	3.18	2.11E-104

96	GO:0021891	olfactory bulb interneuron development	504	3.23	2.63E-104
97	GO:0021781	glial cell fate commitment	426	3.64	1.68E-103
98	GO:0061036	positive regulation of cartilage development	428	3.61	4.28E-103
99	GO:0072088	nephron epithelium morphogenesis	525	3.08	1.77E-101
100	GO:0035115	embryonic forelimb morphogenesis	522	3.08	6.56E-101

Supplementary Table 22. Functional enrichment of sarcopterygian CNEs using GREAT binomial test over genomic regions - GO "molecular function"

No.	Term ID	Description	No. of observed regions	Fold Enrichment	Bonferroni P-Value
1	GO:0043565	sequence-specific DNA binding	7,246	2.46	<1E-300
2	GO:0001071	nucleic acid binding transcription factor activity	9,095	2.27	<1E-300
3	GO:0003700	sequence-specific DNA binding transcription factor activity	9,054	2.27	<1E-300
4	GO:0000981	sequence-specific DNA binding RNA polymerase II transcription factor activity	2,360	2.65	<1E-300
5	GO:0044212	transcription regulatory region DNA binding	3,069	2.36	<1E-300
6	GO:0000975	regulatory region DNA binding	3,087	2.34	<1E-300
7	GO:0003705	sequence-specific distal enhancer binding RNA polymerase II transcription factor activity	1,409	2.71	3.66E-226
8	GO:0000976	transcription regulatory region sequence-specific DNA binding	1,313	2.67	2.23E-205
9	GO:0000982	RNA polymerase II core promoter proximal region sequence-specific DNA binding transcription factor activity	942	2.57	1.37E-136
10	GO:0001077	RNA polymerase II core promoter proximal region sequence-specific DNA binding transcription factor activity involved in positive regulation of transcription	756	2.76	1.47E-123

11	GO:0001012	RNA polymerase II regulatory region DNA binding	566	2.85	3.46E-97
12	GO:0000977	RNA polymerase II regulatory region sequence-specific DNA binding	561	2.86	8.00E-97
13	GO:0001159	core promoter proximal region DNA binding	429	3.17	1.38E-86
14	GO:0000987	core promoter proximal region sequence-specific DNA binding	423	3.20	2.56E-86
15	GO:0003707	steroid hormone receptor activity	795	2.19	3.60E-82
16	GO:0008301	DNA bending activity	671	2.32	8.44E-79
17	GO:0004879	ligand-dependent nuclear receptor activity	803	2.11	2.10E-76
18	GO:0000978	RNA polymerase II core promoter proximal region sequence-specific DNA binding	316	3.51	4.53E-73
19	GO:0003908	methylated-DNA-[protein]-cysteine S-methyltransferase activity	136	7.85	8.64E-69
20	GO:0008172	S-methyltransferase activity	139	5.71	5.04E-54
21	GO:0009008	DNA-methyltransferase activity	140	5.65	7.46E-54
22	GO:0001047	core promoter binding	394	2.51	2.05E-53
23	GO:0003706	ligand-regulated transcription factor activity	317	2.82	7.73E-53
24	GO:0035326	enhancer binding	346	2.67	1.63E-52
25	GO:0001972	retinoic acid binding	212	3.68	1.05E-51
26	GO:0001158	enhancer sequence-specific DNA binding	328	2.69	2.56E-50
27	GO:0001046	core promoter sequence-	342	2.60	3.51E-49

		specific DNA binding			
28	GO:0070888	E-box binding	340	2.58	2.07E-48
29	GO:0005501	retinoid binding	239	3.19	8.99E-48
30	GO:0070742	C2H2 zinc finger domain binding	227	3.24	2.12E-46
31	GO:0016524	latrotoxin receptor activity	185	3.74	7.31E-46
32	GO:0019840	isoprenoid binding	242	3.04	4.69E-45
33	GO:0035198	miRNA binding	173	3.63	5.34E-41
34	GO:0033613	activating transcription factor binding	337	2.18	1.83E-33
35	GO:0001078	RNA polymerase II core promoter proximal region sequence-specific DNA binding transcription factor activity involved in negative regulation of transcription	284	2.26	1.91E-30
36	GO:0045499	chemorepellent activity	294	2.22	3.13E-30
37	GO:0035515	oxidative RNA demethylase activity	52	9.46	5.75E-29
38	GO:0035516	oxidative DNA demethylase activity	52	9.46	5.75E-29
39	GO:0001102	RNA polymerase II activating transcription factor binding	188	2.70	2.88E-28
40	GO:0008046	axon guidance receptor activity	330	2.05	3.57E-28
41	GO:0001105	RNA polymerase II transcription coactivator activity	333	2.01	7.00E-27
42	GO:0070491	repressing transcription factor binding	274	2.15	4.18E-26
43	GO:0043425	bHLH transcription factor binding	164	2.78	7.89E-26

44	GO:0044323	retinoic acid-responsive element binding	85	4.47	4.43E-25
45	GO:0070016	armadillo repeat domain binding	123	3.29	6.33E-25
46	GO:0003708	retinoic acid receptor activity	140	2.96	2.65E-24
47	GO:0016922	ligand-dependent nuclear receptor binding	198	2.39	1.47E-23
48	GO:0005113	patched binding	100	3.57	1.80E-22
49	GO:0005115	receptor tyrosine kinase-like orphan receptor binding	69	4.75	1.88E-21
50	GO:0001191	RNA polymerase II transcription factor binding transcription factor activity involved in negative regulation of transcription	190	2.34	1.95E-21
51	GO:0008267	poly-glutamine tract binding	43	8.22	2.72E-21
52	GO:0050682	AF-2 domain binding	110	3.21	2.88E-21
53	GO:0035514	DNA demethylase activity	55	5.86	8.13E-21
54	GO:0043734	DNA-N1-methyladenine dioxygenase activity	55	5.86	8.13E-21
55	GO:0004407	histone deacetylase activity	140	2.66	4.38E-20
56	GO:0031490	chromatin DNA binding	157	2.41	1.60E-18
57	GO:0019213	deacetylase activity	157	2.36	9.15E-18
58	GO:0005110	frizzled-2 binding	87	3.34	1.48E-17
59	GO:0003680	AT DNA binding	151	2.34	9.76E-17
60	GO:0005042	netrin receptor activity	119	2.63	2.00E-16
61	GO:0008190	eukaryotic initiation factor 4E binding	58	4.40	3.09E-16
62	GO:0005109	frizzled binding	178	2.02	1.03E-13

63	GO:0000983	RNA polymerase II core promoter sequence-specific DNA binding transcription factor activity	92	2.77	1.41E-13
64	GO:0004658	propionyl-CoA carboxylase activity	47	4.49	3.91E-13
65	GO:0050692	DBD domain binding	50	4.17	8.10E-13
66	GO:0003839	gamma-glutamylcyclotransferase activity	37	5.59	9.25E-13
67	GO:0001011	sequence-specific DNA binding RNA polymerase recruiting transcription factor activity	60	3.51	1.89E-12
68	GO:0001087	TFIIB-class binding transcription factor activity	60	3.51	1.89E-12
69	GO:0001093	TFIIB-class transcription factor binding	60	3.51	1.89E-12
70	GO:0000979	RNA polymerase II core promoter sequence-specific DNA binding	103	2.48	2.28E-12
71	GO:0002151	G-quadruplex RNA binding	60	3.47	3.32E-12
72	GO:0015183	L-aspartate transmembrane transporter activity	51	3.92	4.52E-12
73	GO:0035035	histone acetyltransferase binding	90	2.56	2.76E-11
74	GO:0043237	laminin-1 binding	122	2.13	2.40E-10
75	GO:0001106	RNA polymerase II transcription corepressor activity	134	2.05	2.56E-10
76	GO:0004132	dCMP deaminase activity	34	4.65	2.32E-09
77	GO:0031870	thromboxane A2 receptor binding	18	9.51	8.66E-09

78	GO:0071837	HMG box domain binding	77	2.48	9.07E-09
79	GO:0008508	bile acid:sodium symporter activity	93	2.26	1.02E-08
80	GO:0009374	biotin binding	42	3.59	2.04E-08
81	GO:0005112	Notch binding	105	2.08	4.25E-08
82	GO:0001099	basal RNA polymerase II transcription machinery binding	67	2.56	5.84E-08
83	GO:0034236	protein kinase A catalytic subunit binding	112	2.01	7.05E-08
84	GO:0016421	CoA carboxylase activity	51	3.00	7.24E-08
85	GO:0070696	transmembrane receptor protein serine/threonine kinase binding	105	2.03	1.86E-07
86	GO:0016885	ligase activity, forming carbon-carbon bonds	51	2.75	1.42E-06
87	GO:0097161	DH domain binding	62	2.46	1.55E-06
88	GO:0070700	BMP receptor binding	95	2.02	1.81E-06
89	GO:0004966	galanin receptor activity	48	2.77	3.70E-06
90	GO:0047021	15-hydroxyprostaglandin dehydrogenase (NADP+) activity	35	3.42	3.75E-06
91	GO:0050221	prostaglandin-E2 9-reductase activity	35	3.42	3.75E-06
92	GO:0047710	bis(5'-adenosyl)-triphosphatase activity	64	2.34	5.23E-06
93	GO:0001671	ATPase activator activity	75	2.17	5.55E-06
94	GO:0032183	SUMO binding	64	2.32	7.61E-06
95	GO:0035500	MH2 domain binding	30	3.73	7.85E-06
96	GO:0035501	MH1 domain binding	30	3.73	7.85E-06

97	GO:0031369	translation initiation factor binding	78	2.10	1.22E-05
98	GO:0008607	phosphorylase kinase regulator activity	33	3.33	2.00E-05
99	GO:0005114	type II transforming growth factor beta receptor binding	60	2.32	2.27E-05
100	GO:0005148	prolactin receptor binding	66	2.19	4.09E-05

Supplementary Table 23. Functional enrichment of sarcopterygian CNEs using GREAT binomial test over genomic regions – HGNC gene families

No.	Description	No. of observed regions	Fold Enrichment	Bonferroni P-Value
1	SOX	555	3.13	3.37E-111
2	BHLH	911	2.28	7.94E-105
3	ZFHX	163	9.89	3.86E-98
4	IRX, TALE	304	4.59	1.01E-97
5	FOX	567	2.62	3.67E-85
6	NKL	420	2.88	9.72E-74
7	TALE	385	2.73	1.28E-61
8	POU	351	2.71	1.19E-55
9	PRD	365	2.59	4.26E-53
10	HOXL	210	3.72	1.01E-52
11	LIM	229	3.40	4.51E-51
12	ZNF, ZFHX	251	3.10	5.30E-49
13	TSHZ, ZFHX	132	3.91	7.35E-35
14	TBX	234	2.47	8.46E-31
15	PAX, PRD	113	3.84	4.21E-29
16	EFN	179	2.73	2.75E-28
17	PTPE	105	3.67	1.94E-25
18	CUT	163	2.70	3.81E-25
19	SIX	77	4.46	2.61E-23
20	PHF, FANC	91	3.85	2.61E-23
21	WNT	138	2.47	8.60E-18

22	PROX	50	3.93	1.07E-12
23	ZMYND	101	2.44	1.94E-12
24	NFAT	60	3.30	4.10E-12
25	SKOR	53	3.11	1.17E-09
26	MEF2	65	2.44	1.17E-07
27	SP	67	2.10	2.12E-05
28	IFN, CD	16	5.29	6.82E-05
29	MYOIII	43	2.36	2.58E-04
30	CHCHD	45	2.26	4.55E-04
31	FBXW, WDR	45	2.05	4.95E-03
32	FATP	40	2.15	5.43E-03

Supplementary Table 24. Top 10 human genes with the highest numbers of tetrapod or sarcopterygian CNEs

Tetrapod CNEs				
No.	**Gene ID**	**No. CNEs**	**Name**	**Description**
1	ENSG00000117114	136	LPHN2	latrophilin 2
2	ENSG00000218336	97	ODZ3	odz, odd Oz/ten-m homolog 3 (Drosophila)
3	ENSG00000169554	94	ZEB2	zinc finger E-box binding homeobox 2
4	ENSG00000184226	92	PCDH9	protocadherin 9
5	ENSG00000155093	85	PTPRN2	protein tyrosine phosphatase, receptor type, N polypeptide 2
6	ENSG00000185008	84	ROBO2	roundabout, axon guidance receptor, homolog 2 (Drosophila)
7	ENSG00000145934	80	ODZ2	odz, odd Oz/ten-m homolog 2 (Drosophila)
8	ENSG00000184349	80	EFNA5	ephrin-A5
9	ENSG00000158321	79	AUTS2	autism susceptibility candidate 2
10	ENSG00000205148	79	AC016251.1	Putative uncharacterized protein FLJ46792
Sarcopterygian CNEs				
No.	**Gene ID**	**No. CNEs**	**Name**	**Description**
1	ENSG00000117114	183	LPHN2	latrophilin 2
2	ENSG00000169554	182	ZEB2	zinc finger E-box binding homeobox 2
3	ENSG00000218336	178	ODZ3	odz, odd Oz/ten-m homolog 3 (Drosophila)
4	ENSG00000091656	173	ZFHX4	zinc finger homeobox 4
5	ENSG00000078328	158	RBFOX1	RNA binding protein, fox-1 homolog (C. elegans)

				1
6	ENSG00000185008	153	ROBO2	roundabout, axon guidance receptor, homolog 2 (Drosophila)
7	ENSG00000153707	139	PTPRD	protein tyrosine phosphatase, receptor type, D
8	ENSG00000170430	134	MGMT	O-6-methylguanine-DNA methyltransferase
9	ENSG00000143995	133	MEIS1	Meis homeobox 1
10	ENSG00000164330	132	EBF1	early B-cell factor 1

51 Amemiya, C., Ota, T. & Litman, G. W. in *Analysis of Nonmammalian Genomes* eds E. Lai & B. W. Birren) 223-256 (Academic Press, 1996).

52 Bruton, M. N. & Coutouvidis, S. E. An inventory of all known specimens of the coelacanth *Latimeria chalumnae* with comments on trends in the catches. . *Env. Biol. Fishes* **32**, 371-390 (1991).

53 Trewavas, E. The presence in Africa East of the Rift Valleys of two species of *Protopterus, P. annectans* and *P. amphibius*. . *Annales du Musee du Congo Belge Tervuren (Belgique) Sciences Zoologiques, serie 4* **1**, 83-100 (1954).

54 Williams, L. J. *et al.* Paired-end sequencing of Fosmid libraries by Illumina. *Genome Res*, doi:gr.138925.112 [pii]
10.1101/gr.138925.112 (2012).

55 Gnerre, S. *et al.* High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proc Natl Acad Sci U S A* **108**, 1513-1518, doi:1017351108 [pii]
10.1073/pnas.1017351108 (2011).

56 Levin, J. Z. *et al.* Comprehensive comparative analysis of strand-specific RNA sequencing methods. *Nat Methods* **7**, 709-715, doi:nmeth.1491 [pii]
10.1038/nmeth.1491 (2010).

57 Grabherr, M. G. *et al.* Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol* **29**, 644-652, doi:nbt.1883 [pii]
10.1038/nbt.1883 (2011).

58 Smit, A. F., Hubley, R. & Green, P. *RepeatMasker Open-3.0*, 1996-2010).

59 Morgulis, A., Gertz, E. M., Schaffer, A. A. & Agarwala, R. A fast and symmetric DUST implementation to mask low-complexity DNA sequences. *J Comput Biol* **13**, 1028-1040, doi:10.1089/cmb.2006.13.1028 (2006).

60 Sayers, E. W. *et al.* Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* **38**, D5-16, doi:gkp967 [pii]
10.1093/nar/gkp967 (2010).

61 Birney, E., Clamp, M. & Durbin, R. GeneWise and Genomewise. *Genome Res* **14**, 988-995, doi:10.1101/gr.1865504
14/5/988 [pii] (2004).

62 Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J Mol Biol* **215**, 403-410, doi:10.1016/S0022-2836(05)80360-2
S0022-2836(05)80360-2 [pii] (1990).

63 Holt, C. & Yandell, M. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics* **12**, 491, doi:1471-2105-12-491 [pii]
10.1186/1471-2105-12-491 (2011).

64 Jurka, J. *et al.* Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet Genome Res* **110**, 462-467, doi:84979 [pii]
10.1159/000084979 (2005).

65 Korf, I. Serial BLAST searching. *Bioinformatics* **19**, 1492-1496 (2003).

66 Cantarel, B. L. *et al.* MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Res* **18**, 188-196, doi:gr.6743907 [pii]
10.1101/gr.6743907 (2008).

67 Korf, I. Gene finding in novel genomes. *BMC Bioinformatics* **5**, 59, doi:10.1186/1471-2105-5-59
1471-2105-5-59 [pii] (2004).

68 Stanke, M. & Waack, S. Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics* **19 Suppl 2**, ii215-225 (2003).

69 Stanke, M., Diekhans, M., Baertsch, R. & Haussler, D. Using native and syntenically mapped cDNA alignments to improve de novo gene finding. *Bioinformatics* **24**, 637-644, doi:btn013 [pii] 10.1093/bioinformatics/btn013 (2008).
70 Quevillon, E. *et al.* InterProScan: protein domains identifier. *Nucleic Acids Res* **33**, W116-120, doi:33/suppl_2/W116 [pii] 10.1093/nar/gki442 (2005).
71 Eilbeck, K., Moore, B., Holt, C. & Yandell, M. Quantitative measures for the management and comparison of annotated genomes. *BMC Bioinformatics* **10**, 67, doi:1471-2105-10-67 [pii] 10.1186/1471-2105-10-67 (2009).
72 Parra, G., Bradnam, K. & Korf, I. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics* **23**, 1061-1067, doi:btm071 [pii] 10.1093/bioinformatics/btm071 (2007).
73 Trapnell, C., Pachter, L. & Salzberg, S. L. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**, 1105-1111, doi:btp120 [pii] 10.1093/bioinformatics/btp120 (2009).
74 Trapnell, C. *et al.* Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol* **28**, 511-515, doi:nbt.1621 [pii] 10.1038/nbt.1621 (2010).
75 Kong, L. *et al.* CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine. *Nucleic Acids Res* **35**, W345-349, doi:35/suppl_2/W345 [pii] 10.1093/nar/gkm391 (2007).
76 Siepel, A. *et al.* Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* **15**, 1034-1050, doi:gr.3715005 [pii] 10.1101/gr.3715005 (2005).
77 Cabili, M. N. *et al.* Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev* **25**, 1915-1927, doi:gad.17446611 [pii] 10.1101/gad.17446611 (2011).
78 Belgard, T. G. *et al.* A transcriptomic atlas of mouse neocortical layers. *Neuron* **71**, 605-616, doi:S0896-6273(11)00601-5 [pii] 10.1016/j.neuron.2011.06.039 (2011).
79 Ulitsky, I., Shkumatava, A., Jan, C. H., Sive, H. & Bartel, D. P. Conserved function of lincRNAs in vertebrate embryonic development despite rapid sequence evolution. *Cell* **147**, 1537-1550, doi:S0092-8674(11)01450-4 [pii] 10.1016/j.cell.2011.11.055 (2011).
80 Pauli, A. *et al.* Systematic identification of long noncoding RNAs expressed during zebrafish embryogenesis. *Genome Res* **22**, 577-591, doi:gr.133009.111 [pii] 10.1101/gr.133009.111 (2012).
81 Gruber, A. R., Findeiss, S., Washietl, S., Hofacker, I. L. & Stadler, P. F. Rnaz 2.0: Improved Noncoding Rna Detection. *Pac Symp Biocomput* **15**, 69-79, doi:9789814295291_0009 [pii] (2010).
82 Li, L., Stoeckert, C. J., Jr. & Roos, D. S. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res* **13**, 2178-2189, doi:10.1101/gr.1224503 13/9/2178 [pii] (2003).
83 Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**, 3389-3402, doi:gka562 [pii] (1997).
84 Notredame, C., Higgins, D. G. & Heringa, J. T-Coffee: A novel method for fast and accurate multiple sequence alignment. *J Mol Biol* **302**, 205-217, doi:10.1006/jmbi.2000.4042

S0022-2836(00)94042-7 [pii] (2000).
85 Guindon, S. *et al.* New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst Biol* **59**, 307-321, doi:syq010 [pii]
10.1093/sysbio/syq010 (2010).
86 Abascal, F., Zardoya, R. & Posada, D. ProtTest: selection of best-fit models of protein evolution. *Bioinformatics* **21**, 2104-2105, doi:bti263 [pii]
10.1093/bioinformatics/bti263 (2005).
87 Anisimova, M. & Gascuel, O. Approximate likelihood-ratio test for branches: A fast, accurate, and powerful alternative. *Syst Biol* **55**, 539-552, doi:T808388N86673K61 [pii]
10.1080/10635150600755453 (2006).
88 Huerta-Cepas, J., Dopazo, J. & Gabaldon, T. ETE: a python Environment for Tree Exploration. *BMC Bioinformatics* **11**, 24, doi:1471-2105-11-24 [pii]
10.1186/1471-2105-11-24 (2010).
89 Wu, X. & Watson, M. CORNA: testing gene lists for regulation by microRNAs. *Bioinformatics* **25**, 832-833, doi:btp059 [pii]
10.1093/bioinformatics/btp059 (2009).
90 Philippe, H. *et al.* Resolving difficult phylogenetic questions: why more sequences are not enough. *PLoS Biol* **9**, e1000602, doi:10.1371/journal.pbio.1000602 (2011).
91 Laurin-Lemay, S., Brinkmann, H. & Philippe, H. Origin of land plants: impact of sequence contamination and missing data. *Curr Biol* (In Press).
92 Stajich, J. E. *et al.* The Bioperl toolkit: Perl modules for the life sciences. *Genome Res* **12**, 1611-1618, doi:10.1101/gr.361602 (2002).
93 Edgar, R. C. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26**, 2460-2461, doi:btq461 [pii]
10.1093/bioinformatics/btq461 (2010).
94 Ebersberger, I., Strauss, S. & von Haeseler, A. HaMStR: profile hidden markov model based search for orthologs in ESTs. *BMC Evol Biol* **9**, 157, doi:1471-2148-9-157 [pii]
10.1186/1471-2148-9-157 (2009).
95 Chevreux, B. *et al.* Using the miraEST assembler for reliable and automated mRNA transcript assembly and SNP detection in sequenced ESTs. *Genome Res* **14**, 1147-1159, doi:10.1101/gr.1917404
1917404 [pii] (2004).
96 Prosdocimi, E. M. *et al.* Errors in ribosomal sequence datasets generated using PCR-coupled 'panbacterial' pyrosequencing, and the establishment of an improved approach. *Mol Cell Probes*, doi:S0890-8508(12)00076-X [pii]
10.1016/j.mcp.2012.07.003 (2012).
97 Eddy, S. R. Profile hidden Markov models. *Bioinformatics* **14**, 755-763, doi:btb114 [pii] (1998).
98 Castresana, J. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol Biol Evol* **17**, 540-552 (2000).
99 Roure, B., Rodriguez-Ezpeleta, N. & Philippe, H. SCaFoS: a tool for selection, concatenation and fusion of sequences for phylogenomics. *BMC Evol Biol* **7 Suppl 1**, S2, doi:1471-2148-7-S1-S2 [pii]
10.1186/1471-2148-7-S1-S2 (2007).
100 Philippe, H. *et al.* Phylogenomics revives traditional views on deep animal relationships. *Curr Biol* **19**, 706-712, doi:S0960-9822(09)00805-7 [pii]
10.1016/j.cub.2009.02.052 (2009).
101 Stamatakis, A., Ludwig, T. & Meier, H. RAxML-III: a fast program for maximum likelihood-based inference of large phylogenetic trees. *Bioinformatics* **21**, 456-463, doi:bti191 [pii]
10.1093/bioinformatics/bti191 (2005).

102 Lartillot, N., Lepage, T. & Blanquart, S. PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. *Bioinformatics* **25**, 2286-2288, doi:btp368 [pii]
10.1093/bioinformatics/btp368 (2009).
103 Lartillot, N. & Philippe, H. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Mol Biol Evol* **21**, 1095-1109, doi:10.1093/molbev/msh112 msh112 [pii] (2004).
104 Lartillot, N. & Philippe, H. Improvement of molecular phylogenetic inference and the phylogeny of Bilateria. *Philos Trans R Soc Lond B Biol Sci* **363**, 1463-1472, doi:888L46874V1LG457 [pii]
10.1098/rstb.2007.2236 (2008).
105 Tajima, F. Simple methods for testing the molecular evolutionary clock hypothesis. *Genetics* **135**, 599-607 (1993).
106 Takezaki, N., Rzhetsky, A. & Nei, M. Phylogenetic test of the molecular clock and linearized trees. *Mol Biol Evol* **12**, 823-833 (1995).
107 Venditti, C., Meade, A. & Pagel, M. Detecting the node-density artifact in phylogeny reconstruction. *Syst Biol* **55**, 637-643 (2006).
108 Webster, A. J., Payne, R. J. & Pagel, M. Molecular phylogenies link rates of evolution and speciation. *Science* **301**, 478, doi:10.1126/science.1083202
301/5632/478 [pii] (2003).
109 Paradis, E., Claude, J. & Strimmer, K. APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics* **20**, 289-290 (2004).
110 Harmon, L. J., Weir, J. T., Brock, C. D., Glor, R. E. & Challenger, W. GEIGER: investigating evolutionary radiations. *Bioinformatics* **24**, 129-131, doi:btm538 [pii]
10.1093/bioinformatics/btm538 (2008).
111 Kumar, S. & Filipski, A. in *Encyclopedia of Life Sciences* (Macmillan Reference Ltd., 2001).
112 Zhang, Z., Schwartz, S., Wagner, L. & Miller, W. A greedy algorithm for aligning DNA sequences. *J Comput Biol* **7**, 203-214, doi:10.1089/10665270050081478 (2000).
113 Vilella, A. J. *et al.* EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. *Genome research* **19**, 327-335, doi:10.1101/gr.073585.107 (2009).
114 Catchen, J. M., Conery, J. S. & Postlethwait, J. H. Automated identification of conserved synteny after whole-genome duplication. *Genome research* **19**, 1497-1505, doi:10.1101/gr.090480.108 (2009).
115 Muffato, M., Louis, A., Poisnel, C. E. & Roest Crollius, H. Genomicus: a database and a browser to study gene synteny in modern and ancestral genomes. *Bioinformatics* **26**, 1119-1121, doi:10.1093/bioinformatics/btq079 (2010).
116 Harris, R. S. *Improved pairwise alignment of genomic DNA*, The Pennsylvania State University, (2007).
117 Blanchette, M. *et al.* Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res* **14**, 708-715, doi:10.1101/gr.1933104
14/4/708 [pii] (2004).
118 Wang, J., Lee, A. P., Kodzius, R., Brenner, S. & Venkatesh, B. Large number of ultraconserved elements were already present in the jawed vertebrate ancestor. *Mol Biol Evol* **26**, 487-490, doi:msn278 [pii]
10.1093/molbev/msn278 (2009).
119 Lee, A. P., Kerk, S. Y., Tan, Y. Y., Brenner, S. & Venkatesh, B. Ancient vertebrate conserved noncoding elements have been evolving rapidly in teleost fishes. *Mol Biol Evol* **28**, 1205-1215, doi:msq304 [pii]
10.1093/molbev/msq304 (2011).

120 Visel, A. *et al.* ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature* **457**, 854-858, doi:nature07730 [pii]
10.1038/nature07730 (2009).
121 McLean, C. Y. *et al.* GREAT improves functional interpretation of cis-regulatory regions. *Nat Biotechnol* **28**, 495-501, doi:nbt.1630 [pii]
10.1038/nbt.1630 (2010).
122 Loots, G. G., Ovcharenko, I., Pachter, L., Dubchak, I. & Rubin, E. M. rVista for comparative sequence-based discovery of functional transcription factor binding sites. *Genome Res* **12**, 832-839, doi:10.1101/gr.225502. Article published online before print in April 2002 (2002).
123 Brudno, M. *et al.* Glocal alignment: finding rearrangements during alignment. *Bioinformatics* **19 Suppl 1**, i54-62 (2003).
124 Loytynoja, A. & Goldman, N. An algorithm for progressive multiple alignment of sequences with insertions. *Proc Natl Acad Sci U S A* **102**, 10557-10562, doi:0409137102 [pii]
10.1073/pnas.0409137102 (2005).
125 Loytynoja, A. & Goldman, N. Phylogeny-aware gap placement prevents errors in sequence alignment and evolutionary analysis. *Science* **320**, 1632-1635, doi:320/5883/1632 [pii]
10.1126/science.1158395 (2008).
126 Fletcher, W. & Yang, Z. The effect of insertions, deletions, and alignment errors on the branch-site test of positive selection. *Mol Biol Evol* **27**, 2257-2267, doi:msq115 [pii]
10.1093/molbev/msq115 (2010).
127 Pond, S. L., Frost, S. D. & Muse, S. V. HyPhy: hypothesis testing using phylogenies. *Bioinformatics* **21**, 676-679, doi:bti079 [pii]
10.1093/bioinformatics/bti079 (2005).
128 Uchikawa, M., Ishida, Y., Takemoto, T., Kamachi, Y. & Kondoh, H. Functional analysis of chicken Sox2 enhancers highlights an array of diverse regulatory elements that are conserved in mammals. *Dev Cell* **4**, 509-519, doi:S1534580703000881 [pii] (2003).
129 Sauka-Spengler, T. & Barembaum, M. Gain- and loss-of-function approaches in the chick embryo. *Methods Cell Biol* **87**, 237-256, doi:S0091-679X(08)00212-4 [pii]
10.1016/S0091-679X(08)00212-4 (2008).
130 Smith, J. J., Sumiyama, K. & Amemiya, C. T. A living fossil in the genome of a living fossil: Harbinger transposons in the coelacanth genome. *Mol Biol Evol* **29**, 985-993, doi:msr267 [pii]
10.1093/molbev/msr267 (2012).
131 Gibbs, P. D. & Schmale, M. C. GFP as a Genetic Marker Scorable Throughout the Life Cycle of Transgenic Zebra Fish. *Mar Biotechnol (NY)* **2**, 107-125, doi:10.1007/s101269900014 [pii] (2000).
132 Danke, J. *et al.* Genome resource for the Indonesian coelacanth, Latimeria menadoensis. *J Exp Zool A Comp Exp Biol* **301**, 228-234, doi:10.1002/jez.a.20024 (2004).
133 Amemiya, C. T. *et al.* VH gene organization in a relict species, the coelacanth Latimeria chalumnae: evolutionary implications. *Proc Natl Acad Sci U S A* **90**, 6661-6665 (1993).
134 Ota, T., Rast, J. P., Litman, G. W. & Amemiya, C. T. Lineage-restricted retention of a primitive immunoglobulin heavy chain isotype within the Dipnoi reveals an evolutionary paradox. *Proc Natl Acad Sci U S A* **100**, 2501-2506, doi:10.1073/pnas.0538029100
0538029100 [pii] (2003).
135 Fjell, C. D., Bosdet, I., Schein, J. E., Jones, S. J. & Marra, M. A. Internet Contig Explorer (iCE)--a tool for visualizing clone fingerprint maps. *Genome Res* **13**, 1244-1249, doi:10.1101/gr.819303
13/6a/1244 [pii] (2003).
136 Schein, J. E. *et al.* in *Bacterial Artificial Chromosomes* Vol. 1: Library Construction, Physical Mapping, and Sequencing eds S. Zhao & M. Stodolsky) (Humana Press, Inc. , 2004).

137 Ewing, B. & Green, P. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res* **8**, 186-194 (1998).
138 Al-Shahrour, F. *et al.* BABELOMICS: a systems biology perspective in the functional annotation of genome-scale experiments. *Nucleic Acids Res* **34**, W472-476, doi:34/suppl_2/W472 [pii]
10.1093/nar/gkl172 (2006).
139 Krishnan, A., Almen, M. S., Fredriksson, R. & Schioth, H. B. The origin of GPCRs: identification of mammalian like Rhodopsin, Adhesion, Glutamate and Frizzled GPCRs in fungi. *PLoS One* **7**, e29817, doi:10.1371/journal.pone.0029817
PONE-D-11-18298 [pii] (2012).
140 Kamesh, N., Aradhyam, G. K. & Manoj, N. The repertoire of G protein-coupled receptors in the sea squirt Ciona intestinalis. *BMC Evol Biol* **8**, 129, doi:1471-2148-8-129 [pii]
10.1186/1471-2148-8-129 (2008).
141 Palczewski, K. *et al.* Crystal structure of rhodopsin: A G protein-coupled receptor. *Science* **289**, 739-745, doi:8725 [pii] (2000).
142 Fredriksson, R. & Schioth, H. B. The repertoire of G-protein-coupled receptors in fully sequenced genomes. *Mol Pharmacol* **67**, 1414-1425, doi:mol.104.009001 [pii]
10.1124/mol.104.009001 (2005).
143 Metpally, R. P. & Sowdhamini, R. Genome wide survey of G protein-coupled receptors in Tetraodon nigroviridis. *BMC Evol Biol* **5**, 41, doi:1471-2148-5-41 [pii]
10.1186/1471-2148-5-41 (2005).
144 Speca, D. J. *et al.* Functional identification of a goldfish odorant receptor. *Neuron* **23**, 487-498, doi:S0896-6273(00)80802-8 [pii] (1999).
145 Semyonov, J., Park, J. I., Chang, C. L. & Hsu, S. Y. GPCR genes are preferentially retained after whole genome duplication. *PLoS One* **3**, e1903, doi:10.1371/journal.pone.0001903 (2008).
146 Yokoyama, S. & Tada, T. Evolutionary dynamics of rhodopsin type 2 opsins in vertebrates. *Mol Biol Evol* **27**, 133-141, doi:msp217 [pii]
10.1093/molbev/msp217 (2010).
147 Philippe, H. & Roure, B. Difficult phylogenetic questions: more data, maybe; better methods, certainly. *BMC Biol* **9**, 91, doi:1741-7007-9-91 [pii]
10.1186/1741-7007-9-91 (2011).
148 Degnan, J. H. & Rosenberg, N. A. Gene tree discordance, phylogenetic inference and the multispecies coalescent. *Trends Ecol Evol* **24**, 332-340, doi:S0169-5347(09)00084-6 [pii]
10.1016/j.tree.2009.01.009 (2009).
149 Shan, Y. & Gras, R. 43 genes support the lungfish-coelacanth grouping related to the closest living relative of tetrapods with the Bayesian method under the coalescence model. *BMC Res Notes* **4**, 49, doi:1756-0500-4-49 [pii]
10.1186/1756-0500-4-49 (2011).
150 Liu, L. BEST: Bayesian estimation of species trees under the coalescent model. *Bioinformatics* **24**, 2542-2543, doi:btn484 [pii]
10.1093/bioinformatics/btn484 (2008).
151 Alfoldi, J. *et al.* The genome of the green anole lizard and a comparative analysis with birds and mammals. *Nature* **477**, 587-591, doi:Doi 10.1038/Nature10390 (2011).
152 Aparicio, S. *et al.* Whole-genome shotgun assembly and analysis of the genome of Fugu rubripes. *Science* **297**, 1301-1310, doi:10.1126/science.1072104
1072104 [pii] (2002).
153 Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12**, doi:Artn 323
Doi 10.1186/1471-2105-12-323 (2011).

154 Trapnell, C. *et al.* Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* **28**, 511-U174, doi:Doi 10.1038/Nbt.1621 (2010).
155 Grabherr, M. G. *et al.* Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* **29**, 644-U130, doi:Doi 10.1038/Nbt.1883 (2011).
156 Bogart, J. P., Balon, E. K. & Bruton, M. N. The chromosomes of the living coelacanth and their remarkable similarity to those of one of the most ancient frogs. *J Hered* **85**, 322-325 (1994).
157 Voss, S. R. *et al.* Origin of amphibian and avian chromosomes by fission, fusion, and retention of ancestral chromosomes. *Genome Res* **21**, 1306-1312, doi:gr.116491.110 [pii]
10.1101/gr.116491.110 (2011).
158 Smith, J. J. & Voss, S. R. Gene order data from a model amphibian (Ambystoma): new perspectives on vertebrate genome structure and evolution. *BMC Genomics* **7**, 219, doi:1471-2164-7-219 [pii]
10.1186/1471-2164-7-219 (2006).
159 Hennessy, F., Nicoll, W. S., Zimmermann, R., Cheetham, M. E. & Blatch, G. L. Not all J domains are created equal: implications for the specificity of Hsp40-Hsp70 interactions. *Protein Sci* **14**, 1697-1709, doi:14/7/1697 [pii]
10.1110/ps.051406805 (2005).
160 Kampinga, H. H. *et al.* Guidelines for the nomenclature of the human heat shock proteins. *Cell Stress Chaperones* **14**, 105-111, doi:10.1007/s12192-008-0068-7 (2009).
161 Kampinga, H. H. & Garrido, C. HSPBs: Small proteins with big implications in human disease. *Int J Biochem Cell Biol*, doi:S1357-2725(12)00204-X [pii]
10.1016/j.biocel.2012.06.005 (2012).
162 Modisakeng, K. W., Dorrington, R. A. & Blatch, G. L. Isolation of genes encoding heat shock protein 70 (hsp70s) from the coelacanth, Latimeria chalumnae. . *South African Journal of Science* **100**, 683-686 (2004).
163 Modisakeng, K. W., Amemiya, C., Dorrington, R. A. & Blatch, G. L. Molecular Biology studies on the coelacanth: a review. *South African Journal of Science* **102**, 479-485 (2006).
164 Modisakeng, K. W. *et al.* Isolation of a Latimeria menadoensis heat shock protein 70 (Lmhsp70) that has all the features of an inducible gene and encodes a functional molecular chaperone. *Mol Genet Genomics* **282**, 185-196, doi:10.1007/s00438-009-0456-4 (2009).
165 Hennessy, F., Cheetham, M. E., Dirr, H. W. & Blatch, G. L. Analysis of the levels of conservation of the J domain among the various types of DnaJ-like proteins. *Cell Stress Chaperones* **5**, 347-358 (2000).
166 Mahalingam, D. *et al.* Targeting HSP90 for cancer therapy. *Br J Cancer* **100**, 1523-1529, doi:6605066 [pii]
10.1038/sj.bjc.6605066 (2009).
167 Odunuga, O. O., Longshaw, V. M. & Blatch, G. L. Hop: more than an Hsp70/Hsp90 adaptor protein. *Bioessays* **26**, 1058-1068, doi:10.1002/bies.20107 (2004).
168 Henry, J. T. & Crosson, S. Ligand-binding PAS domains in a genomic, cellular, and structural context. *Annu Rev Microbiol* **65**, 261-286, doi:10.1146/annurev-micro-121809-151631 (2011).
169 McIntosh, B. E., Hogenesch, J. B. & Bradfield, C. A. Mammalian Per-Arnt-Sim proteins in environmental adaptation. *Annu Rev Physiol* **72**, 625-645, doi:10.1146/annurev-physiol-021909-135922 (2010).
170 Postlethwait, J., Amores, A., Cresko, W., Singer, A. & Yan, Y. L. Subfunction partitioning, the teleost radiation and the annotation of the human genome. *Trends Genet* **20**, 481-490, doi:10.1016/j.tig.2004.08.001
S0168-9525(04)00213-6 [pii] (2004).

171 Hahn, M. E. *et al.* Unexpected diversity of aryl hydrocarbon receptors in non-mammalian vertebrates: insights from comparative genomics. *J Exp Zool A Comp Exp Biol* **305**, 693-706, doi:10.1002/jez.a.323 (2006).
172 Hubbard, T. J. *et al.* Ensembl 2009. *Nucleic Acids Res* **37**, D690-697, doi:gkn828 [pii]
10.1093/nar/gkn828 (2009).
173 Woolfe, A. *et al.* Highly conserved non-coding sequences are associated with vertebrate development. *PLoS Biol* **3**, e7, doi:10.1371/journal.pbio.0030007 (2005).
174 Navratilova, P. *et al.* Systematic human/zebrafish comparative identification of cis-regulatory activity around vertebrate developmental transcription factor genes. *Dev Biol* **327**, 526-540, doi:S0012-1606(08)01320-1 [pii]
10.1016/j.ydbio.2008.10.044 (2009).
175 Lee, T. I. *et al.* Control of developmental regulators by Polycomb in human embryonic stem cells. *Cell* **125**, 301-313, doi:S0092-8674(06)00384-9 [pii]
10.1016/j.cell.2006.02.043 (2006).
176 Kim, T. H. *et al.* Analysis of the vertebrate insulator protein CTCF-binding sites in the human genome. *Cell* **128**, 1231-1245, doi:S0092-8674(07)00205-X [pii]
10.1016/j.cell.2006.12.048 (2007).
177 Xie, X. *et al.* Systematic discovery of regulatory motifs in conserved regions of the human genome, including thousands of CTCF insulator sites. *Proc Natl Acad Sci U S A* **104**, 7145-7150, doi:0701811104 [pii]
10.1073/pnas.0701811104 (2007).
178 Niimura, Y. & Nei, M. Evolutionary dynamics of olfactory receptor genes in fishes and tetrapods. *Proc Natl Acad Sci U S A* **102**, 6039-6044, doi:0501922102 [pii]
10.1073/pnas.0501922102 (2005).
179 Niimura, Y. On the origin and evolution of vertebrate olfactory receptor genes: comparative genome analysis among 23 chordate species. *Genome Biol Evol* **1**, 34-44, doi:10.1093/gbe/evp003 (2009).
180 Zhang, J. *et al.* Loss of fish actinotrichia proteins and the fin-to-limb transition. *Nature* **466**, 234-237, doi:10.1038/nature09137 (2010).
181 Millot, J. & Anthony, T. *Anatomie de Latimeria chalumnae. Tome I. Squelette, Muscles et Fromations de soutien*. (C.N.R.S., 1958).
182 Johanson, Z. *et al.* Fish fingers: digit homologues in sarcopterygian fish fins. *J Exp Zool B Mol Dev Evol* **308**, 757-768, doi:10.1002/jez.b.21197 (2007).
183 Matsuoka, T. *et al.* Neural crest origins of the neck and shoulder. *Nature* **436**, 347-355, doi:nature03837 [pii]
10.1038/nature03837 (2005).
184 Walthall, J. C. & Ashley-Ross, M. A. Postcranial myology of the California newt, Taricha torosa. *Anat Rec A Discov Mol Cell Evol Biol* **288**, 46-57, doi:10.1002/ar.a.20279 (2006).
185 Diogo, R., Abdala, V., Aziz, M. A., Lonergan, N. & Wood, B. A. From fish to modern humans--comparative anatomy, homologies and evolution of the pectoral and forelimb musculature. *J Anat* **214**, 694-716, doi:JOA1067 [pii]
10.1111/j.1469-7580.2009.01067.x (2009).
186 Abdala, V. & Diogo, R. Comparative anatomy, homologies and evolution of the pectoral and forelimb musculature of tetrapods with special attention to extant limbed amphibians and reptiles. *J Anat* **217**, 536-573, doi:JOA1278 [pii]
10.1111/j.1469-7580.2010.01278.x (2010).
187 Shubin, N. H., Daeschler, E. B. & Jenkins, F. A., Jr. The pectoral fin of Tiktaalik roseae and the origin of the tetrapod limb. *Nature* **440**, 764-771, doi:nature04637 [pii]

10.1038/nature04637 (2006).
188 Amemiya, C. T. & Litman, G. W. Complete nucleotide sequence of an immunoglobulin heavy-chain gene and analysis of immunoglobulin gene organization in a primitive teleost species. *Proc Natl Acad Sci U S A* **87**, 811-815 (1990).
189 Strong, S. J. *et al.* A novel multigene family encodes diversified variable regions. *Proc Natl Acad Sci U S A* **96**, 15080-15085 (1999).
190 Yoder, J. A. *et al.* Resolution of the novel immune-type receptor gene cluster in zebrafish. *Proc Natl Acad Sci U S A* **101**, 15706-15711, doi:0405242101 [pii]
10.1073/pnas.0405242101 (2004).
191 Yoder, J. A. *et al.* Immune-type receptor genes in zebrafish share genetic and functional properties with genes encoded by the mammalian leukocyte receptor cluster. *Proc Natl Acad Sci U S A* **98**, 6771-6776, doi:10.1073/pnas.121101598
121101598 [pii] (2001).
192 Cannon, J. P. *et al.* A bony fish immunological receptor of the NITR multigene family mediates allogeneic recognition. *Immunity* **29**, 228-237, doi:S1074-7613(08)00321-X [pii]
10.1016/j.immuni.2008.05.018 (2008).
193 Schreeder, D. M. *et al.* Cutting edge: FcR-like 6 is an MHC class II receptor. *J Immunol* **185**, 23-27, doi:jimmunol.1000832 [pii]
10.4049/jimmunol.1000832 (2010).
194 Viertlboeck, B. C., Schweinsberg, S., Schmitt, R., Herberg, F. W. & Gobel, T. W. The chicken leukocyte receptor complex encodes a family of different affinity FcY receptors. *J Immunol* **182**, 6985-6992, doi:182/11/6985 [pii]
10.4049/jimmunol.0803060 (2009).
195 Burmester, T., Weich, B., Reinhardt, S. & Hankeln, T. A vertebrate globin expressed in the brain. *Nature* **407**, 520-523, doi:10.1038/35035093 (2000).
196 Burmester, T., Weich, B., Reinhardt, S. & Hankeln, T. A vertebrate globin expressed in the brain. *Nature* **407**, 520-523 (2000).
197 Schmidt, M. *et al.* Cytoglobin is a respiratory protein in connective tissue and neurons, which is up-regulated by hypoxia. *Journal of Biological Chemistry* **279**, 8063-8069 (2004).
198 Blank, M. *et al.* Oxygen supply from the bird's eye perspective: Globin E is a respiratory protein in the chicken retina. *J. Biol. Chem.* **286**, 26507–26515 (2011).
199 Fuchs, C., Burmester, T. & Hankeln, T. The amphibian globin gene repertoire as revealed by the *Xenopus* genome. *Cytogenet. Genome Res.* **112**, 296-306 (2006).
200 Roesner, A., Fuchs, C., Hankeln, T. & Burmester, T. A globin gene of ancient evolutionary origin in lower vertebrates: evidence for two distinct globin families in animals. *Mol. Biol. Evol.* **22**, 12-20 (2005).
201 Hoogewijs, D. *et al.* Androglobin: a chimeric globin in metazoans that is preferentially expressed in mammalian testes. *Mol. Biol. Evol.*, in press (2012).
202 Burmester, T. *et al.* Neuroglobin and cytoglobin: genes, proteins and evolution. *IUBMB Life* **56**, 703-707 (2004).
203 Burmester, T. & Hankeln, T. What is the function of neuroglobin? *J. Exp. Biol.* **212**, 1423-1428 (2009).
204 Katoh, K., Kuma, K., Toh, H. & Miyata, T. MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res.* **33**, 511-518, doi:33/2/511 [pii]
10.1093/nar/gki198 (2005).
205 Huelsenbeck, J. P. & Ronquist, F. MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* **17**, 754-755 (2001).

206 Whelan, S. & Goldman, N. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Mol Biol Evol* **18**, 691-699 (2001).
207 Goldstone, J. V. *et al.* The chemical defensome: environmental sensing and response genes in the Strongylocentrotus purpuratus genome. *Dev Biol* **300**, 366-384, doi:S0012-1606(06)01158-4 [pii]
10.1016/j.ydbio.2006.08.066 (2006).
208 Goldstone, J. V. *et al.* Identification and developmental expression of the full complement of Cytochrome P450 genes in Zebrafish. *BMC Genomics* **11**, 643, doi:1471-2164-11-643 [pii]
10.1186/1471-2164-11-643 (2010).
209 Nelson, D. R. *et al.* Comparison of cytochrome P450 (CYP) genes from the mouse and human genomes, including nomenclature recommendations for genes, pseudogenes and alternative-splice variants. *Pharmacogenetics* **14**, 1-18 (2004).
210 Alfoldi, J. *et al.* The genome of the green anole lizard and a comparative analysis with birds and mammals. *Nature* **477**, 587-591, doi:nature10390 [pii]
10.1038/nature10390 (2011).
211 Mullins, M. C. *et al.* Genes establishing dorsoventral pattern formation in the zebrafish embryo: the ventral specifying genes. *Development* **123**, 81-93 (1996).
212 Bauer, H., Lele, Z., Rauch, G. J., Geisler, R. & Hammerschmidt, M. The type I serine/threonine kinase receptor Alk8/Lost-a-fin is required for Bmp2b/7 signal transduction during dorsoventral patterning of the zebrafish embryo. *Development* **128**, 849-858 (2001).
213 Murayama, E., Herbomel, P., Kawakami, A., Takeda, H. & Nagasawa, H. Otolith matrix proteins OMP-1 and Otolin-1 are necessary for normal otolith growth and their correct anchoring onto the sensory maculae. *Mechanisms of development* **122**, 791-803, doi:10.1016/j.mod.2005.03.002 (2005).
214 Beck, F., Erler, T., Russell, A. & James, R. Expression of Cdx-2 in the mouse embryo and placenta: possible role in patterning of the extra-embryonic membranes. *Dev Dyn* **204**, 219-227, doi:10.1002/aja.1002040302 (1995).
215 Epstein, M., Pillemer, G., Yelin, R., Yisraeli, J. K. & Fainsod, A. Patterning of the embryo along the anterior-posterior axis: the role of the caudal genes. *Development* **124**, 3805-3814 (1997).
216 Young, T. *et al.* Cdx and Hox genes differentially regulate posterior axial growth in mammalian embryos. *Dev Cell* **17**, 516-526, doi:S1534-5807(09)00347-5 [pii]
10.1016/j.devcel.2009.08.010 (2009).
217 Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860-921 (2001).
218 Waterston, R. H. *et al.* Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**, 520-562, doi:Doi 10.1038/Nature01262 (2002).
219 Hillier, L. W. *et al.* Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature* **432**, 695-716, doi:Doi 10.1038/Nature03154 (2004).
220 Lindblad-Toh, K. *et al.* Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature* **438**, 803-819, doi:Doi 10.1038/Nature04338 (2005).
221 Hellsten, U. *et al.* The genome of the western clawed frog Xenopus tropicalis. *Science* **328**, 633-636, doi:DOI 10.1126/science.1183670 (2010).
222 Jones, F. C. *et al.* The genomic basis of adaptive evolution in threespine sticklebacks. *Nature* **484**, 55-61, doi:Doi 10.1038/Nature10944 (2012).
223 Kasahara, M. *et al.* The medaka draft genome and insights into vertebrate genome evolution. *Nature* **447**, 714-719, doi:Doi 10.1038/Nature05846 (2007).