doi:10.1038/nature12826

Supplementary Notes

Supplementary Figures

Supplementary Tables

Note: The following supplementary tables are not part of this file and are presented as separate tables:

Summary of major findings

Supplementary Note I. Genome sequencing, assembly and annotation

- The *C. milii* genome was sequenced from a single male individual to a depth of ~19.25× using Roche 454 Titanium and ABI 3730 technologies and assembled using the CABOG 6.1 assembler. The total length of the contigs, N50 contig size and N50 scaffold size are 937 Mb, 46.6 kb and 4.52 Mb, respectively.
- In addition, RNA-seq data was generated from 10 tissues on an Illumina platform.
- The genome assembly is of high quality, as cumulative aligned coverage of 71 finished BACs was 94% with 0.16% base discrepancy.
- Interspersed repeats and low-complexity regions make up 28.2% and 1.8% of the genome, respectively.
- A total of 17,449 protein-coding genes were found in the genome assembly, with an additional 1,423 protein-coding genes predicted in RNA-seq transcripts.
- *C. milii* has a heterozygosity value of 0.00233 per bp, which is similar to that of another wild-stock species, the Atlantic cod (0.00209).
- Only 5.5 Mb of the *C. milii* genome resides on segmental duplications, which represents ~0.6% of the genome.

Supplementary Note II. GC content and isochores

- The *C. milii* genome is less heterogeneous in GC content than human, chicken and zebrafish, but more heterogeneous than lizard, *X. tropicalis*, fugu, stickleback and medaka.
- Approximately 46% of the genome is organized into isochores that are represented by only three families L2, H1 and H2 (average GC levels of 39.9%, 43.0% and 44.4%, respectively).

Supplementary Note III. Characterization of miRNA genes in *C. milii*

- Through small RNA sequencing and homology searches, 693 miRNA gene loci were identified in the *C. milii* genome.
- Tissue-specific expression profiles of *C. milii* miRNAs are similar to those observed in zebrafish and other vertebrates.
- A total of 131 novel miRNAs were predicted in *C. milii*, seven of which are highly expressed in multiple tissues.
- The *C. milii* genome encodes a higher number of known miRNA families than sea lamprey, hagfish, zebrafish, *X. tropicalis* and chicken but fewer families than mammals, indicating that an expansion of miRNA families has occurred in mammalian lineages after the split from other vertebrates.
- 22 out of 136 miRNA families conserved in *C. milii* and mammals have been secondarily lost in teleost fishes. One of these, mir-150, is involved in B-cell development in mice and human while another one, mir-33, is known to regulate cholesterol metabolism in the liver and pancreas.

Supplementary Note IV. Origin and evolution of conserved noncoding elements (CNEs) in vertebrates

- A total of 63,877 'gnathostome CNEs' (average size of 271 bp) were identified based on whole-genome alignments of *C. milii* and 11 bony vertebrates.
- A subset of 1,687 gnathostome CNEs are conserved in all 12 gnathostome genomes. They are associated mainly with transcription factor, chromatin-binding and protein dimerization activity genes as well as with genes involved in central nervous system development.
- Less than 0.6% of gnathostome CNEs could be found in the sea lamprey, sea squirt and amphioxus. This indicates that the emergence of gnathostomes was accompanied by the recruitment of a massive number of CNEs and the assembly of novel gene regulatory networks built around them.
- The teleost ancestor had lost nearly eight times more gnathostome CNEs than the tetrapod ancestor, possibly due to the higher rate of nucleotide substitution in teleost genomes.

Supplementary Note V. Phylogenomics of vertebrates

- Phylogenomic analysis of 699 strict one-to-one orthologues from 13 chordates (including *C. milii*) using Maximum likelihood and Bayesian Inference supported *C. milii* as a sister group to bony vertebrates with maximal support (Bootstrap percent 100 and Posterior Probability 1.0).
- Alternative topologies were rejected with high confidence (AU and NP values <0.03).
- The pattern of intron gains and losses provided independent unequivocal support for *C. milii* as a sister group to bony vertebrates.

Supplementary Note VI. Rate of molecular evolution

- Relative Rate tests based on the genome-wide set of 699 orthologous protein-coding genes indicated that *C. milii* protein coding genes were evolving significantly slower than those of other vertebrates examined (p-value <0.01 for all comparisons), including the coelacanth.
- Two-Cluster tests provided further evidence that protein-coding genes of *C. milii* were evolving significantly slower than those of coelacanth, tetrapods, teleosts and sea lamprey (Z-stat:14.18, 10.93, 20.24 and 27.93, respectively; CP: 99.96%).
- Analysis of neutral evolutionary rates in four-fold degenerate sites indicated that the neutral evolutionary rate of *C. milii* is also the lowest for all vertebrates.

Supplementary Note VII. Intron evolution in vertebrates

- Our analysis of approximately 40,000 intron positions in 3,603 sets of orthologous genes from *C. milii* and nine bony vertebrates is the most extensive study of intron-exon evolution in vertebrates to date.
- The vast majority of intron positions in conserved protein-coding regions are intact in the genomes of *C. milii* and all other vertebrates studied.

- The number of intron changes observed in the *C. milii* lineage since it diverged from the gnathostome ancestor is smaller than in any bony vertebrate lineage, which reflects the lower rates of molecular evolution in *C. milii*.
- Intron losses outnumbered intron gains in most of the branches with the exception of stickleback, in which gains (603) outnumbered losses (126). The gains and losses in the stickleback lineage are the largest recorded in any vertebrate lineage.

Supplementary Note VIII. Large-scale synteny analysis

- 93% of *C. milii* scaffolds show conserved synteny with single chromosomes in chicken. Many *C. milii* scaffolds showing synteny with two or more human chromosomes correspond to single chicken chromosomes, highlighting instances of interchromosomal rearrangements in the human lineage.
- Seven novel syntenic relationships between chicken and human chromosomes were identified through analysis of one-to-one conserved syntenic blocks between *C. milii*, chicken and human.
- Synteny of many large blocks of genes in *C. milii* is extensively conserved in tetrapods. The largest block is a 10 Mb *C. milii* scaffold containing 148 syntenic genes that corresponds to a 45 Mb region on human chromosome 2.
- 82 of 86 *C. milii* scaffolds with homology to chicken microchromosomes show correspondence to a single chicken microchromosome each, suggesting that the chromosomal organization observed in *C. milii* and chicken represents the ancestral vertebrate form.
- A substantially higher number of interchromosomal rearrangements than previously known in the medaka and zebrafish lineages were identified by comparing *C. milii* scaffolds with medaka and zebrafish chromosomes.

Supplementary Note IX. Evolution of protein domains and gene families

- The *C. milii* genome encodes the greatest proportion of proteins containing the immunoglobulin and B-box zinc finger domains of all bony vertebrates such as human and stickleback.
- Six protein domains present in *C. milii* and teleost fishes are lost in tetrapods, exemplified by the 'sea anemone cytotoxic protein' domain which is a part of actinoporins.
- We identified 34 genes that are present in *C. milii* and teleost fishes but lost specifically in tetrapods. These genes have functions related to an ancestral aquatic lifestyle, and include innate immune system genes, fin and lateral line development genes, and olfactory receptor genes.
- 271 genes present in *C. milii* and tetrapods are lost specifically in teleost fishes. Interestingly, 104 of these genes are associated with human genetic diseases, indicating their non-redundant functions in non-teleosts.
- There are 27 *C. milii* genes that are not present in bony vertebrates. These are ancient eukaryotic genes that are still retained in *C. milii* and some invertebrates but lost in bony vertebrates. One of these is the *isopenicillin N epimerase* (*Ipne*) gene, the first instance of the presence of a gene involved in antibiotic synthesis in a vertebrate. Another intriguing instance is the presence of a cephalotoxin-like gene in the *C. milii*

genome. A related gene has been reported only in the cuttlefish, *Sepia esculenta* and ESTs are found in the spiny dogfish and lungfish, indicating its limited distribution in vertebrate genomes.
- Surprisingly, *C. milii* has only six OR-like genes, which is the lowest number observed among all chordates.
- By contrast, the genome has a larger repertoire of vomeronasal receptor genes (4 V1R and 33 V2R genes) similar to that of teleost fishes.

Supplementary Note X. Genes involved in bone formation

- Almost all genes known to be involved in the formation of bone are present in *C. milii*, except for a family of genes that encode the secretory calcium-binding phosphoproteins (SCPPs).
- Zebrafish contains a single bone-specific SCPP gene, *spp1* (*osteopontin*). Manipulation of the activity of this gene in zebrafish using morpholinos and CRISPR/Cas9 system resulted in the reduction of bone formation, supporting the hypothesis that the absence of this gene family in cartilaginous fishes explains the absence of ossified endoskeleton.

Supplementary Note XI. Analysis of the immune system of *C. milii*

- The presence of elaborate innate immune functions in *C. milii* is supported by an essentially modern form of the complement system, a diverse repertoire of pathogen receptors, such as TLR- and NOD-like receptors and intracytoplasmic helicases, upstream and downstream effectors of the inflammasomes, and the basic components of the interferon system; a notable exception is the apparent lack of a TLR4-related receptor and associated components of this signalling pathway among the ten identifiable TLR-like genes.
- The immunoglobulin (Ig) genes are in the cluster-type organization, *TCR* genes are found in the typical translocon organization. Ig heavy (H) genes are linked to TCR loci, likely an ancestral feature of antigen receptors.
- The presence of four MHC paralogous groups in the *C. milii* genome is compatible with two rounds of genome duplication in the ancestor of gnathostomes.
- Consistent with the lack of lymph nodes and germinal centres as well as the relatively long lag-time required to generate humoral immunity, the genes encoding the mammalian regulators of secondary lymphoid tissue formation, TNFRSF3 (LTβR) and its ligands (TNFSF1 [LTα] and TNFSF3 [LTβ]) are absent from the *C. milii* genome, as is a critical cytokine of follicular helper T cells, IL21. By contrast, key determinants of formation and function of spleen (such as HOX11) and thymus (FOXN1) regulating differentiation of thymic epithelial cells, and AIRE, a key regulator of central tolerance) are present.
- All hallmarks of the cytotoxic CD8 lineage of T cells are present in the *C. milii* genome.
- Surprisingly, however, despite the presence of polymorphic MHC class II and invariant chain genes, a bona fide *CD4* gene is absent, as are genes encoding transcription factors regulating the differentiation of several T helper lineages and several of their key effector cytokines.

- The gene encoding FOXP3, the essential regulator of the CD4+ regulatory T (T_{reg}) cells, while present, lacks the structural hallmarks of its mammalian orthologues; in support of the lack of bona fide T_{reg} cells, the gene encoding IL2, a key regulator of T_{reg} cells in mammals, and its specific receptor, IL2RA are absent.
- The genes encoding cytokines of the T helper lineage (IL4, IL9, IL13, IL17E/IL25, IL31) are absent from an otherwise seemingly modern complement of interleukin genes, whereas the gene encoding IFNγ, a classical Th1 cytokine, is present in *C. milii* and nurse shark.
- Several members of the IL10 family of anti-inflammatory cytokines are encoded in the in *C. milii* genome, suggesting that the balance between pro- and anti-inflammatory functions can be achieved without a dedicated regulatory T cell subset.
- *RORC*, the gene encoding RORγt, and genes encoding IL23 and IL23RA, which potentiate Th17 responses, are not present in cartilaginous fishes, suggesting that T_H17 cells are not present; thus, in cartilaginous fish IL17 and IL22 might be furnished by non-lymphoid cells.
- The lack of the RORγt transcription factor suggests that cartilaginous fishes might only possess group 1 innate lymphoid cells.

Supplementary Note I. Genome sequencing, assembly and annotation

I.1 Sequencing and assembly

The elephant shark (*Callorhinchus milii*) DNA for shotgun sequencing, and for the bacterial artificial chromosome (BAC) library, is derived from the testis of a single male caught in Hobart, Tasmania, Australia. All sequences were generated on the Roche 454 Titanium instrument with the exception of the BAC-end, fosmid-end and plasmid-end sequences that were generated on the ABI3730 instrument (Institute of Molecular and Cell Biology (IMCB), Biopolis, Singapore; and Craig Venter Institute (JCVI), Rockville, MD). Sequenced genome coverage for each read type is as follows: BAC End Sequences, 0.10×; 40kb fosmid, 0.02×; 3-4kb plasmids, 1.38×; 454 Fragment, 11.26×; 454 3kb, 4.0×; and 454 8kb, 2.49×. The approximate average depth of coverage is 19.25× and the approximate estimated size of the genome is ~1 Gb.

Assembly version 6.1.2 was built with all sequence data, using the CABOG 6.1 assembler[38]. Post assembly, sequences of 71 finished BACs (NCBI 20/4/2011) were merged into the 6.1.2 assembly. The top scaffold that each BAC mapped to was identified by MEGABLAST (-e 1e-20 -W 200 -p 98). Contigs of the top scaffold that the BAC mapped to were identified by BLASTN (-W 150 -F F). A Perl script was used to create a new contig for each BAC, extend the contig if the 5' and 3' overlapping contigs were longer than the BAC sequence and adjust flanking gaps accordingly. We then sorted scaffolds by decreasing length, assigned new sequence identifiers to contigs and scaffolds, and extended 20-bp and 50-bp gaps to 100-bp as per NCBI's guidelines. This Whole Genome Shotgun project has been deposited at DDBJ/EMBL/GenBank under the accession AAVX00000000. The version described in this paper is the second version, AAVX02000000. In the final assembly, referred to as Callorhinchus_milii-6.1.3, there were 5,393 contigs with an N50 contig length of 46.6 kb (**Supplementary Table I.1**). There were 60 scaffolds with the N50 scaffold length of 4.52 Mb. A total of 937 Mb was assembled in contigs. The overall repeat content is 28%, ~3 times higher than that in similar size genomes of chicken and turkey (~10%) [39,40].

RNA sequencing (RNA-seq) and transcript assembly

For more accurate gene annotation of the reference assembly we have generated RNA-seq from 10 tissues of *C. milii* (brain, gills, heart, intestine, kidney, liver, muscle, ovary, spleen and testis) and assembled them into transcripts. In addition, RNA-seq was generated from the thymus and spleen of nurse shark (*Ginglymostoma cirratum*), an elasmobranch.

Total RNA was extracted using the TRIzol reagent (Invitrogen, Carlsbad, USA), treated with DNAse (TaKaRa Bio Inc, Shiga, Japan) and purified using RNeasy Mini Kit (QIAGEN, Hilden). DNase-treated total RNA was subjected to polyA selection using DynaBeads Oligo dT (Invitrogen, Carlsbad, USA). The polyA selected RNA was used to prepare a strand-specific RNA-seq library using the Script-Seq mRNA preparation kit (Epicentre, Madison, USA) as per the manufacturer's protocol with the following modification: Phusion polymerase (NEB, Ipswich, USA) was used in place of Epicentre's FailSafe PCR mix. The library was purified with QIAGEN MinElute PCR purification kit (Qiagen,Valencia, USA) and eluted with 11µL Elution Buffer. The quality and quantity of the library was analyzed on an Agilent 2100 Bioanalyzer. The library was diluted to a concentration of 8 pM. Cluster generation was performed on a cBOT machine. Each library was paired-end sequenced (2× 76 cycles) in at least two lanes of the Illumina GAIIx platform according to the manufacturer's specifications. The *C. milii* and *G. cirratum* RNA-seq data were submitted to SRA under accession numbers SRA054255 and SRA062964, respectively.

Illumina reads for each tissue were assembled *de novo* using Trinity version r2011-07-13 [41]. Trinity transcripts ≥ 350 bp from each tissue were processed to remove mitochondrial sequences, anti-sense transcripts and transcripts with frameshifts. Mitochondrial sequences were identified by BLASTN search of the transcripts against *C. milii* mitochondrion genome (GenBank HM147137.1). Anti-sense transcripts and transcripts with frameshifts were identified by BLASTX search (1×10^{-7} cut-off) against NCBI RefSeq database. The BLASTX result was also used to classify transcripts as protein-coding ($<1\times10^{-7}$) or non-coding ($>1\times10^{-7}$). Only the longest transcript of each Trinity component was retained for both categories. Transcripts with ≥ 80% protein coverage are considered full-length, while transcripts with <80% protein coverage are considered partial transcripts.

In addition to *de novo* assembly of the reads using Trinity, the reads were also assembled onto the *C. milii* genome assembly using Tophat (version 1.3.1) [42] ; and Cufflinks [43]. First, Illumina RNA-seq reads for each tissue were aligned to the genome assembly using Tophat. The following parameters were used for Tophat alignment: --library-type fr-secondstrand --mate-inner-dist 140 --max-intron-length 100000 --min-intron-length 50. Next, Cufflinks was used to assemble the spliced fragment alignments derived from the Tophat step. The following parameters were used for Cufflinks assembly of the transcripts: --multi-read-correct --library-type fr-secondstrand --max-intron-length 100000 --min-intron-length 50 --pre-mrna-fraction 0.5 --overhang-tolerance 0 --min-isoform-fraction 0.3. Finally, the Cufflink transcripts were processed in the same way as the Trinity transcripts.

Genome assembly quality evaluation

To assess the structural accuracy of the assembly, we analyzed the alignments of the 8.1Mb of finished *C. milii* BAC clone sequence against 67,833 contigs from 21,219 scaffolds of assembly version 6.1.2. Cumulative aligned coverage of 71 BACs was 94%. Overall high quality base discrepancy (SNPs) was 0.16% that translates to 1 SNP per 669 bases. No order discordances (misordered sequence contigs within a supercontig) were observed. For all assembly contigs the cumulative aligned sequence coverage for all BACs (8.1Mb) was 90%. We estimated coverage of the *C. milii* genome by aligning 6,778 *C. milii* full-length cDNAs (GenBank accession numbers JX052268-JX053440 and JX207142-JX212746) to the assembly using BLAT with default parameters. 88% of these sequences aligned over >=90% of their length and 93% aligned over >=80% of their length.

Identification and annotation of repetitive sequences

RepeatScout (using 16-mer frequencies) and PILER-DR (default >94% identity, >400 bp) were run on the 30 longest scaffolds of the assembly (total 330 Mb). In addition, proteins encoded by LINEs, DNA transposons and LTR elements were obtained from RepBase and NCBI and searched against the entire assembly using TBLASTN (-e 1e-3 –F “m S”) to identify genomic regions that are potential interspersed repeats. The results from the RepeatScout, PILER-DF and TBLASTN analyses were clustered together with known *Callorhinchus* repeats [44], using CDHIT-EST at minimum 94% identity and 90% coverage of shorter sequence (–l 100 -c 0.94 -aS 0.9 -G 0 -r 1). For each cluster that was not represented by a known *Callorhinchus* repeat, sequences were oriented to the same strand as the representative sequence and aligned using DIALIGN (with translation to peptide segments and masking of unaligned bases) to obtain consensus sequences. The consensus sequences were filtered of protein-coding genes (using BLASTX against NR at E < 0.01), simple repeats (using trf, nseg and mdust), known *Callorhinchus* repeats (RepeatMasker < 20% or 200 unmasked bases) and low-copy number repeats (sequences that occurred fewer than 10 times at minimum 50% coverage in the *C. milii* genome using RepeatMasker). A total of 88 novel interspersed repeats remained. The low-complexity regions were identified using

DUST. Based on these analyses, *C. milii* contains 28.2% interspersed repeats and 1.8% low-complexity repeats. The types and extent of interspersed repeats are given in **Supplementary Table I.2**.

I.2 Ensembl genome annotation

Conceptual translations from C. milii transcripts

Conceptual translations of Trinity and Cufflinks transcripts were obtained by predicting open reading frames based on BLASTX results against RefSeq vertebrate proteins, and GenBank proteins of *Chondrichthyes* and *Petromyzon marinus* at E < 1e-3. The conceptual translations were clustered at 100% identity, >=25 amino acids using CDHIT, to reduce redundancy across tissue types and assembly methodology. The conceptual translations were searched against vertebrate proteins using BLASTP E < 1e-3 to calculate protein coverage and "length ratio" (length of *C. milii* protein to length of Refseq protein). Conceptual translations that had >=80% coverage of top BLASTP hit or >=40% coverage with >=65% length ratio, were selected for building genes in the Ensembl pipeline. These transcripts were submitted to NCBI under accession numbers JW861113-JW881738, KA353634-KA353668 (BioProject ID PRJNA168475).

Protein-coding genes

The genome was masked using RepeatMasker with the parameters "-s –nolow" and a combined library of 3,109 repetitive elements obtained from RepBase (RepeatMasker edition; release 20110419) (repeats from human, chicken, *Xenopus tropicalis*, zebrafish, medaka, fugu and *Chondrichthyes)*, and 88 novel repetitive elements identified in the *C. milii*.

C. milii proteins (192 obtained from NCBI; and ~65,000 conceptual translations derived from RNA-seq data) were searched against the genome assembly using Pmatch, while ~126,000 RefSeq vertebrate proteins (from human, opossum, platypus, chicken, *Xenopus tropicalis*, medaka, zebrafish, other *Chondrichthyes, Petromyzon marinus*) and ~105,000 RefSeq invertebrate proteins (*Branchiostoma floridae, Ciona intestinalis, Strongylocentrotus purpuratus, Drosophila melanogaster, Caenorhabditis elegans and Nematostella vectensis*) were searched against the genome assembly and *ab initio* Genscan predictions using BLAST. Gene models were built using Genewise and Exonerate, and the best prediction per query protein selected from either algorithm using the ‘BestTargetted’ module in the Ensembl pipeline. Gene models were selected based on the following priority criteria: firstly *C. milii* proteins, secondly vertebrate proteins and lastly invertebrate proteins, using the ‘LayerAnnotation’ module. UTRs were added to the gene models using transcripts obtained through RNA-seq. A total of 17,449 protein-coding genes (26,799 transcripts) and 260 pseudogenes were predicted that were located on 2,954 scaffolds. Half of the protein-coding genes were located on scaffolds that had at least 53 genes, permitting us to carry out long-range gene synteny analysis. In addition to Ensembl prediction, we identified 1,423 unique protein-coding sequences based on RNA-seq transcripts that are either partial or missing in the genome assembly.

Non-coding RNA genes

Non-coding RNA genes were predicted by BLAST and Infernal search against Rfam 10.0 and miRBase 17.0. microRNAs were predicted using miRDeep on small RNA sequences (obtained on Illumina platform) and added to those found through homology. The small RNAs predicted by the Ensembl include 747 miRNA, 215 snoRNA, 77 snRNA and 46 rRNA.

Elephant Shark Genome Browser

Annotation of the *C. milii* genome was carried out using version 59 of the Ensembl annotation pipeline code. The annotation is hosted on the website http://esharkgenome.imcb.a-star.edu.sg/.

I.3 SNP calling

For SNP calling, we retained positions covered by at least two read alignments bearing the same high-quality (Phred score > 27) mutation relative to the assembly. Repetitive and multi-allelic (> 3) sites were ignored. In addition to these criteria, we specifically explored several combinations of other quality-related filters, such as imposing different thresholds of coverage, allele balance, strand bias, minimum distance to an indel or between variants. We used the transition/transversion ratio to assess the quality of each filter. We concluded that, in our data, the optimal SNP calling results from a combination of five criteria: (i) we considered only scaffolds larger than 50 Kb, (ii) we excluded duplicated sequence (being conservative, in this case we took as duplicated all segments with more than 4 windows that have a normalized mean coverage higher than 3, instead of 10 windows), (iii) we imposed that any site should be covered by 5 to 15 reads; (iv) we discarded indel regions (defined as follows: first, for each read we merged the segments of its indels and their 5 bp flanking regions; second, we discarded fragments that are supported by only one read) and (v) SNPs should be separated by a minimum distance of three bp.

Applying the described SNP calling, we discovered 402,090 SNPs in a callable genome of 172,697,509 bp (~18% of the genome), with a transition/transversion ratio of 1.87. If we account for only the coding sequence, the transition/transversion ratio increases to 3.03 (**Supplementary Table I.3**).

I.4 Heterozygosity

The proportion of SNPs in the callable genome would give us an estimation of the overall heterozygosity present in that species. We found that the global heterozygosity is 0.00233 per bp. We also calculated the heterozygosity per bp in 1-kb windows of callable sequence, and the median heterozygosity for all windows is 0.002 (**Supplementary Fig. I.1**). Finally, we plotted the distribution of the mean heterozygosity in 100 windows across the assembly and conclude that the heterozygosity is homogeneous across the entire genome with no traces of inbreeding and runs of homozygosity (**Supplementary Fig. I.2**).

Polymorphism data are currently available for other fishes such as the medaka and the Atlantic cod genomes [45,46]. Kasahara et al. identified SNPs in medaka fish by comparing genomic sequences of two individuals from inbred lines representing two different populations in Japan (coverages were 10.6× and 2.8×, respectively), which allowed them to identify 16.4 million SNPs. Star et al. (2011) sequenced a single heterozygous male Atlantic cod to 40× coverage using 454 reads. One million SNPs were obtained by mapping 454 and Illumina reads to the assembly generated. The authors selected SNPs where 3 reads shared polymorphism and with no additional SNPs in a 5-bp window on either side of the SNP position. They combined the number of SNPs from 454 (603,555 SNPs) and Illumina (873,847 SNPs) reads to obtain a total of 1,047,875 SNPs (429,527 of them were common).

The heterozygosity we estimated for *C. milii* is not strictly comparable to the polymorphism reported for medaka fish because two individuals derived from different inbred lines representing different populations were used. This is likely to explain the high number of SNPs cataloged. However, the use of a single individual makes our heterozygosity estimates more comparable with those obtained for the Atlantic cod, and we found that both vertebrates have similar values of heterozygosity (**Supplementary Table I.4**).

I.5 Segmental duplication

Structural variation and gene duplications are thought to have a profound effect on phenotypic adaptations in humans and other vertebrates. It has long been argued that segmental duplications have played a significant role in gene and genome evolution. We explored the extent and quality of segmental duplications (> 10 Kb and > 94% identity) in the *C. milii* genome assembly.

Method

Our analyses were restricted to the non-repetitive portion of the genome or repetitive regions with higher divergence that allowed us to map the reads uniquely. We made use of the raw read data that was used to assemble the genome (33,281,009 reads sequenced on Roche 454 platform, that represent a raw coverage of 17.60× of the assembly after removal of the TCAG key sequences of each read). First, we explored the read length and quality distributions in order to incorporate into our analyses the major proportion of long high-quality reads (**Supplementary Fig. I.3**) and retained only reads longer than 200 bp (32,629,870 reads). As expected, the qualities values per cycle are predominantly low at the end of the reads (data not shown).

Reads were mapped to a soft-masked version of the genome via local alignments with megaBLAST (parameters -D 3 -p 94 -F m -U T -s 220 -R T) [47]. To improve the accuracy of the alignments around indels, we realigned each pair of aligned sequences using an optimal global alignment algorithm. Finally, we kept the alignments that fulfilled the following criteria: (i) a sequence identity higher than 94%, (ii) alignment length >= 200 bp, (iii) at least 100 bases in non-repetitive regions, (iv) at least 100 bases with Phred scores >= 27 and (v) a proportion of aligned bases relative to the read length of at least 40%.

With the final mappings, 67% of the non-repetitive genome (45% of the whole genome) is covered with 5 to 15 reads (**Supplementary Fig. I.4**). Note that the reduction of coverage from the initial 17.60× coverage results from the use of high-quality local alignments only. On the other hand, the vast majority of reads are almost identical to the reference genome as expected. Not surprisingly, longer scaffolds are more similar to the mapped reads (**Supplementary Fig. I.5**).

Duplication analysis

We screened the elephant shark assembly for segmental duplications using the whole genome shotgun sequence (WSSD) approach [48]. This strategy is based on finding genomic fragments with an excess of read depth by, (1) estimating the number of copies of the sequences in the assembly, based on the fact that regions with a higher copy number will translate into a higher sequencing coverage, (2) partitioning the genome into segments with significantly the same copy number and, (3) selecting as duplications those segments with a remarkably high number of copies and larger than approximately 10 Kb. More precisely, scaffolds were split into windows of 1 kb of non-repetitive sequence and for each window we calculated the mean coverage from the per-base coverage values. For all windows, mean coverage values were normalized by subtracting the mean coverage across windows and [49] dividing by its standard deviation calculated exclusively from those windows with mean coverage lower than 20× (in order to avoid potentially hidden repeats that would inflate the coverage) (**Supplementary Fig. I.6**). The R package DNAcopy was then applied with the default parameters in order to, firstly, smooth possible outlier windows with mean coverage notably different from adjacent windows and, secondly, segment the genome into intervals of markedly different mean coverage. For the resulting intervals (n=17,711) , we applied parameters previously used [50,51] and selected as duplications those intervals with more than 10 windows (i.e. at least 10 Kb of non-repetitive sequence) and an interval median for the normalized mean coverage larger than three, which corresponds to more than 3 standard deviations from the empirical distribution of mean coverage across windows.

Overall, we estimated a total of 5.5 Mb residing on duplications, which represents ~0.6% of the genome (**Supplementary Table I.5**). In terms of length, half of the duplications were shorter than 30 Kb and <1% of the duplications were longer than 100 Kb (data not shown). The distribution of duplications was not homogeneous across scaffolds. Only 0.9% of the scaffolds harbored at least one duplication and, of these, there was an excess of scaffolds almost totally duplicated: out of the 161 scaffolds harboring at least one duplication, 154 (96%) had more than 90% of their sequence duplicated (**Supplementary Fig. I.7a**); however, this was mainly restricted to the shortest scaffolds (**Supplementary Fig. I.7b**). A list of scaffolds with the number of duplications and the percentage of their sequence considered as duplicated can be found in the **Supplementary Table I.6.**

Thus, we aimed at identifying genes that have expanded in the elephant shark genome. Duplicated genes were identified by looking for genes that completely or partially overlap segmental duplications. Of the 18,808 genes (including protein-coding and RNA genes) predicted in the *C. milii* assembly, we found a total of 265 genes as being completely (245 genes) or partially (20 genes) covered by segmental duplications, which altogether represents ~1.4% of the genes (see **Supplementary Table I.7**). However, we note that most of the completely duplicated genes are located on short scaffolds, and hence might be an artifact of the fragmented state of the assembly.

I.6 Identification of potential sex-chromosome

Very little work has been done on sex-determining mechanisms in chondrichthyans (cartilaginous fishes). Analysis of karyotypes in elasmobranch cartilaginous fishes has suggested that male heterogamy is the major sex-determining mechanism [52]. To verify if male is indeed the heterogametic sex in *C. milii*, we mapped 454 fragment reads to the assembly using BLAT and counted average fold sequence coverage for each scaffold. Our expectation was that assembled scaffolds linked with the X and Y chromosome would be present with half the expected coverage. However, no such scaffolds were identified. Our large-scale synteny analysis has shown that many genes on *C. milii* scaffold_2 (17 Mb) have orthologues on the human X chromosome. If scaffold_2 were indeed associated with X chromosome in *C. milii*, the observed sequence coverage of 8.36 for scaffold_2 compared to 8.38 for all scaffolds suggests that the male sex chromosomes are not heterogametic in *C. milii*. Further analysis will be needed to resolve the origin of sex chromosomes and associated questions about the sex determination mechanism in *C. milii*.

Supplementary Note II. GC content and isochores

II.1 Methods

We obtained average GC content for the entire *C. milii* genome using UCSC tool 'faCount'. To determine if the genome is heterogeneous in GC content, we calculated the standard deviation of GC content by first computing the GC content in non-overlapping 3-kb windows on all scaffolds. Windows that had >20% missing bases were discarded. To compare the level of GC heterogeneity of the *C. milii* with that of other vertebrates, this process was repeated for the genomes of human (GRCh37 assembly), chicken (WASHUC2), lizard (AnoCar2), *X. tropicalis* (JGI v4.2), zebrafish (Zv9), medaka (MEDAKA1), stickleback (BROADS1) and fugu (FUGU5). Note that the assemblies of *Xenopus* and fugu are available in the form of scaffolds whereas the others are in chromosomes.

To determine if GC content was heterogeneous at the individual scaffold level, we applied the method described by Fujita et al. [53] for the green anole lizard genome. We first segmented the scaffolds into 300-kb non-overlapping regions and further segmented each region into fifteen 20-kb windows. Windows that had >20% missing bases were discarded, along with their encompassing region. GC content was computed for the remaining windows. Kruskal-Wallis test was conducted on each scaffold to determine if the mean ranks of the groups of windows were significantly different ($P < 0.05$), which would indicate that there is GC content heterogeneity in that scaffold.

Previously, Fujita et al. [53] searched for isochores in the lizard genome and found that it contained few isochores, and these isochores made up only ~20% of the genome, a proportion that was much less than other tetrapods like human (71%) and chicken (54%). We applied the method used by Fujita et al. [53] to identify isochores in *C. milii*, and additionally in lizard and stickleback genomes for comparative purposes. We ran a Bayesian algorithm that infers isochore structure from DNA sequence and a Monte Carlo expectation-maximization algorithm that infers the hyperparameters of the isochore model [54], namely the distribution of the isochore length in each isochore family (lambda), transition matrix of isochore families (P), the prior distribution of the isochore mean per family (eta) and the prior distribution of the segment variances (a, b). To facilitate our comparison with previously published findings [53], we used the same average GC contents (for K=2 families, eta=(0.35, 0.45); for K=3, eta=(0.37, 0.44, 0.50); for K=4, eta=(0.39, 0.44, 0.48, 0.53)), and initial transition probabilities of the hidden Markov model (0.1 for each family). We ran this algorithm for 100 iterations, each time doing 100 simulations on the isochore model. The algorithm identified putative isochores and classified them into specific families with an average GC content and coordinates along the scaffold or chromosome. These putative isochores were merged into larger ones, whenever adjacent regions belonged to the same isochore family. To determine if these putative isochores were homogeneous in GC content with respect to the rest of the scaffold/chromosome, one-tailed F-tests were carried out to compare the variances of the GC content of 3-kb windows in each putative isochore to that of the windows of the entire scaffold or chromosome. Prior to the F-test, arcsine-transformation was carried out on the GC-values to adjust for the non-normality of the data. Bonferroni-corrected P-values were computed, and putative isochores that were longer than 300 kb (based on the classical definition of isochores) and had GC variances less than the scaffold/chromosome on which they resided (P-value < 0.05) were classified as long homogeneous isochores.

To correlate gene density with isochore families, we used the genomic locations for all protein-coding genes of lizard and stickleback from Ensembl database (release 68) and genomic locations of *C. milii* genes from our own gene set. We carried out an overlap analysis of the isochores against genic regions based on genomic coordinates, and calculated the total amount of overlap relative to the total amount of isochoric sequence in each isochore family to obtain the gene density.

II.2 Results

The average GC content of the *C. milii* genome is 42.3%, comparable to the GC content of tetrapods (39.9 – 41.3%) (**Supplementary Fig. II.1** and **Supplementary Table II.1**). Compared to tetrapods, teleost fish genomes display more variation in their GC content (36.6 – 45.5%). The order of the genomes based on decreasing heterogeneity (decreasing standard deviations) is human, chicken, zebrafish, *C. milii*, fugu, *X. tropicalis*, stickleback, medaka and lizard (**Supplementary Table II.1**).

To determine whether GC content was homogeneous in individual scaffolds of *C. milii*, we analysed the scaffolds larger than 4 Mb, which is approximately the N50 scaffold size. There are 70 such scaffolds (total 532.0 Mb). **Supplementary Table II.2** shows the results of the Kruskal-Wallis tests conducted on the GC contents of groups of 20-kb windows in 300-kb regions. There is indeed heterogeneity in GC content across 61 of 70 scaffolds, showing that the *C. milii* genome is heterogeneous not only at the genome level, but also at the scaffold level.

We searched for isochores in *C. milii* scaffolds. Isochores are defined as genomic regions that are greater than 300 kb with reasonably homogeneous GC composition [55]. These isochores can be classified into a few families of characteristic GC levels and are reported to be linked to basic biological properties such as gene density, replication timing and recombination [56]. However, the complement of isochore families differs in various vertebrates. For example, human genome has five isochore families: L1, L2, H1, H2 and H3 with mean GC content of 36.0%, 38.9%, 43.1%, 48.7% and 54.5% respectively[55], while the relative abundance of each of these families differ in the chicken genome, with an underrepresentation of L1 family, and the presence of a high-GC H4 family. On the other hand, teleost fishes have at most 2 isochore families. For instance, zebrafish has L1 and L2 families while stickleback and *Tetraodon* have H1 and H2 families [57].

We ran a Bayesian algorithm [54] to identify isochores in *C. milii*, lizard and stickleback, with the latter two genomes included for comparison. We determined the most appropriate number of isochore families (K) by carrying out likelihood ratio tests of the likelihood values returned from the final iterations of the algorithm using different models with K=2 to K=4. For *C. milii* and lizard, K=3 was accepted (chi-squared test P-value < 0.05) in more than half of *C. milii* scaffolds (13/20) and all lizard macrochromosomes respectively, whereas for stickleback, K=2 was accepted (chi-squared test P-value < 0.05) in more than half of the linkage groups (12/21). Putative isochores identified by the algorithm were post-processed to identify isochores that are >300 kb and possess higher GC homogeneity than the rest of the scaffold or chromosome in which they reside. In *C. milii*, 246 isochores were identified (**Supplementary Table II.3**). These isochores make up 244.3 Mb in total, 46% of the total 532 Mb in the 70 scaffolds analysed. In lizard, 470 isochores were identified (data not shown), approximating the number identified by Fujita et al. [53]. The lizard isochores total 235.3 Mb and make up 22% of the 1,082 Mb analysed. In stickleback, 71 isochores were identified (data not shown). These isochores made up 232.4 Mb, 58% of the 401 Mb genomic sequences analysed. The isochore families, their average GC levels and relative proportions among all isochoric regions in *C. milii*, lizard and stickleback are shown in **Supplementary Table II.4**.

To determine if there is an overrepresentation of protein-coding genes in isochoric regions of high GC content, we compared the genomic locations of isochores and protein-coding genes. The density of genes in H1 isochores is higher than that in L2 isochores in both *C. milii* and lizard (**Supplementary Table II.5**). However, there is no increase in the abundance of genes in H2 isochores compared to H1 isochores of *C. milii*. The gene density does not show an increase in the H2 isochores of stickleback compared to H1 isochores, which is in contrast to the results of Costantini et al. [55], possibly due to the difference in methods used.

In summary, the *C. milii* genome is more heterogeneous in GC content than the bony vertebrate genomes investigated, except for human, chicken and zebrafish. There are 246 isochores in the 70 largest scaffolds of the *C. milii* genome. The amount of isochoric sequence in this sample of the genome suggests that isochores make up ~46% of the entire *C. milii* genome. The isochores in the *C. milii* genome fall within only three families: L2, H1 and H2 (average GC levels of 39.9%, 43.0% and 44.4% respectively) with L2 and H1 families accounting for most of the isochoric sequence (73% and 25% respectively).

Acknowledgements

We thank Prof. Paul Fearnhead and Dr. Matthew Fujita for kindly providing R package and scripts respectively to carry out the identification of isochores.

Supplementary Note III. Characterization of miRNA genes in *C. milii*

III.1 Methods

Small RNA library preparation and sequencing

Small RNA libraries were prepared with the "Small RNA v1.5 Sample Prep Kit" following the manufacturer's instructions (Illumina, San Diego, CA). Briefly, total RNA was isolated by Trizol extraction. The RNA was ligated with 3′ RNA adapter which is specifically modified to target microRNA (miRNAs) and other small RNAs that have a 3′ hydroxyl group resulting from cleavage by Dicer and other RNA processing enzymes, and then with 5′ RNA adapter at the 5′ end of RNAs with a phosphate group. Reverse transcription followed by PCR was performed to select for adapter-ligated fragments. The double-stranded DNA libraries were size-selected by PAGE purification (6% TBE PAGE). Libraries were prepared either for single-plex or multiplexed runs. For multiplexed libraries, specific barcode sequences were incorporated into the RNA adapters and comprised the first 4 bases sequenced. Libraries were either loaded singly or pooled (n=4) at a concentration of 8 pM on an Illumina Genome Analyzer IIx and sequenced for 36 cycles following the manufacturer's protocols.

Data Analysis

The image analysis and base calling were done using Illumina's RTA Pipeline and sequence files were generated for each sample. Pre-processing was performed using the Biopieces (http://www.biopieces.org) package. First, adapter sequences were trimmed with the remove_adapter script, sequences were further filtered by length (between 20 to 24 nt) and unique sequences were counted to produce the mature miRNAs sequences. Multiplexed samples were demultiplexed using custom Perl scripts. Next, miRDeep (version 2) [58] and miRBase Release 19 [59] were used in combination with the *C. milii* genome assembly to predict known and novel miRNAs. A miRDeep score cutoff of 1.0 was used and 548 non-overlapping miRNA genes were predicted in the genome assembly. Potential miRNAs that were not expressed in the tissues examined were predicted by homology search with BLASTN analysis (e-value 1e-03). The tissue expression of miRNA was quantified using the quantifier script of the miRDeep package and then counts were quantile-normalized with Matlab (MathWorks, Natick, MA). The miRNA sequences reported have been deposited in GenBank under accession numbers JX994303 - JX994995.

III.2 Results

Three groups of small RNAs have recently emerged as key regulators of gene expression in eukaryotes. These include microRNAs (miRNAs), short interfering RNAs (siRNAs) and Piwi-associated RNAs (piRNAs). miRNAs are short single-stranded RNAs that regulate gene expression post-transcriptionally by destabilizing or inhibiting efficient translation of mRNAs. Both siRNA and piRNA are involved in silencing of transposons. Among these, miRNAs are unusual in that they are continuously added on in each lineage and once they become part of a gene regulatory network, they are rarely secondarily lost in the descendant lineages [60,61]. The dramatic expansion of miRNA families in bilaterian animals and their stabilizing role in gene regulatory networks has led to the suggestion that they are instrumental in the evolution of organismal complexity [62,63]. Analysis of miRNA families in vertebrates have shown that a majority of miRNAs found in gnathostomes evolved in the stem vertebrate lineage before the divergence of jawless vertebrates and gnathostomes, and that their expression patterns in the shared major tissues are conserved in the two lineages [60]. The largest number of miRNAs cloned to date from chondrichthyans (cartilaginous fishes) is that of the catshark (*Scyliorhinus canicula*), and includes 107 miRNAs belonging to 91 families [60].

Small RNA libraries of *C. milii* were cloned from brain, blood, eye, gills, heart, intestine, kidney, liver, muscle, ovary/uterus, pancreas, rectal gland, skin, spleen, testis and uterus, and sequenced on an Illumina sequencer. The sequences were analysed against the *C. milii* genome assembly using miRDeep. In addition, a homology search was performed by BLASTN using precursor miRNA sequences from miRBase (release 19) against the *C. milii* genome. A total of 693 miRNA gene loci were identified with 548 predicted by miRDeep, and another 145 predicted by homology searches (**Supplementary Table III.1**). Of these, 562 miRNA gene loci have orthologues in miRBase and the remaining 131 miRNA loci are novel. Of the 562 known miRNA, 302 belong to 136 miRNA families. Among the 131 novel miRNA, nine could be assigned to four novel families based on the similarity of their mature sequences.

Most of *C. milii* miRNAs are located in inter-genic regions (66.7% n=462) (**Supplementary Table III.2**) and intronic regions (26.9% n=187) (**Supplementary Table III.3**) with a small fraction found to overlap annotated coding exons (6.3% n=44) (**Supplementary Table III.4**). Approximately 18% of the miRNAs (98) are clustered within 3kb of another miRNA, with the cluster size ranging from 2 to 7 miRNAs (**Supplementary Table III.5**). The top 10 most highly expressed miRNAs families in each tissue profiled in this study are shown in **Supplementary Fig. III.1**. To further define tissue-specific expression, we identified miRNAs whose expression levels were 20-fold greater than their median expression across the other tissues (**Supplementary Table III.6**). Overall, the tissue-specific expression profiles of *C. milii* miRNAs are similar to those observed in zebrafish and other vertebrates [60].

Novel miRNAs in C. milii

A total of 131 novel miRNAs were predicted by miRDeep that are currently not observed in other species in miRBase (Release 19). Most of these novel miRNAs are expressed at levels when compared to known miRNAs (**Supplementary Fig. III.2**); they were identified by us mainly because of the deep sequencing of a number of different tissues. However, we did identify seven novel miRNAs (NOVEL_scaffold108_27224, JX994798; NOVEL_scaffold317_36270, JX994395; NOVEL_scaffold98_26104, JX994372; NOVEL_scaffold10135_43909, JX994714; NOVEL_scaffold11_6499, JX994412; NOVEL_scaffold22_11007, JX994414; NOVEL_scaffold176_32152, JX994420) that are highly expressed in a number of tissues (**Supplementary Fig. III.2**). For instance, NOVEL_scaffold108_27224 is among the top 10 most highly expressed miRNA in kidney, pancreas, spleen and testis while NOVEL_scaffold98_26104 is among the top 10 most highly expressed miRNA in kidney, pancreas and testis (**Supplementary Fig. III.1**).

Secondary loss of miRNA families

C. milii encodes more known families of miRNA (136 families) than sea lamprey (83 families), hagfish (69 families), zebrafish (94 families), *X. tropicalis* (86 families) and chicken (123 families) but fewer families than mouse (269 families) and human (558 families) (**Supplementary Fig. III.3**). This indicates a spurt in the expansion of miRNA families in mammalian lineages after they split from other vertebrates. In addition, very few families of ancient vertebrate miRNAs have been secondarily lost in mammals. For instance, only 5 families shared between *C. milii* and zebrafish, and 3 families shared between *C. milii* and chicken are lost in mouse and human. On the other hand, zebrafish has lost 22 families of miRNA (**Supplementary Table III.7**) that are conserved in *C. milii*, mouse and human. To verify if these losses are specific to zebrafish or common to teleost fishes, we carried out detailed searches of zebrafish, stickleback and fugu genome assemblies. Of the 22 families, 21 are lost in the three teleosts whereas only one family (mir-33) is lost specifically in the zebrafish lineage (**Supplementary Table III.7**). One of the lost miRNA families, mir-150, is involved in B-cell development in mice and human [64,65]. In *C. milii*, this family is expressed predominantly in the spleen and blood suggesting that it may have a similar function in *C. milii*. Two of the miRNAs lost in zebrafish (mir-33A and B) are located in the introns of *sterol-regulatory element-binding factor-1* and *-2* (*SREBF-1* and *-2*) genes in human and mouse. These miRNAs are known to regulate cholesterol metabolism in the liver and pancreas [66,67]. In *C. milii*, they are present in orthologous introns, and are expressed specifically in the liver and pancreas (**Supplementary Table III.7**).

Supplementary Note IV. Origin and evolution of conserved noncoding elements in vertebrates

IV.1 Methods

The *C. milii* genome was repeat-masked using WindowMasker (version 2011-10-19). The soft-masked genome sequences of 11 bony vertebrates – human (hg19), mouse (mm10), cow (bosTau7), opossum (monDom5), chicken (galGal3), lizard (anoCar2), *Xenopus tropicalis* (xenTro3), fugu (fr3), stickleback (gasAcu1), medaka (oryLat2) and zebrafish (danRer7) – were obtained from the UCSC Genome Browser website (http://hgdownload.cse.ucsc.edu/downloads.html). The bony vertebrate genomes were split into 100 Mb subsequences (with 10 kb overlaps). Pairwise alignment of *C. milii* and each of these genomes was carried out using LASTZ [68] with parameters H=2000, Y=3400, L=6000, K=2200, Q= "HoxD55.q", default gap opening and gap extension penalties. MULTIZ [69] was used to generate a multiple alignment which we term as the "12-way alignment". A neutral model was obtained from the 12-way alignment by running PhyloFit [70] on four-fold degenerate sites of protein-coding genes using general reversible "REV" substitution model (**Supplementary Fig. IV.1**).

The 12-way alignment, neutral model and the following parameters – target coverage of input alignments 0.3, average length of conserved sequence 45 bp, the conserved model defined as rho=0.3x of the neutral model – were used to run PhastCons [71] to predict conserved elements (CEs). We filtered these CEs, by firstly removing elements shorter than 60 bp or with log-likelihood (LOD) score lower than 30, and secondly classifying them into the following classes, requiring at least 30% CE coverage: protein-coding exons, UTRs, pseudogenes, ncRNA genes, other transcribed sequences and conserved noncoding elements (CNEs). To identify CEs that overlap protein-coding exons, UTRs, pseudogenes and ncRNA genes, the genomic coordinates of CEs were compared against the coordinates of *C. milii* genes. To eliminate other potentially protein-coding or transcribed sequences, a BLASTX search (E-value cutoff <1e-5) of CEs was carried out against Ensembl release 69 proteins of human, opossum, chicken, *X. tropicalis*, fugu and zebrafish. In addition, genomic coordinates of CEs were compared to the coordinates of *C. milii* spliced RNA-seq transcripts (identified by BLAT alignments of RNA-seq transcripts to the *C. milii* genome assembly) to filter out any other transcribed sequence. The remaining CEs after filtering represent a genome-wide set of conserved noncoding elements (CNEs). Repetitive *C. milii* sequences (softmasked bases exceeding 50%) were identified and excluded from this set of CNEs. Each CNE was assigned to the gene with the nearest transcription start site (TSS) located within 2 Mb of the CNE. For genes with multiple transcripts, the average coordinate of all TSSs was used. Only CNEs present in *C. milii* and at least two bony vertebrates were considered as gnathostome CNEs (gCNEs) since those present in *C. milii* and only one of the bony vertebrates could have evolved independently in the two lineages.

To ascertain the biological significance of CNEs, we compared the human CNEs (excluding sequences located on chrY, chrUn and unordered chromosomes) with two sets of functional sequences: ChIP-seq regions bound by the widely expressed transcriptional coactivator p300 in mouse embryonic tissue (mm9 assembly) [72] that have been “lifted over” to the human genome (hg19 assembly, 4,528 p300-binding sites), and experimentally validated tissue-specific transcriptional enhancers obtained from the VISTA Enhancer Browser [73] (800 enhancers). Enrichment of functional sequences in CNEs was determined using a binomial distribution of the overlaps between the functional regions and 1,000 sets of randomly selected noncoding and nonrepetitive regions in the human genome. One-tailed p-values were calculated to test the hypothesis that the number of overlaps in the CNE set was similar to that in the random sequence set.

To trace the origin of gCNEs, we searched the CNEs against the assemblies of the jawless vertebrate *Petromyzon marinus* v7 (GCA_000148955.1), the tunicate *Ciona intestinalis* version KH (GCA_000224145.1) and *Branchiostoma floridae* v2 (GCA_000003815.1) using BLASTN (E-value cutoff <1e-5). To trace the evolution of gCNEs in different lineages, we first counted the CNEs present in each of the extant species (≥30% coverage) and then determined CNEs that were present in the most recent common ancestors of the four teleost fishes and the seven tetrapods by combining the numbers of CNEs in each of the descendant species. We then estimated the number of CNEs lost in each of the lineages. The loss of CNEs in a particular genome was verified by a lack of a BLASTN hit (E-value cutoff <1e-5) in the repeat-masked genome. To provide further evidence that missing CNEs are indeed lost and not undetectable simply because of sequencing gaps, we searched for gap-free syntenic intervals in *C. milii* and zebrafish or human genomes. Gap-free syntenic intervals in two genomes were defined as contiguous genomic regions shorter than 500 kb that are flanked by 2 orthologous alignments of ≥70% identity over 60 bp, and in the same order and orientation. The majority of tetrapod-lost (99.7%) and teleost-lost (89%) gCNEs that could be assigned to syntenic intervals in human and zebrafish, respectively, were located in ungapped syntenic intervals. These high proportions suggest that the absence of gCNEs in teleost fish or tetrapod genomes is largely due to divergence or deletion of gCNE sequences rather than incomplete sequencing of genomes.

Elephant shark orthologues lost in teleost fishes were identified as those *C. milii* genes with no Inparanoid orthologue or reciprocal BLASTP hit (E-value cutoff <1e-5) in zebrafish, medaka, stickleback or fugu. Similarly, *C. milii* orthologues lost in tetrapods were identified as those with no Inparanoid orthologue or reciprocal BLASTP hit (E-value cutoff <1e-5) in human, mouse, cow, opossum, chicken, lizard or *X. tropicalis*.

To determine the functional enrichment of genes associated with pan-gnathostome CNEs, the R-package 'topGO' was used to identify the Gene Ontology (GO) terms that were most significantly associated with the genes. The 'weight01' algorithm was used to account for the GO graph structure and the Fisher's Exact Test was applied to find significantly enriched GO terms. We used the human gene annotation of the pan-gnathostome CNEs because GO annotation of the human genome is more comprehensive than that of the *C. milii* genome. GO annotation of human genes was obtained from Ensembl release 65. Adjustment of p-values was not carried out for multiple testing, because according to the program documentation, p-value adjustment is not recommended for the 'weight01' method.

IV.2 Results

Comparative genomics of vertebrates have indicated that a substantially higher proportion of conserved sequences are located in noncoding rather than protein-coding regions of the genome [71,74]. This implies that a majority of functional elements reside in the noncoding regions of vertebrate genomes. Transgenic assays of conserved noncoding elements (CNEs) have shown that a significant proportion have the potential to drive gene expression [75-77]. *C. milii*, by virtue of its phylogenetic position, is a crucial reference genome to study the origin and evolution of CNEs in vertebrates.

Previous comparison of a 1.4× coverage assembly of the *C. milii* genome and whole genomes of eight bony vertebrates had identified ~8,500 CNEs (≥70% identity over ≥100 bp; total length 1.82 Mb) conserved in *C. milii* and at least one bony vertebrate [78]. With the whole genome sequence of *C. milii* at hand, we set out to identify a comprehensive set of CNEs in gnathostomes. We generated pair-wise alignments followed by a set of multiple alignments for *C. milii* and 11 bony vertebrate genomes (human, mouse, cow, opossum, chicken, lizard, *X. tropicalis*, fugu, stickleback, medaka, zebrafish) using *C. milii* as the reference, and predicted evolutionarily constrained elements using PhastCons. We identified a set of 63,877 'gnathostome CNEs' (gCNEs) that are conserved in *C. milii* and at least two bony vertebrates (**Supplementary Table IV.1**). These gCNEs (average size 271 bp; total length 17.3 Mb) represent a minimal set of noncoding elements that were present in the common ancestor of gnathostomes and that have remained constrained in gnathostome genomes over a period of 450 million years. Comparison of gCNEs in human with p300-binding sites that demarcate enhancers [72], and with transcriptional enhancers verified in transgenic mouse assays [73], showed that the human gCNEs are 12-fold ($p < 10^{-200}$) and 22-fold ($p < 10^{-200}$) enriched for p300-binding sites and transcriptional enhancers, respectively (**Supplementary Table IV.2**), indicating that the predicted gCNEs are enriched for regulatory elements.

Origin of gnathostome CNEs

To determine the origin of gnathostome CNEs, we searched the CNEs present in *C. milii* and at least one bony vertebrate against the genomes of sea lamprey (*Petromyzon marinus*), sea squirt (*Ciona intestinalis*) and amphioxus (*Branchiostoma floridae*). These organisms represent the lineages of jawless vertebrates, urochordates and cephalochordates, respectively. Of the 109,472 CNEs, only 539 (0.5%), 114 (0.1%) and 290 (0.3%) could be found in the genomes of sea lamprey, sea squirt and amphioxus, respectively, indicating that only a minor fraction of the CNEs (≤0.6%) originated in chordates that arose before gnathostomes. Even considering that the currently available sea lamprey genome is incomplete (genome was sequenced using liver DNA which contains only 80% of the germline genome), the whole genome of the sea lamprey is likely to contain <1.0% of gnathostome CNEs. This indicates that the emergence of gnathostomes was accompanied by the recruitment of a massive number of CNEs (putative *cis*-regulatory elements) and the assembly of elaborate gene regulatory networks built around them. Such gene regulatory networks likely underlie the more complex morphological and physiological phenotypes of gnathostomes compared to jawless vertebrates and non-vertebrate chordates.

Pan-gnathostome CNEs

A subset of 1,687 gCNEs are conserved in all 12 gnathostome genomes studied. These "pan-gnathostome CNEs" (average size of 378 ± 251 bp) are significantly longer than all gCNEs (average size 271 ± 201 bp; two-tailed t-test p-value = 2.32×10^{-9}). These CNEs are associated with 636 genes in the human genome. GeneOntology analysis shows that they tend to be located near genes with transcription factor, chromatin-binding and protein hetero- and homo-dimerization activity (**Supplementary Table IV.3**). They were also associated with genes involved in development of the central nervous system (e.g. neuron differentiation, dorsal spinal cord development), limb, kidney and eye (**Supplementary Table IV.4**). This finding is mirrored by the fact that most genes with the highest numbers of pan-gnathostome CNEs are transcription factor-encoding and/or developmental genes (**Supplementary Table IV.5**). Since these CNEs are under extreme selective constraint in gnathostome genomes, the regulatory network of their target genes must also be highly constrained in gnathostomes. Finally, of these 1,687 pan-gnathostome CNEs, only 55 (3.3%) are present in the sea lamprey genome assembly (which is less than 80% complete) or ~5% in the whole genome, compatible with the notion that major morphological and physiological innovations have occurred in the lineage leading to jawed vertebrates.

Evolutionary pattern of gCNEs in bony vertebrate genomes

We investigated the evolutionary pattern of gCNEs in the two major lineages of bony vertebrates, i.e. the teleosts and tetrapods. Interestingly, the ancestor of the four teleosts had lost nearly eight times more gCNEs (22,536) than the tetrapod ancestor (2,870). The 22,536 gCNEs lost in all four teleost fishes are associated with 6,483 genes in the *C. milii* genome (**Supplementary Table IV.6** lists the top 20 genes), while the 2,870 gCNEs lost in the seven tetrapods are associated with 2,212 *C. milii* genes (see **Supplementary Table IV.7** for a list of top 20 genes that have lost the highest number of gCNEs). One possible reason for the loss of gCNEs is that they may have been associated with genes that are lost in that particular lineage. However, we found that only 9% and 12% of the gCNEs lost in teleosts and tetrapods, respectively, are associated with *C. milii* genes that do not have orthologues in these genomes. This indicates that a vast majority of gCNEs (~90%) have been lost despite their putative target genes still being present in the teleost or tetrapod genomes.

The high proportion of gCNE loss in teleosts (35%; 22,536 out of 63,877) could be due to the higher rate of nucleotide substitution in teleost genomes [79,80]. Furthermore, it has been shown protein-coding sequences in teleosts have experienced a higher frequency of indels (predominantly deletions) than human genes [81]. If noncoding regions of teleosts have also experienced this mutation bias, indels could also have contributed to the loss of gCNEs in teleosts. It is widely accepted that *cis*-regulatory mutations and the resulting changes in gene expression patterns have the potential to give rise to phenotypic variations [82,83]. Teleost fishes are the largest (50% of all living vertebrate species) and the most diverse group of vertebrates. They exhibit spectacular variation in their morphology, behaviour and adaptive features. It is possible that the loss/divergence of a large number of gCNEs (putative *cis*-regulatory elements) might have contributed to the phenotypic diversity of teleosts. Further experimental studies are required to verify this hypothesis.

Supplementary Note V. Phylogenomics of vertebrates

V.1 Methods

Orthologue identification

The complete proteome datasets for the following 10 vertebrates were downloaded from Ensembl version 65 (December 2011) - human, mouse, cow, opossum, chicken, lizard, *Xenopus tropicalis*, stickleback, zebrafish and sea lamprey. The coelacanth proteome was obtained from Ensembl version 66 (February 2012, LatCha1) and the amphioxus dataset was downloaded from JGI Genome Portal (http://genome.jgi-psf.org/Brafl1/Brafl1.home.html). The *C. milii* protein dataset is from the present study. For orthologue identification, only the longest isoform of every protein sequence was retained. InParanoid [84] was run with default settings (i.e., minimum 50% alignment span, minimum 25% alignment coverage, minimum BLASTP score of 40 bits, minimum inparalog confidence level of 0.05) to identify orthologues between each pair of species. MultiParanoid [85] was used to create multi-species clusters of InParanoid orthologues. A custom Perl script was used to generate a dataset comprising strict one-to-one orthologues (core orthologues) from the 13 chordates. Although the sea lamprey genome assembly is considered to be incomplete, the relative completeness of the other genomes to which sea lamprey is compared, mitigates the risk of assigning orthologs incorrectly.

Phylogenomic analyses using a genome-wide set of chordate core-orthologues

In total there were 699 one-to-one orthologues in the 13 chordate dataset. Individual protein alignments of these ‘core’ orthologues were generated using ClustalW. A concatenated alignment was then prepared by merging individual alignments of these 699 proteins. The concatenated alignment was trimmed using Gblocks version 0.91b [86] with auto settings (minimum number of sequences for a conserved position = 7; minimum number of sequences for a flank position = 11; minimum number of contiguous non-conserved positions = 8; minimum length of a block = 10; allowed gap positions = none). The combined length of the trimmed amino acid alignment was 237,907 positions.

For phylogenetic analyses we used Maximum Likelihood (ML) and Bayesian inference (BI) using RAxML [87] and MrBayes [88], respectively. The best-fit substitution model for the alignment was determined using ModelGenerator [89].

Maximum Likelihood

We used the rapid bootstrapping algorithm plus a thorough ML search (-f a option) as implemented in RAxML-7.2.6 [87]. Bootstrap support values/percentages were determined using 100 replicates. A JTT (Jones-Taylor-Thornton) amino acid substitution model [90] with Gamma model of rate heterogeneity (PROTGAMMAJTTF option) as recommended by ModelGenerator was used for the ML run.

Bayesian Inference

We used the parallel (MPI) version [91] of MrBayes 3.2.1 for Bayesian inference. A JTT substitution model with gamma distributed rates was used for the analysis. Two independent runs starting from different random trees were run for 1 million generations with sampling every 100 generations. A consensus tree was generated from all sampled trees excluding the first 2500 samples (corresponds to 25% of the samples) which were discarded as 'burn-in'. Number of chains and processors used were 4 and 8, respectively. Check-pointing (check-frequency of 20,000) was used to help in resumption of the analyses in case of a crash or timeout.

Testing of alternate topologies

We tested the likelihoods of alternate topologies using CONSEL [92]. Site-wise log-likelihood values for the topologies being tested were generated using the "-f g" option implemented in RAxML-7.2.6 [87]. These values were used as input for CONSEL. The following five topologies out of the various possible permutations were specifically tested:

1. Chondrichthyes (represented by *C. milii*) sister to [Sarcopterygii (lobe-finned fishes and tetrapods) + Actinopterygii (represented by stickleback and zebrafish)]
2. Actinopterygii sister to (Chondrichthyes + Sarcopterygii)
3. Tetrapods sister to (lobe-finned fishes, (Teleostei, Chondrichthyes))
4. Tetrapods sister to (Teleostei, (lobe-finned fishes, Chondrichthyes))
5. Tetrapods sister to (Chondrichthyes, (lobe-finned fishes, Teleostei))

Topology 1 is the traditional view where Chondrichthyes (cartilaginous fishes) represent the sister group of bony vertebrates. Topology 2 was suggested as an alternative tree in a previous study [93] and was also implied by a tree generated using 271 full-length cDNA sequences (198 kb alignment) from *C. milii* [94]. Topologies 3 to 5 were supported by the phylogenetic analysis of protein-coding genes from whole mitochondrial genome sequences [95].

V.2 Results

Morphological and paleontological studies have placed Chondrichthyes (cartilaginous fishes) as a sister group to all other gnathostomes [96]. However, molecular phylogenetic analyses based on mitochondrial sequences or nuclear genes have produced conflicting topologies. Molecular phylogeny based on whole mitochondrial genome sequences [95,97,98] split gnathostomes into two clades – Tetrapods and Pisces with the latter including all bony fishes such as lobe-finned fishes and Actinopterygians, and placed chondrichthyans at a terminal position in the piscine branch. On the other hand, analyses of nuclear protein-coding genes have supported the traditional view with varying degrees of support. A traditional tree with a split between chondrichthyans and bony fishes was recovered with a set of 35 nuclear genes [99]. However, this study did not include any lobe-finned fish, and tetrapods were represented by human only. A similar topology was obtained using a smaller dataset of seven nuclear genes that included lungfishes as representatives of lobe-finned fishes [100]. A subsequent study using 14 nuclear genes placed chondrichthyans as sister group to bony vertebrates but could not reject alternative hypotheses (i.e., the ray-finned fishes as sister group to other jawed vertebrates or the monophyly of Pisces) [93]. More recently, analysis of 78 nuclear genes (21,749 amino acid positions) derived from ESTs from lungfish and chondrichthyans supported the traditional tree [101].

The availability of the whole genome sequence of *C. milii* provided a unique opportunity to address this controversy using a phylogenomics approach. We generated Maximum Likelihood (ML) and Bayesian inference (BI) trees using a genome-scale dataset comprising one-to-one orthologues from 13 chordates. This dataset, comprising 699 genes (concatenated amino acid alignment length: 237,907 positions), is an order of magnitude larger than previously used datasets. Both the ML and BI trees gave identical topologies with strong bootstrap (BS) and posterior probability (PP) support for all nodes (**Supplementary Fig. V.1**). With both phylogenetic methods, *C. milii* emerged as a sister group to the remaining gnathostomes with maximal support (BS 100, PP 1.0), consistent with the traditional morphology-based vertebrate phylogeny. The two trees also showed strong support for monophyly of bony vertebrates (BS 98, PP 1.0) which are divided into Sarcopterygii (BS 100, PP 1.0) and Actinopterygii (BS 100, PP 1.0). We evaluated the likelihood of five alternative topologies using CONSEL (see **Supplementary Table V.1**). Topology testing using CONSEL indicated that the topology with *C. milii* as a sister to the bony vertebrates was the most likely topology (approximately unbiased test, AU: 0.972, bootstrap probability, NP: 0.970). In fact, all CONSEL tests gave unambiguous support for this topology (**Supplementary Table V.1**). The second topology with Actinopterygii as sister to Chondrichthyes + Sarcopterygii (AU: 0.028) was rejected based on the 5% significance level. The remaining three topologies, that were previously proposed based on the analysis of mitochondrial sequences, were also significantly rejected (AU: 4e-37, 9e-43 and 1e-74 for topology 2, 3 and 4, respectively). Thus, both the phylogenomic analysis and topology testing provided robust support to the traditional phylogenetic tree with a split between Chondrichthyes and bony vertebrates.

Another promising character state for phylogenetic analysis is the gain/loss of introns which are rare events [102,103]. Indeed, for more than 99% of intron positions with observed changes in *C. milii* and other vertebrates (see section on Intron evolution), the phylogenetic distribution could be explained by a single intron gain or loss. To address the controversy regarding the phylogenetic position of chondrichthyans, we used slow-evolving invertebrate outgroups to search for intron gain/loss events supporting alternative hypotheses for the deepest divergences within gnathostomes (see **Supplementary Note VII**). The resultant data was unequivocal, with 13 gains and 10 losses supporting *C. milii* as a sister group to bony vertebrates, and no character supporting an alternative phylogeny in which *C. milii* groups with teleost fish ($P = 2 \times 10^{-7}$). Thus, the gain and loss of introns provided independent support for Chondrichthyes as a sister group to bony vertebrates.

Supplementary Note VI. Rate of molecular evolution

VI.1 Methods

Tajima's Relative Rate Test

We used the concatenated protein alignment (237,907 positions) obtained from the core orthologue dataset for 13 chordates. Sea lamprey was used as outgroup for testing the relative rates of *C. milii* proteins with those of the remaining gnathostomes. The relative rate test (RRT) is based on three sequences – two ingroups and one outgroup. A significantly lower number of unique differences (based on p-value) in either of the ingroup sequence would suggest that the particular ingroup is evolving at a slower rate compared to the other ingroup.

Two-Cluster Analysis

We performed the Two-Cluster test as implemented in the program LINTRE [104]. The Two-Cluster Test can be considered as an extension of the RRT that addresses the comparison of multiple sequences [105]. Pairwise distances between individual taxa were calculated from the 13-chordate ML tree (**Supplementary Fig. V.1**) using the R-package 'ape'. The corresponding 13 chordate protein alignment was converted to an appropriate format for LINTRE. A tree file was prepared manually based on the distances calculated from the ML tree (see LINTRE documentation). Since direct comparison of *C. milii* with specific taxa (or groups) was not possible in the complete dataset, we performed additional pairwise comparisons between *C. milii* and the taxa/group of interest. Specifically, we were interested in comparing *C. milii* with the coelacanth, tetrapods (represented by human, mouse, cow, opossum, chicken, lizard, *Xenopus*), teleosts (represented by stickleback and zebrafish) and sea lamprey. For these pairwise tests, only the sequences of interest were retained in the concatenated alignment and the input tree was adjusted accordingly. The resultant Two-Cluster output tables indicated which of the taxa/groups under consideration was evolving significantly faster or slower based on Z-statistics. Lintre calculates standard errors based on variances and covariances of the pairwise distances. For confirmation, we also calculated standard error values based on 100 bootstrap replicates using MEGA5 [106] and 100 random trees selected from the Bayesian samples excluding burn-in.

Since there is a tendency of branch lengths to be underestimated in regions of a tree containing few species (node-density artifact) [107], we checked for the presence of this phenomenon in our Bayesian tree using the node-density artifact analyzer (http://www.evolution.reading.ac.uk/pe/index.html). However, we did not find this artifact in our tree.

Neutral tree based on four-fold degenerate sites

To compare the neutral nucleotide mutation rate for the different species, we generated a neutral tree based on an alignment of four-fold degenerate (4D) sites from 13 chordates. We used the topology obtained from our phylogenomic analyses (**Supplementary Fig. V.1**) as an input for RAxML-based optimization of branch lengths for the 4D alignment. Codon alignments of the coding sequences based on the individual protein alignments were generated using PAL2NAL [108]. A concatenated coding sequence alignment was generated from them. An alignment of 4D sites was extracted from this coding sequence alignment using the 'do.4d' option ('read.msa' function) as implemented in the RPHAST package [109]. The extracted 4D alignment contained 30,179 positions. We used the "-f e" option (optimize model+branch lengths for the input tree under GAMMA) in RAxML-7.2.6 [87] to generate a neutral tree for this alignment. A General Time Reversible nucleotide substitution model [110] with Gamma model of rate heterogeneity (GTRGAMMA) as deduced by ModelGenerator [89] was used for the analysis. Pairwise distances to the outgroup (amphioxus) were calculated from the neutral tree using the 'cophenetic.phylo' function as implemented in the R-package 'ape' [111].

VI.2 Results

Previous studies based on a few mitochondrial and nuclear protein-coding genes have shown that the nucleotide substitution rate in elasmobranch sharks is an order of magnitude slower than that of mammals [112,113]. To determine if this is a genome-wide phenomenon of chondrichthyans, we estimated and compared the molecular evolutionary rate of *C. milii* and other vertebrates using genome-wide set of 699 orthologous protein-coding genes. We first tested the evolutionary rates of protein-coding sequences from 13 chordates using the concatenated amino acid alignment (237,907 positions). Sea lamprey was used as the outgroup for comparing evolutionary rates in *C. milii* with other gnathostomes. Tajima's Relative Rate tests indicated that the *C. milii* protein-coding genes were evolving slower than not only mammals but also all other gnathostomes examined including the coelacanth (p-value < 0.01 for all comparisons; **Supplementary Table VI.1**). Comparison of relative rate of *C. milii* proteins with those of sea lamprey using amphioxus as the outgroup revealed that

C. milii proteins were evolving significantly slower than those of sea lamprey as well (χ^2: 794.53, p-value: 8.3E-175).

Additionally, we performed the Two-Cluster Test as implemented in the program LINTRE [104]. This test also indicated that protein sequences in the *C. milii* lineage were evolving significantly slower than those in other gnathostomes (Z-stat: 14.98, confidence probability CP: 99.96%; **Supplementary Table VI.2**, tree 1, node 23). In order to make a direct comparison of *C. milii* with other gnathostome groups, we performed a pairwise Two Cluster test involving the two ingroups of interest (*C. milii* and either coelacanth, tetrapods, teleosts or sea lamprey) and outgroups sea lamprey and/or amphioxus. These pairwise Two-Cluster Tests provided further evidence that *C. milii* sequences were evolving significantly slower than those of coelacanth, tetrapods, teleosts and sea lamprey (Z-stat: 14.18, 10.93, 20.24, and 27.93, respectively and CP: 99.96% (**Supplementary Table VI.2**). We also checked whether the distances to outgroup (from the ML tree, **Supplementary Fig. V.1**) were significantly different for *C. milii* compared to other vertebrate groups using Z-statistics. For this purpose, standard errors were calculated by comparison with 100 randomly sampled Bayesian trees and 100 bootstrap replicates from MEGA5. This analysis further confirmed that *C. milii* (0.93 substitutions per site) is slower-evolving than coelacanth, tetrapods, teleost fishes and sea lamprey (0.96, 1.02, 1.05 and 1.04 substitutions per site, respectively) (**Supplementary Table VI.3**). Thus, both RRT and Two-Cluster Tests show that the *C. milii* protein-coding sequences are evolving significantly slower than those of all other vertebrates.

To determine whether the slow evolutionary rate of protein-coding genes in *C. milii* is a reflection of the neutral nucleotide mutation rate, we generated a tree based on 4D sites. Based on the branch lengths of the neutral tree (**Fig. 2**), *C. milii* appears to be the slowest-evolving species among the vertebrates. To confirm this, we examined the actual distances to the outgroup (**Supplementary Table VI.4**) and found that *C. milii* had the smallest pairwise distance to amphioxus (2.06 substitutions per 4D site) followed by coelacanth (2.121 substitutions per 4D site) and sea lamprey (2.125 substitutions per 4D site). This confirms that the neutral evolutionary rate is also the lowest for *C. milii*. Recently, the genome of the western painted turtle was sequenced and shown to have a lower substitution rate relative to other amniote species analyzed [114]. In order to compare the neutral substitution rates of the turtle with *C. milii*, we generated a neutral tree using 4D sites extracted from 399 one-to-one orthologues from 14 chordate species including those from turtle (**Supplementary Fig.**

VI.1). This neutral tree also showed that the neutral substitution rate of *C. milii* is the lowest among all the vertebrates.

Our analyses have shown that the overall molecular evolutionary rate of *C. milii* is the lowest among vertebrates. The reason(s) for the slow evolutionary rate is not clear. Several physiological and environmental factors have been proposed to explain the inter-specific variation in molecular evolutionary rates, including body size, weight-specific metabolic rate, generation time (or number of germ line replication events per generation), DNA repair efficiency, and exposure to mutagens such as UV radiation [115-117]. *C. milii* is a moderate-sized cartilaginous fish (maximum length ~120 cm) that normally lives off southern Australia and New Zealand at depths of 200 to 500 m [118], and visits shallow bays and estuaries only during spring for spawning. Males and females attain maturity around three and four years, respectively [119]. It is a benthic forager feeding mainly on crustaceans, molluscs, echinoderms and polychaetes [120]. Although the metabolic rate of *C. milii* is not known, it is likely to be low, because the metabolic rates of its close relatives, the elasmobranchs are 5 to 10 times lower than mammals [121]. Additionally, elasmobranchs with a less-active lifestyle have a lower metabolic rate than the more active ones [122]. A higher metabolic rate is hypothesized to result in an increased mutation rate owing to DNA damage by mutagenic by-products of oxidative respiration, and increased rate of DNA synthesis and nucleotide replacement [116,123,124]. Consequently, a lower metabolic rate should result in a decreased mutation rate.

Supplementary Note VII. Intron evolution in vertebrates

VII.1 Methods

We extracted intron-exon structures from the GTF and genomic files of *C. milii* and nine other vertebrates – human, mouse, cow, opossum, chicken, Anolis lizard, *X. tropicalis*, stickleback and zebrafish (downloaded from Ensembl version 65). A set of 3,603 orthologous genes among these vertebrates was identified by using InParanoid and Multiparanoid (see **Supplementary Note V**). The orthologues were aligned at the protein level using ClustalW and positions corresponding to intron positions were mapped using a custom Perl script. Only introns longer than 50 nucleotides with characteristic U2 or U12 splicing boundaries (GT/AG, GC/AG or AT/AC) were considered.

We used the following method to obtain an initial set of discordant (i.e., not present in all studied vertebrates) intron positions: For each pair of species, we extracted the pairwise alignment from the 10-species ClustalW alignment. We then identified positions at which one species had an intron but at which the other species (i) lacked that intron position; (ii) lacked an intron within 10 codons; and (iii) had a conserved local alignment. To determine whether two species have conserved local alignment, we considered the region of pairwise alignment including the 10 aligned amino acid positions both up- and downstream of the intron position (that is, not including positions at which both species contained gaps). A region was scored as conserved if the region: (i) is not dominated by gaps (defined as ≤10 gaps within the closest 10 aligned amino acid positions) and (ii) has ≥50% amino acid identity within the 10 aligned positions. Next, for each intron position for which both up- and downstream regions passed this automatic filtering process, two researchers independently manually analyzed the region, which led to elimination of 7% of intron positions (located within low-complexity, repetitive regions or regions of uncertain alignment). In addition, many species-specific introns were found to be highly dubious, with short intron lengths that were multiples of three nucleotides and did not contain in-frame stop codons - these were removed.

This left 1,052 discordant intron positions. We then set out to determine whether these positions had undergone intron gains or losses, and the phylogenetic position of the change(s). Phylogenetic distribution of 939 positions could be determined automatically by a custom Perl script. For the remaining 113 introns, manual inspection was required to ascertain the history of intron gain and loss due to various alignment and annotation difficulties (e.g., large number of positions in the alignment that lack sequence similarity for some species [typically only one species]), suggesting that the aligned regions for that species were not homologous to the regions for the other species. Intron gain and loss was inferred based on these phylogenetic distributions using parsimony. Supporting the approach of parsimony was the finding that 99% of discordant intron positions exhibited a phylogenetic pattern consistent with a single loss or gain.

Emboldened by the highly parsimonious phylogenetic patterns observed, we next used intron losses and gains to study early gnathostome phylogeny. Two main competing hypotheses for early gnathostome phylogeny were tested: grouping teleost fish either with tetrapods (Tel+Tet) or with elephant shark (Tel+ES). The former implies a split between Chondrichthyes and bony vertebrates and supports the traditional phylogeny based on morphological characters. We studied invertebrate intron-exon structures to trace the history of intron loss and gain within early gnathostomes in order to determine their phylogenetic relationships. Intron positions found in elephant shark and invertebrates but not in teleosts or tetrapods lend support to the "Tel+Tet" hypothesis (i.e., loss in the Tel+Tet ancestor), as do introns found only in teleosts and tetrapods (i.e., gain in the Tel+Tet ancestor). Conversely, intron positions restricted to invertebrates and tetrapods (i.e., loss in Tel+ES ancestor) or to teleosts and elephant shark (i.e., gain in Tel+ES ancestor) lend support to the "Tel+ES" hypothesis.

To seek support for either hypothesis, we studied 49 intron positions whose distribution within the ten vertebrates made them potentially informative (25 *C. milii*-specific, 20 bony vertebrate-specific and four tetrapod-specific). For each intron, we performed protein alignments with mapped intron positions of vertebrates and the orthologue of the slow-evolving cephalochordate amphioxus (*Branchiostoma floridae*), as described above. This clearly indicated intron presence or absence for most introns. In other cases, the gene or gene region appears to be absent from the current set of amphioxus gene predictions. In these cases it was necessary to use TBlastN searches against the genome itself to find the region in the amphioxus genome, which showed clear intron presence/absence in several additional cases. In cases where amphioxus lacked the intron, we studied additional slow-evolving invertebrate species for which genomes are available (the lophotrochozoan *Lottia gigantea,* the cnidarian *Nematostella vectensis*, and the placozoan *Trichoplax adhaerens,* downloaded from http://www.jgi.doe.gov/), using the same method. In total, this analysis yielded 23 phylogenetically informative characters, all of them supporting the "Tel+Tet" hypothesis (13 intron gains and 10 losses), and none supporting the "Tel+ES" (for the two Tel+ES introns for which presence/absence could be determined in invertebrates, the introns were present in various invertebrates, consistent with a single loss in the tetrapod ancestor, regardless of phylogeny). **Supplementary Table VII.1** summarizes these discordant introns that have undergone a change in the deepest branches within gnathostomes: 15 changes on the *C. milii* branch (9 gains and 6 losses), 23 changes on the Tel+Tet branch (13 gains and 10 losses), and seven changes that could not be directionalized because of insufficient or conflicting evidence in invertebrate outgroups.

VII.2 Results

To investigate the evolution of vertebrate gene structures, we compared approximately 40,000 intron positions in 3,603 sets of orthologous genes from *C. milii* and nine bony vertebrates, comprising the most extensive study of intron-exon evolution in vertebrates. The vast majority of intron positions in conserved protein-coding regions were intact between *C. milii* and all other vertebrate species. *C. milii* showed presence/absence differences from other vertebrate species at 43 positions, 15 of which were found to be due to changes in the *C. milii* lineage, indicating a low rate of change (15 changes per ~40,000 sites in ~450 My, or roughly 10^{-6} per site per My). The number of changes since the gnathostome ancestor is smaller in *C. milii* than in any bony vertebrate (**Fig. 2**), which is consistent with the lower rates of molecular evolution of *C. milii*. The higher rate of changes in bony vertebrates however, appears to mostly reflect a higher rate of change in the osteichthyan ancestor, the rates of change in *C. milii* being otherwise comparable to those in tetrapods. Rates of change were much higher overall in the two teleost fishes, and particularly in stickleback - 68% of all changes (and 83% of gains) were specific to stickleback, and 83% of all changes occurred in teleost fishes. Intron losses outnumbered intron gains in 13/17 branches, with the striking exception of stickleback, in which gains outnumbered losses 5-to-1 (603/126) (**Fig. 2**). Thus, the present study has identified the largest number of gains and losses of introns recorded in any vertebrate lineage. Previous genome-wide studies had found evidence mainly for loss of introns in mammals and teleosts [125-127] whereas our study shows that gain of introns is also prevalent in some subset of vertebrate lineages.

Supplementary Note VIII. Large-scale synteny analysis

VIII.1 Identification of syntenic blocks

The InParanoid orthologue gene sets for human-*C. milii*, chicken-*C. milii*, medaka-*C. milii* and zebrafish-*C. milii* were used for this analysis. i-ADHoRe v3.0 [128] was used to identify orthologous regions/syntenic blocks in the genome pairs compared. The following parameters were used: "alignment_method= gg4, anchor_points = 3, tandem_gap = 15, gap_size = 30, cluster_gap = 35, max_gaps_in_alignment = 35, q_value = 0.75, prob_cutoff = 0.01, level_2_only = false, multiple_hypothesis_correction = FDR". The program first identifies homologous regions (segments) in two genomes that contain at least three homologous genes (anchorpoints) with the anchorpoints separated by at most 30 non-homologous genes ('gap_size'). These form the base-clusters (with minimum quality factor of 0.75 and probability cut-off of 0.01), which are then grouped into larger syntenic blocks ('multiplicons') if they are within 35 genes ('cluster_gap') of each other. Considering only the non-redundant 'multiplicons' (syntenic blocks) and their corresponding 'anchor points' (homologous genes of the syntenic segments), syntenic blocks between *C. milii* scaffolds and chromosomes of other genomes were identified, and the number of orthologous genes in the syntenic blocks was tabulated.

VIII.2 Large-scale synteny conservation

By virtue of its phylogenetic position, *C. milii* is an ideal outgroup for inferring large-scale chromosomal rearrangements in bony vertebrate lineages and for inferring the ancestral gnathostome linkage groups. We carried out large-scale synteny comparison between *C. milii* and representative tetrapods (human and chicken) and teleost fishes (medaka and zebrafish). Comparisons with tetrapods revealed that 72% and 93% of *C. milii* syntenic scaffolds show conserved synteny with single chromosomes in human and chicken, respectively (**Supplementary Table VIII.1 and VIII.2**; **Supplementary Fig. VIII.1**). Interestingly, a majority of *C. milii* scaffolds showing synteny with two or more human chromosomes (21% out of 28%) correspond to single chicken chromosomes (**Supplementary Table VIII.2**), highlighting previously reported instances of interchromosomal rearrangements in the mammalian lineage [129,130]. On the other hand, there is one instance of *C. milii*-human one-to-one correspondence which shows a one-to-two relation with chicken chromosomes, a previously identified rearrangement in the chicken lineage [39,129]. Comparisons with teleost fishes showed that 88% and 74% of *C. milii* syntenic scaffolds show one-to-one or one-to-two correspondence with medaka and zebrafish chromosomes, respectively (**Supplementary Table VIII.3 and VIII.4**). Because teleost fishes have undergone an additional round of whole genome duplication (3R) [80,131], one-to-two correspondences were also considered as conserved synteny. Overall, these comparisons show that the *C. milii* genome has experienced a lower rate of interchromosomal rearrangements, comparable to the chicken genome which possesses the most stable karyotype among tetrapods [132,133]. Based on a comparison between human and medaka genomes using *Ciona* and sea urchin as outgroups, Nakatani et al. [130] had previously reconstructed an ancestral gnathostome karyotype comprising 40 proto-chromosomes. Our analyses of one-to-one conserved syntenic blocks between *C. milii*, chicken and human has identified seven novel syntenic relationships between chicken and human chromosomes that do not correspond to any of the reconstructed gnathostome proto-chromosomes (**Supplementary Table VIII.5**). These syntenic regions potentially represent additional ancestral gnathostome proto-chromosomes.

In addition to the one-to-one correspondence with tetrapod chromosomes, synteny of many large blocks of genes on *C. milii* scaffolds is extensively conserved in tetrapods. This is illustrated by the syntenic regions *C. milii* scaffold_14 (~10 Mb) and human chr_2q23.3 - 2q33.1 (45 Mb) (**Supplementary Fig. VIII.2**) that contain 148 syntenic genes including the HOXD gene cluster. HOXD cluster genes are regulated by evolutionarily conserved, long-range enhancers and global control regions located outside the HOX cluster and spread over several flanking non-Hox genes [134-136]. The HOXD locus thus represents a typical genomic regulatory block (GRB) characterized by large genomic regions containing several conserved regulatory elements and their target genes interspersed with 'bystander' genes [137]. The conserved syntenic regions identified in this study extend far beyond the previously identified GRB at the HOXD locus and raise the possibility of a much larger GRB than previously identified [136]. The extensive syntenic blocks conserved between the *C. milii* and human genomes should help to delineate many more potential GRBs in the human genome.

Avian karyotypes typically contain a large number of microchromosomes that are apparently derived from similar microchromosomes in the tetrapod ancestor [133,138]. Although the karyotype of *C. milii* is yet to be determined, it is likely to comprise mainly microchromosomes like those of its closely related holocephalan, the ratfish (*Hydrolagus colliei*) which contains 29 pairs of dot-like chromosomes resembling the avian microchromosomes [139]. Interestingly, 82 out of 86 *C. milii* scaffolds with homology to chicken microchromosomes show correspondence to a single chicken microchromosome each (**Supplementary Table VIII.6**). This suggests that the organization of avian microchromosomes might reflect the ancestral genome organization as exemplified by cartilaginous fishes. The remaining four *C. milii* scaffolds are each syntenic to a macro- and a microchromosome of chicken (scaffold_6: chromosomes 1 and 20; scaffold_19: chromosomes 4 and 13; scaffold_45: chromosomes 7 and 17; scaffold_109: chromosomes 3 and 14) (**Supplementary Table VIII.6**). These split linkages indicate fissions/translocations in the early tetrapod lineage or fusions/translocations in the *C. milii* lineage.

To assess the extent of interchromosomal rearrangements in teleosts, we compared gene clusters on *C. milii* scaffolds corresponding to a single chicken chromosome (hence representing ancestral gnathostome linkage groups) with medaka and zebrafish chromosomes. We identified 10 and 30 *C. milii* scaffolds that each showed correspondence to more than two medaka and zebrafish chromosomes, respectively (**Supplementary Table VIII.7 and VIII.8**). Based on a comparison of medaka and human genomes, it has been proposed that the teleost ancestor contained 13 proto-chromosomes before the 3R, and that there were eight major chromosomal rearrangements after the 3R but prior to the divergence of teleosts [45]. In addition, the zebrafish lineage is thought to have experienced approximately 14 major interchromosomal rearrangements, whereas no major rearrangements occurred in the lineage leading to medaka [45]. In the present study we have identified 8 and 25 additional interchromosomal rearrangements in the medaka and zebrafish lineages respectively, that were missed in the previous study (**Supplementary Table VIII.7 and VIII.8**). For example, "*Ancestral Chr-b*" reconstructed in the previous study [45] is syntenic to medaka chromosomes Ola11 and Ola16. However, a *C. milii* scaffold corresponds to these chromosomes in addition to two other chromosomes (Ola4 and Ola20) (**Supplementary Table VIII.7**). Thus, our comparisons revealed a substantially higher number of interchromosomal rearrangements in the teleost lineage than previously identified based on teleost-tetrapod comparison alone.

Supplementary Note IX. Evolution of protein domains and gene families

IXa. Protein domain analysis

Since this is the first cartilaginous fish genome to be sequenced, we carried out detailed characterization of *C. milii* proteins based on their domains and compared them with proteins from various bony vertebrates. This analysis should provide a comprehensive insight into the evolution of protein families in gnathostomes. In addition to *C. milii*, protein sequences from the following representative bony vertebrates were analyzed: human, mouse, cow, opossum, anole lizard, chicken, *X. tropicalis*, zebrafish and stickleback (Ensembl release 65).

Methods

The proteins were searched against the PFAM database (version 26) using HMMER version 3.0 (http://hmmer.janelia.org/) and the number of unique domains in each protein was counted. For genes with multiple isoforms, only the longest protein was considered. As each domain can be represented by more than one profile hidden Markov model, we combined the counts of proteins that mapped to different profile models under the same domain. For each genome, the percentage of proteins that mapped to a particular domain was calculated. The top 100 most abundant protein domains in the *C. milii* are given in **Supplementary Table IX.1**. A comparison of the top 50 protein domains in the *C. milii*, stickleback (a representative teleost) and human (a representative tetrapod) is shown in **Supplementary Fig. IX.1**.

Results

The *C. milii* genome encodes more proteins containing the immunoglobulin domain and B-box zinc finger domain than stickleback and human, whereas the human genome has an abundance of 7-TM receptor (rhodopsin family), C2H2-type zinc finger, zinc-finger double domain, olfactory receptor, KRAB box, serpentine type 7-TM GPCR chemoreceptor and immunoglobulin C1-set domains. The stickleback genome encodes more proteins with protein-kinase domain, NACHT domain, SPRY and SPRY-associated domain than *C. milii* and human. These comparisons highlight the protein families that have expanded independently in the three lineages.

A notable instance of a domain lost in teleosts is the progesterone receptor (PGR) domain. Even though teleost genomes encode a PGR protein[140], the PGR domain itself has substantially diverged. This may be related to the fact that while progesterone is the endogenous ligand for PGR in tetrapods and cartilaginous fishes[141], 17,20β-DHP and 20β-S are the main ligands for PGR in teleosts[142]. Our analysis also identified six domains shared by *C. milii* and teleost fishes but lost in tetrapods (**Supplementary Table IX.5**). One of these domains, 'sea anemone cytotoxic protein' is found in actinoporins, a highly potent family of pore-forming toxins produced by sea anemones [143]. It would be interesting to see what role these pore-forming toxin-like proteins are playing in *C. milii* and teleost fishes.

IXb. Evolution of protein-coding gene families in vertebrates

Methods

Genes lost specifically in tetrapods and teleost fishes

We used Ensembl Biomart to extract human orthologues of chicken, anole lizard, *Xenopus tropicalis*, zebrafish, medaka, stickleback and fugu genes. Human-tetrapod and human-teleost union sets were prepared from these orthologues. Human-tetrapod orthologues not present in teleost fishes were identified by comparing the human-tetrapod list with the human-teleost list. This 'tetrapod-specific' set was compared with elephant shark-human InParanoid orthologues to obtain elephant shark genes present in tetrapods but absent in teleost fishes. To identify genes lost in tetrapods, the zebrafish orthologues of medaka, stickleback, fugu, human, chicken, anole lizard and *X. tropicalis* were extracted from Ensembl Biomart. Zebrafish-teleost and zebrafish-tetrapod union sets were prepared and zebrafish-teleost genes not present in tetrapods were obtained by comparing the two union sets. Comparison of these 'teleost-specific' genes with elephant shark-zebrafish InParanoid orthologues highlighted the genes common to elephant shark and teleost fishes but lost in tetrapods. These sets were further refined by BLAST searches against the NCBI NR database and filtering proteins that matched any tetrapod or teleost protein in the respective analysis. The identified genes were annotated by searching for associated human (lost in teleosts) or zebrafish (lost in tetrapods) Gene Ontology (GO, biological process) terms using Ensembl Biomart. Additionally, we looked for the function of these genes by searching literature and databases such as GeneCards (v 3.0; http://www.genecards.org/), The Human Protein Atlas (http://www.proteinatlas.org/) (for genes lost in teleost fishes) and ZFIN (http://zfin.org/) (for genes lost in tetrapods).

Genes absent in bony vertebrates

We generated a union set of elephant shark InParanoid orthologues from thirteen bony vertebrates (human, mouse, cow, opossum, chicken, anole lizard, *X. tropicalis*, African coelacanth, zebrafish, stickleback, medaka, fugu and *Tetraodon nigrovirilis*). Comparison of this ‘elephant shark-bony vertebrate’ set with elephant shark genes identified a set of elephant shark genes that do not have a bony-vertebrate orthologue. Using custom Perl scripts, protein sequences of these genes were BLAST searched against the NCBI NR database (E-value threshold of ≤1e-5) to exclude genes that had a bony-vertebrate protein as the top hit. The remaining gene set was further curated by BLAST searches against the NCBI NR database and manual inspection of the alignments to exclude low complexity sequences and sequences whose hits included a bony vertebrate in the top 30 hits. To verify that these genes in the elephant shark assembly are not the result of contamination, we looked for their expression in the elephant shark by searching Trinity and Cufflinks RNA-seq transcripts obtained from 10 different tissues. The GO terms for these genes were extracted using Ensembl genome annotation pipeline. Finally, we performed domain prediction using SMART and searched literature to get an idea about the possible functions of these genes.

To identify genes present in tetrapods but lost in teleost fishes, we used the criterion that the gene should be present in human and at least one other tetrapod (chicken, anole lizard or *X. tropicalis*). This set was then compared with elephant shark-human orthologues to obtain the set of genes present in elephant shark and tetrapods but lost in teleost fishes. For the second set, i.e. genes present in elephant shark and teleost fishes but lost in tetrapods, we required that the gene should be present in zebrafish and at least one other teleost fish (medaka, stickleback or fugu). This set was compared to the elephant shark-zebrafish orthologues to obtain the genes present in elephant shark and teleost fishes but lost in tetrapods. Additionally, manual curation was done to ensure that the identified genes were indeed absent in teleost fishes or tetrapods.

Results

Tetrapod-specific gene losses

Functional annotation of the zebrafish orthologues of the 34 genes lost specifically in tetrapods (**Supplementary Table IX.7**) highlighted several genes that are specific to the aquatic lifestyle such as the innate immune system genes, fin and lateral line development genes, and olfactory receptor genes. The immune system-related genes include a member of the finTRIM (fish novel TRIM) family previously identified in teleosts whose expression is induced by viruses [144]. This gene family has expanded in teleost fishes [144,145] indicating that it plays an important role in innate immunity. This and the other immune system genes found in elephant shark and teleost fishes (cathepsin L.1-like gene, caspase 8-like gene, interferon phi gene and Rho-class glutathione S-transferase gene) are likely to be specific for counteracting aquatic pathogens. Three of the tetrapod-specific losses are related to fin development and include two actinodin genes. A previous study has shown that loss of these actinodin genes in tetrapods is linked to the fin-to-limb transition [146]. The third fin-related gene encodes an uncharacterized protein that is expressed in the zebrafish apical ectodermal ridge, pectoral fin bud and other regions of the fin (ENSDARG00000008732; **Supplementary Table IX.7**). This gene could be of interest in view of the high regeneration capacity of these organs. Two of the tetrapod-specific losses are represented by putative lateral-line genes. One of these (ENSDARG00000089429; **Supplementary Table IX.7**) with expression in the lateral line ganglion and neuromasts of zebrafish encodes an uncharacterized protein. The second (ENSDARG00000086369; **Supplementary Table IX.7**) encodes a fibroblast growth factor receptor 1-like protein, a component of the FGF signaling pathway that has been implicated in lateral line development [147]. The 'genes lost in tetrapods' set also includes an olfactory receptor gene which has expanded to a family of 76 genes in zebrafish (**Supplementary Table IX.7**). This gene family belongs to the ζ class of olfactory receptors which are specific receptors for aquatic odorants [148].

Invertebrate genes present in *C.milii* but lost in bony vertebrates

Our analysis identified 27 *C. milii* genes that have homologues in invertebrates but not in bony vertebrates (Supplementary Table VII.8). One of these is the *isopenicillin N epimerase* (*Ipne*) gene found widely in bacteria, fungi and invertebrates such as amphioxus, sea urchin, Pacific oyster, sea anemone and Trichoplax. Thus, this ancient gene has been apparently lost multiple times in invertebrates and vertebrates. The bacterial and fungal IPNEs convert isopenicillin N to penicillin N[149]. To our knowledge, the *C. milii Ipne* gene is the first instance of the presence of a gene involved in antibiotic synthesis in a vertebrate.

An intriguing instance is the cephalotoxin-like gene previously identified only in the cuttlefish, *Sepia esculenta* [150]. The multi-exonic *C. milii* gene encodes a protein with a similar domain organization (transmembrane, EGF, CCP, TSP1 and LDLa domains) to the cuttlefish protein. However, unlike the cuttlefish transcript that is expressed specifically in the posterior salivary glands, we observe multi-tissue expression (gills, intestine, kidney, liver, muscle, spleen and testis) in *C. milii*. Further detailed searches identified ESTs for this gene in the spiny dogfish, *Squalus acanthias* (accession number EC093606.1) and the marbled lungfish, *Protopterus aethiopicus* (FL669404.1 and FL669393.1) indicating its presence in a lobe-finned fish lineage besides cartilaginous fishes. Additionally, we searched the whole-genome sequences of the sea lamprey, African coelacanth, spotted gar and teleost fishes (zebrafish, medaka, stickleback, tilapia, *Tetraodon* and fugu) by TBLASTN using cuttlefish and *C. milii* protein sequences as queries. However, we did not find any homologs in these genomes. The intriguing pattern of distribution of this cephalotoxin-like gene raises questions about the origin of this gene in vertebrates: was it transferred horizontally from cuttlefish to the gnathostome ancestor (prey to predator) and subsequently lost multiple times in gnathostomes or is it an ancient gene that has been lost multiple times in invertebrates and vertebrates? Analysis of additional invertebrate and vertebrate genomes should clarify the origin of this unusual gene. In any case, the presence of cephalotoxin-like protein may explain why there is only one known predator of *C. milii*, the sevengill shark [151].

IXc. Olfactory and vomeronasal receptor genes

Olfactory receptor genes

Olfaction is vital for finding food, choosing mates and identifying offspring, and avoiding predators. Olfactory receptors (ORs) are responsible for detection of odorant molecules present in the environment. Vertebrate ORs belong to the rhodopsin-like G protein-coupled receptor (GPCR) superfamily, which is the largest family within the GPCRs. These proteins contain seven transmembrane α-helices typical of all GPCRs. Genome-scale analyses have revealed the presence of OR-like genes not only in vertebrates, but also in the nonvertebrate chordate amphioxus [148]. These studies have shown that the repertoire of OR genes within different groups is highly variable. Among mammals, humans contain 387 functional OR genes whereas opossum, rats and mice contain 1188, 1207 and 1035 functional OR genes, respectively [152]. Among non-mammalian vertebrates, chicken, *Xenopus* and zebrafish possess 211, 824 and 154 functional OR genes respectively, whereas the jawless vertebrate sea lamprey contains 32 functional OR genes [148]. Among the nonvertebrate chordates, amphioxus contains 31 functional OR genes, whereas tunicates, the closest phylogenetic group of vertebrates, lack vertebrate-type OR genes [148]. Previously, six OR-like genes have been identified from the 1.4× assembly of elephant shark including three which were truncated [148]. The availability of a high-quality assembled genome sequence of *C. milii* provided an opportunity to perform a genome-wide search in elephant shark to uncover its OR gene repertoire.

Olfactory receptor-like (OR-like) genes were searched in the *C. milii* genome by TBlastN using representative OR proteins as query and an E-value cut-off of 1e-10. Aligned regions in the genome identified by TBlastN were extracted with an additional 1 kb region on both 5' and 3' ends. The extracted sequences were used for BlastX against the NCBI non-redundant (nr) database to identify putative ORs. Amino acid sequences (longest open-reading frame) were extracted for the same region and were used for domain prediction using the SMART web-server [153]. A multiple alignment was then generated using the putative elephant shark OR-like genes and known ORs from a previous study [148] using the E-INS-i strategy as implemented in MAFFT version 6.864b [154]. Gaps in the alignment were removed using Gblocks [86]. A neighbor-joining tree was generated for the trimmed alignment using ClustalW (**Supplementary Fig. IX.2a**). Sequences which did not fall within a Type 1 or Type 2 clade were not considered an OR-like gene. In total, we could identify 6 OR-like genes (CmOR1, CmOR2, Cm-theta1, Cm-theta2.1, Cm-theta2.2 and Cm-kappa1) in the elephant shark genome. These genes are the same as those identified in a previous study [148], but the availability of a high quality genome assembly enabled us to obtain full-length sequences for the three truncated OR-like sequences. Of the six OR-like genes identified in elephant shark, only two are real ORs (CmOR1 and CmOR2). The remaining genes belong to groups θ1, θ2 or κ and it has been suggested that these groups are non-ORs [148]. CmOR1 lies within the Group η (eta) clade, whereas CmOR2 belongs to Group ζ (zeta) - both these groups are likely to be specialized for detection of water-soluble odorants [148].

Overall, *C.milii* contains the least number of OR genes among vertebrates. The small number of OR genes in *C. milii* could be related to its greater reliance on electroreceptors rather than ORs for seeking food on the ocean floor. *C. milii* possesses a characteristic fleshy, plough-shaped snout that is studded with ampullae of Lorenzini (electroreceptors) (**Supplementary Fig. IX.2b**). The snout is used to detect bioelectric fields generated by buried crustaceans, molluscs, echinoderms, and polychaetes[155]. Interestingly, the monotreme platypus possesses known electroreceptor capabilities in the bill and displays a similar distorted ratio of olfactory to vomeronasal receptor genes[156].

Vomeronasal receptor genes

Besides the main olfactory system (MOS), many tetrapods possess an accessory olfactory system known as the vomeronasal system (VNS). Central to the VNS is the vomeronasal or Jacobson's organ (VNO) located in the nasal cavity and the accessory olfactory bulb (AOB) located in the brain. The VNO has sensory neurons expressing vomeronasal receptors that are responsible for detection of intraspecific pheromonal cues and some environmental odorants [157]. Unlike tetrapods, which have two anatomically segregated olfactory tissues/organs (main olfactory epithelium and VNO), fishes have a single olfactory organ known as the olfactory rosette. Interestingly, although morphological traits of the VNS are found only in tetrapods, VNS related genes have been identified in teleost fishes [158,159]. In chondrichthyans, the olfactory organs are located in laterally placed cartilaginous chambers or sacs on the ventral surface of the head, anterior to the mouth [155]. Like teleost fishes, there is no distinct VNO in elephant shark. However, genes that express specifically in the VNS have been identified previously in elephant shark [160] suggesting that, like teleost fishes, two separate olfactory signaling pathways (main and vomeronasal) exist in cartilaginous fishes as well despite the absence of an anatomically distinct VNO.

The mammalian vomeronasal family can be subdivided into two subfamilies – vomeronasal receptor family 1 and 2 (V1R and V2R, respectively). Since teleost fishes do not have a VNO, it has been proposed that teleost V1R-like genes be termed as *ora* (ORs related to class A GPCRs) and V2R-like genes as *OlfC* (ORs related to class C GPCRs) [161,162]. Genomic surveys have revealed a considerable amount of variation in the number of vomeronasal receptor genes within different vertebrate groups. Rodents such as mouse and rat possess a large number of intact V1R (187 and 106, respectively) and V2R (70 and 59, respectively) genes. Opossum also possess a large number of intact V1R and V2R (98 and 79, respectively) genes. However, humans have lost all functional V2R genes and only retain five intact V1R genes. These five genes are likely to be remnants of an ongoing pseudogenization process in humans [163,164]. On the other hand, chicken, which does not possess a VNO, appears to lack both V1R and V2R genes. Within amphibians, the western clawed frog possesses 21 V1R and the largest number of V2R genes (249) amongst the vertebrates [163]. Amongst teleost fishes, pufferfishes contain just a single intact V1R-like gene, whereas zebrafish contains two intact V1R-like genes. The number of intact V2R-like genes is 4, 18 and 44 for *Tetraodon*, fugu and zebrafish, respectively [163]. In Atlantic salmon, 29 intact *OlfC* (V2R-like) genes were identified [165]. Mining of the 1.4× assembly of the elephant shark genome identified two V1R-like and 32 partial V2R-like sequences [160]. Here we report the full complement of vomeronasal receptor genes in the elephant shark.

Vomeronasal receptor genes were searched in the elephant shark genome by TBLASTN [166] using representative V1R and V2R sequences. Extracted regions of homology were BLASTX-searched against the NCBI non-redundant database to identify putative vomeronasal receptors. Partial V2R amino acid sequences identified previously [160] were also used as query. A multiple alignment was generated using MAFFT version 6.864b (E-INS-i strategy) [154] for the putative V1Rs and V2Rs together with sequences of previously identified vomeronasal receptors from zebrafish and sea lamprey. Non-OR GPCRs were used as outgroups. A neighbor-joining (NJ) tree was built using MEGA5 [106] with Poisson distance correction and 1000 bootstrap replicates for node support (**Supplementary Fig. IX.3**). Our analysis identified four V1R-like and 33 *OlfC* (V2R-like) genes in the elephant shark genome. Of the 33 *OlfC* (V2R-like) genes, one gene (SINCAMP00000008707) is orthologous to zebrafish OlfCx genes while another gene (SINCAMP00000025729) is related to zebrafish *OlfCc1* gene. The remaining 31 genes formed two separate clades comprising 25 and six genes within the OlfC/V2R clade suggesting that these genes have expanded in the elephant shark. Thus, unlike the limited number of olfactory receptor genes, the elephant shark possesses a large repertoire of vomeronasal receptor genes similar to teleost fishes. This presumably reflects the diverse pheromones employed by aquatic vertebrates for attracting mates and for warning conspecifics regarding potential predators and other dangers. Since fertilization is internal in elephant sharks, sex pheromones in particular may play an important role in attracting mature males to gravid females.

Supplementary Note X. Genes involved in bone formation

To determine whether *C. milii* contains the major genes involved in bone formation, we catalogued the genes that are known to be involved in bone formation and maintenance (**Supplementary Table X.1**) and searched for their orthologues in the *C. milii* genome. Starting from signaling pathways and their components that are involved in specification, commitment, patterning and proliferation of skeletal cells (chondrocytes, osteoblasts and osteoclasts), followed by regulatory transcription factors that define and control the behavior of these cell types, and finally the battery of differentiation genes directly involved in the deposition of the matrix that forms cartilage and bone. In the process, we discovered that almost the entire set of cartilage and bone formation genes are present in *C. milii* (**Supplementary Table X.1**), except for a family of genes called the *secretory calcium-binding phosphoprotein* (SCPP) genes that were derived from tandem duplications of the *Sparcl1* gene.

BMP signalling

Discovered for their ability to induce bone formation, BMP ligands control a diverse array of processes during bone development, including formation of mesenchymal condensations, differentiation of osteoblasts and coordination of the three-dimensional patterning of skeletal elements. We searched for the genes that alter the extracellular BMP gradients (*Noggin*, *Chordin*, *Follistatin*, *Gremlin*) and genes that encode BMP ligands (*Bmp2, 4, 5, 6, 7*), their receptors (*BmprI, BmprII*), intracellular transducers (*Smad1, 4, 5, 6, 7, 8*; *Smurf1, 2*), and found that they are intact in the *C. milii* genome (**Supplementary Table X.1**).

Hedgehog signalling

Indian Hedgehog (Ihh), a member of the vertebrate Hedgehog (Hh) family of ligands involved in homeostasis of the intestinal epithelium, is also required for endochondral bone development, controlling chondrocyte proliferation and maturation [167] as well as osteoblast development [168]. Perturbations in Ihh signalling impair long bone development in murine models and in humans [168]. Ihh functions through signaling components that are shared by all Hh ligands (Sonic-, Desert- and Indian Hedgehog), and Ihh activity in chondrogenesis is mediated by the parathyroid hormone signalling pathway (see below). *C. milii* possesses all three *Hh* genes found in mammals. It also contains genes for the components controlling the modification, release and movement of Hh ligands (*Hhat, Dispatched1, Scube2*, *Scube3, Hhip, Ext1, Ext2,* and *ExtL3*), their reception at the membrane and endocytosis (*Gas1, Cdo,* *Ptc1, Ptc2*, *Gpc3*) as well as their intracellular transduction (*Smo, Kif7, Sufu, Pka, Evc, Evc2, Gli1, Gli2, Gli3*) [169]. Therefore, we conclude that components of the Hh signal transduction machinery are present and functional in *C. milii*.

Parathyroid signaling

The parathyroid gene family consists of two members, *Parathyroid hormone* (*PTH*) and *Parathyroid hormone-related protein* (*PTHrP*). PTH has major roles in calcium homeostasis, where it acts to release calcium from bone and restricts its excretion via the kidney [170]. PTHrP on the other hand has a developmental role during endochondral ossification downstream of Ihh. Ihh prevents chondrocyte hypertrophy by positively regulating PTHrP expression, thereby maintaining a chondrocyte population capable of proliferation [167]. In addition, Ihh plays a distinct and possibly direct role in maintaining proliferation to produce long bones [168]. The lack of chondrocyte proliferation or their unscheduled hypertrophy results in premature differentiation of the cartilage without sufficient growth, resulting in shorter bones. We previously reported on the detailed characterization of the *Pth* gene family in *C. milii* and showed that *C. milii* retains three *Pth* gene family members, one of which (*Pth2*) was lost in the lineage leading to bony vertebrates [171]. In addition, we find distinct orthologues of Pth receptors, *Pthr1* and *Pthr2*, in the *C. milii* genome. These findings preclude the simple association of loss of endochondral bones in cartilaginous fishes to the evolution of the PTH gene family.

FGF signalling

FGF signalling was first implicated in skeletal development by the finding that gain-of-function mutations in their receptors resulted in reduced growth of long bones and achondroplasia, the most common form of dwarfism in humans [170]. Since then, a number of genetic studies from mice and human patients have implicated FGF signalling in almost every step of dermal and endochondral bone formation, including the modulation of chondrogenesis. We found intact copies of genes for FGF ligands (e.g. *Fgf1*, *Fgf23* and *Fgf2*), their receptors (*Fgfr1*, *Fgfr2*, *Fgfr3* and *Fgfr4*) and downstream mediators (*Ras*, *Rac1*, *Raf1*, *Mek*, Map-kinases, *Jnk*, *p38*, *Erk1* and *Erk2*), all implicated in skeletal development, in the *C. milii* genome.

RANK-RANKL-OPG pathway

Bone acts as a major source of calcium and goes through cycles of deposition and resorption. Resorption is largely coordinated by osteoclasts through the RANK-RANKL-OPG pathway. Osteocytes and osteoblasts secrete a bone-dissolving factor called RANKL that binds to its receptor RANK on the surface of osteoclasts and their progenitors. This in turn leads to the activation of osteoclasts resulting in bone resorption. To fine-tune this process of resorption, the bone-forming cells also secrete OPG, which acts to antagonize this pathway by binding to the RANKL ligand, thereby preventing it from binding RANK on osteoclasts. The *C. milii* genome contains genes for all three of these molecules.

Transcription factors

A number of transcription factors are known to play regulatory roles in the formation and maintenance of cartilage and bone. These include Sox5, Sox6 and Sox9 which together initiate and orchestrate the formation of cartilage. Other factors such as Bapx1, Sp7, Sp3, Atf4, Twist1, Twist2, Sox8, Atf4, Mef2c, c-FOS, Msx1 and Msx2 play important roles in bone development. We found genes for all of these factors in the *C. milii* genome, including the master regulator for bone development *Runx2*. We also found orthologues of factors that control the differentiation of osteoclasts, *Nfatc1*, *Pu.1/ Spi-1*, and *Mitf*.

Orthologues for genes encoding other factors that have been implicated in the modulation of bone strength by genome-wide association studies, like *Zbtb40*, *Ahsg*, *Sqstm1*, *Lrp5*, *Gpr177* and *Axin1* (the latter three are Wnt pathway components) are also found in *C. milii*.

Retention of the above-mentioned genes (signalling pathway components and *trans*-factors) in *C. milii* does not come as a surprise, as homologues of many of these pathway components and factors are used for the formation and patterning of diverse tissue types even in invertebrates. The rationale for searching these genes was to determine if there was a specific loss of components that largely seem dedicated to bone development.

Bone and cartilage differentiation genes

We next looked for the presence of downstream genes that participate in the terminal differentiation of skeletal cells, their deposition and modulation of extracellular matrix. It is possible that the lack of bone in cartilaginous fishes could be due to some deficiencies in the cartilage that impedes subsequent deposition of bone. Therefore, in addition to genes involved in bone formation, we also searched for genes participating in cartilage differentiation.

Proteoglycan genes

Proteoglycans constitute important regulators for the formation of cartilage and bone. We found intact orthologues for almost all the genes in the following proteoglycan categories - SLRP gene family clusters (*Fmod, Prelp* and *Optc*; *Ecm2, Aspn, Omd* and *Ogn*; *Dcn, Lum, Kera* and *Epyc*; *Bgn* but no *Ecm2l*), the lectican-HAPLN gene family clusters (*Hapln2* and *Bcan*; *Vcan* and *Hapln1*; *Acan* and *Hapln3*; *Ncan* and *Hapln4*) and other non-clustered proteoglycans (*Fn1*, *Lepre1*,*Tuft1*, *Podn*).

Cartilage genes

We were also able to find genes that are involved in cartilage formation such as *Col2a1* (Type II), *Col11a2* (Type XI), *Matrilin-1, -3*, *Mmp1*, *Mmp13*; genes modifying chondroitin sulfate (*GalNAc4S-6ST*) and heparin sulfate (*Hs2st*); *Cspg4*, *Cspg5* and *Chad* in the *C. milii* genome.

Bone differentiation genes

In order to identify genes involved more specifically in bone differentiation, we started by looking for genes encoding proteins that constitute a large portion of the bone matrix. These include *Type I collagen* (*Col1a1* and *Col1a2*), *type X collagen* (*Col10a1*), *osteocalcin* (*Bglap*), *Mgp* (*Bglap* and *Mgp* are closely linked as in teleosts), and *alkaline phosphatase*. All these genes are intact in *C. milii*. Additionally, we could also find other bone-specific genes like *ankylosis protein* (*Ank*), *Alox12*, *Bmp1*, *Cd44, Fam20C* (duplicate copies), *fibromodulin, osteoglycin, Calcium-sensing receptor* (*Casr*), *osteopotentia*, *osteocrin*, *osteoglycin, Sost, Sostdc1*, *Phex, Crtap*, *Cant1*, *Phospho1*, *Phospho2, Atp2b1*, *Enpp1, Sptbn1, Adamts18, Rspo3, Galnt3, Fam3c, Xylt1, Ext2, Papst1*, *Uxs1*, *Has2* and *Entpd5*.

Secretory calcium-binding phosphoprotein (SCPP), Sparc and Sparcl1 gene family

The survey of SCPP genes in teleosts, birds, reptiles and mammals has revealed a close correlation between the complexity of mineralized tissues and the repertoire of SCPP genes. This in turn has led to the hypothesis that the gain and specialization of SCPP genes may have supported the evolution of diverse mineralized tissues in distinct bony vertebrate lineages [172,173]. What has remained unclear, however, is the genetic composition of SCPP genes in cartilaginous fishes, the sister group of bony vertebrates, that lack endochondral bone [174]. We carried out an extensive search for *Sparc*, *Sparcl1* and *SCPP* genes in the *C. milii* genome.

The *C. milii* genome contains both *Sparc* and *Sparcl1* genes. However, there is neither SIBLING nor other SCPP gene in the *Sparcl1* or *Sparc* locus of *C. milii* (**Supplementary Figs. X.1 and X.2**). To verify whether there is an SCPP gene elsewhere in the *C. milii* genome, we did a relaxed search (E<10) of the *C. milii* genome assembly using TBlastN but did not identify any convincing homologues. Interestingly, in mammals and amphibians the non-clustered SCPP gene, *AMEL*, is located in an intron of *ARHGAP6* gene. Although *C. milii* genome contains an orthologue of the *ARHGAP6* gene, its introns do not contain any gene.

To verify if other cartilaginous fishes contain SIBLING or SCPP genes, we searched genomic resources of cartilaginous fishes available in the public domain. This includes 26× coverage survey sequence of the little skate (*Leucoraja erinacea*) [175], ESTs from the embryonic stages of little skate, small-spotted catshark (*Scyliorhinus canicula*) and *C. milii* (~300,000 ESTs) [175]; and elasmobranch ESTs in the NCBI database (~82,000 ESTs). However, we did not find any homologs of SIBLING or SCPP genes in these datasets. We also searched the jawless vertebrate genome resources comprising lamprey and hagfish ESTs in NCBI, and the genome assembly of sea lamprey (Pmarinus_7.0, January 2011), and could not identify any homologues of SIBLING or Pro/Gln-rich SCPP genes. However, in the sea lamprey, we did find two copies of *Sparc* (*SparcA* and *SparcB*; accession numbers ABM21522.1 and ABM21524.1) but no *Sparcl1*. Thus, the lack of SIBLING and other SCPP genes is common to all chondrichthyans and jawless vertebrates.

Functions of SIBLING proteins

The five mammalian SIBLING proteins constitute a significant fraction of the organic matter in bone. Yet their functions, as assayed by single gene knockouts in mice that have resulted in mild phenotypes, have been difficult to interpret. This has been attributed to compensatory and/or redundant mechanisms provided by the five proteins in bone deposition and resorption. Two trends emerge from the analysis in the mouse model: *Dmp1*, *Mepe* and *Ibsp* positively modulate mineralization in dentine and bone, whereas *Dspp* and *Spp1* do so negatively [176]. Nevertheless, it is clear that SIBLING proteins strongly associate with hydroxyapatite and modulate mineral crystallization. A distinct role for SIBLING proteins in direct binding to collagen fibrils and initiating their mineralization has been suggested based on *in vitro* experiments (reviewed in [176,177]). Further support for a critical role of SIBLING genes in ossification comes from unbiased genome-wide association studies (GWAS). Variations in *SPP1*, *MEPE* and *IBSP* loci are strongly associated with bone mineral density and fracture risk in humans [178-180]. The retention/expansion of different complements of *SIBLING* genes in different vertebrate lineages also suggests redundant and compensatory function amongst the mammalian SIBLING proteins.

Knockdown of spp1 in zebrafish

We genetically interfered with the function of the single bone-specific SCPP gene, *spp1* (also known as *osteopontin* or *opn*) in zebrafish by using two different methods: antisense morpholino-mediated knockdown and targeted genetic modifications using the CRISPR/Cas system[181]. Antisense morpholinos were targeted to either the ATG translation start site or the exon2-intron2 (E2-I2) splice junction of *spp1* pre-mRNA. The CRISPR/Cas system was used to target either exon 6 exon 7, or both exons; in the latter situation, large deletions in the order of ~2.6 kb are expected.

Methods for morpholino knockdown

Adult zebrafish (wild type AB strain) were maintained on a 14 hour light/10 hour dark cycle at 28°C in the Agri-Food & Veterinary Authority (AVA, Singapore)-certified IMCB Zebrafish Facility. Fish were raised according to the guidelines of the IMCB fish facility and the Biopolis IACUC protocol #100520. Morpholinos targeting ATG (ATG MO) or the junction of 2nd exon and intron (E2-I2 MO) of zebrafish *spp1* gene were designed and injected into 300 to 400 1-2 cell stage zebrafish embryos per day on two or more days. Either 1 nl or 2 nl of a 0.75 mM morpholino solution was injected. The sequences of the morpholinos used are as follows (lower case letters in control MO denote mismatches to their knock down MOs):
Spp1-ATG-1 (GTGTGCAAAATATTCTGCTCTCTCT),
Spp1-E2-I2 (ACTGATTGTGAACTTACAGGTACAC),
Cntrl-Spp1-ATG (GTcTcCAAAtTATTgTGgTCTCTCT),
Cntrl-Spp1-E2-I2 (ACTcATTcTcAACTTAgAcGTACAC).

The inhibition of splicing by E2-I2 MO was verified by RT-PCR using a combination of 3 primers (P1, exon 2 forward; P2, intron 2 reverse; and P3, exon 3 reverse) (**Supplementary Fig. X.5**). One microgram of total RNA extracted from pools of 10 embryos (4-dpf or 5-dpf) was used to synthesize cDNA with SuperScript® III Reverse Transcriptase (Invitrogen, Carlsbad, USA). The cDNA was used in a PCR reaction with the following primers:
zfSPPEx2F1, 5' GCACACAAAATGAAATCTATTATTG 3' (Exon 2 forward);
zfSPPIn2R1, 5' GCTAAGAACTGATTGTGAACTTAC 3' (Intron 2 reverse);
zfSPPEx3R1, 5' CCCGTTGAACAATTACAAGCTCTTC 3' (Exon 3 reverse).
PCR was performed using DyNAzyme (Finnzymes, Finland) using the following cycling conditions: 35 cycles of 95°C for 30s; 58°C for 1 min; 72°C for 20s; followed by a final extension of 72°C for 5 min.

Methods for CRISPR/Cas mediated genetic modifications

The Cas9 nuclease expression vector (pMLM3613) and the single guide RNA (sgRNA) expression vector (pDR274) were obtained through the nonprofit reagent distribution service Addgene (http://www.addgene.org/crispr/jounglab). Target sites and corresponding oligonucleotide pairs were selected and designed using the ZiFiT Targeter software (http://zifit.partners.org/) for exons 6 and 7 of the *zfspp1* gene.

Exon 6: Target site (forward strand): GGAATCTGAAACAGATGAGA
Oligo 1: TAGG**AATCTGAAACAGATGAGA**
Oligo 2: AAAC**TCTCATCTGTTTCAGATT**
Exon 7: Target site (reverse strand): GGTAGCCCAAACTGTCTCCC
Oligo 1: TAGG**TAGCCCAAACTGTCTCCC**
Oligo 2: AAAC**GGGAGACAGTTTGGGCTA**

Bold: complementary bases for annealing; underlined: overhang for cloning into the *Bsa*I-digested pDR274 (sgRNA expression vector).

Customized sgRNA expression vectors were obtained by cloning the annealed oligonucleotides into *Bsa*I-digested pDR274 vector; integrity of the clones was confirmed by sequencing. *Dra*I-digested sgRNA expression vectors were transcribed using the MAXIscript T7 Kit (Life Technologies). Following DNaseI treatment, the sgRNAs were purified using ammonium acetate-ethanol precipitation. The Cas9 expression vector was digested with *Pme*I and transcribed using the mMESSAGE mMACHINE T7 ULTRA Kit (Life Technologies). Poly(A) tailing and DNaseI treatment were performed according to manufacturer's instructions followed by lithium chloride precipitation for the Cas9-encoding mRNA. Approximately 200 to 300 one-cell stage zebrafish embryos were injected on two different days with 1 nl or 2 nl solution containing ~12.5 ng/µl sgRNA and ~300 ng/µl Cas9 mRNA[181]. The following combinations of sgRNA(s) and/or Cas9 mRNA were used:

1. *zfspp1* exon 6 sgRNA + *Cas9* mRNA
2. *zfspp1* exon 7 sgRNA + *Cas9* mRNA
3. *zfspp1* exon 6 + exon 7 sgRNAs + *Cas9* mRNA (this should cause a deletion of ~2.6 kb)
4. *zfspp1* exon 7 sgRNA alone (control)
5. *Cas9* mRNA alone (control)

Tail clips of normally developing embryos were used for isolation of genomic DNA for genotyping. We used the GoTaq® Green 2× Master Mix (Promega, USA) for genotyping PCR. The following genotyping primers were used:

(1) zfspp1_exon 6_F1: CACGTCAATGCACTCCCACAACAG
(2) zfspp1_exon 6_R1: CGACTCAAACCCATAACCTTGGCAC
(3) zfspp1_exon 7_F1: CGACCAGTGACATTTCACAGTGTTGC
(4) zfspp1_exon 7_R1: CTACTCCCGAGCTAAAACCACTACAG

Expected product sizes for exon 6 and exon 7 primer pairs from wild-type sequence are 315 bp and 487 bp, respectively. The expected size for a double deletion when using both exon 6 and 7 sgRNAs is ~450 bp (primers #1 and #4). PCR products were purified and cloned into pGEM®-T Easy vector (Promega, USA). Multiple clones were sequenced using the BigDye® Terminator Cycle Sequencing Kit (Applied Biosystems, USA) on ABI 3730xl capillary sequencers (Applied Biosystems, USA).

Staining embryos

Five-day old live zebrafish embryos were stained using Alizarin red (Sigma Aldrich, Sweden) that binds to mineralized matrix and fluoresces in the red spectrum. Live embryos were incubated overnight in 30 ml of "fish water" containing 200 µl of 0.5% Alizarin red [182]. Embryos were then rinsed in "fish water" and anesthetized using tricaine (Sigma Aldrich, Sweden). The Alizarin red stained embryo was observed under a compound microscope (Axio imager M2; Carl Zeiss, Germany) and imaged using an attached digital microscope camera (Axiocam; Carl Zeiss, Germany). All manipulations were done on entire images in ImageJ software (National Institutes of Health, USA). RNA *in situ* hybridization for *spp1* expression and Alcian blue staining for visualizing cartilage were done as previously described[183]. The mutants were scored as either ‘normal’ (resembling wild types), ‘mild’ bone-phenotype or ‘strong’ bone-phenotype with the latter showing the most reduction of bone.

Results

RNA *in situ* hybridization in embryos from 1 to 5 dpf showed that *spp1* is expressed specifically in cells surrounding the bone matrix starting from 2 dpf (**Supplementary Fig. X.3**) and precedes the deposition of bone. This provides support for its role in the deposition of bone. We note that the zebrafish *spp1* loss-of-function phenotype observed in our study (**Fig. 4; Supplementary Fig. X.4, X.6 and X.9**) is in contrast to the mouse *Spp1* knockout phenotype, which shows an increase in bone formation[184], possibly because of redundancy and compensation within the SIBLING gene family in mammals[176,185].

Brightfield images of *Cas9* mRNA and sgRNA injected embryos showed that their overall development is comparable to that of wild type embryos (**Supplementary Fig. X.7**). Alcian blue staining of cartilage showed that despite reduction in bone formation, cartilage formation in morpholino injected embryos (**Supplementary Fig. X.5**) and *Cas9* mRNA+ sgRNA injected embryos was similar to that in wild type embryos (**Supplementary Fig. X.5 and X.8**).

Supplementary Note XI. Analysis of the immune system of *C. milii*

XI.1 Strategy of analysis

The genome assemblies of *C. milii* and the transcriptomes derived from several organs (see **Supplementary Note I**) were interrogated for the presence or absence of representative genes relevant to mammalian innate and adaptive immune facilities. Human sequences were used as initial queries and subsequently complemented by sequences of teleost and chondrichthyans (cartilaginous fishes) as required. When the *C. milii* databases indicated the potential absence of relevant sequences, the thymus and spleen transcriptomes of the nurse shark *Ginglymostoma cirratum* (see **Supplementary Note I**) and the databases of other cartilaginous fishes (little skate, catshark and dogfish)[175] were additionally examined.

XI.2. Antigen recognition

XI.2.1. Antigen receptor genes

Key features The structures of *Ig* and *TCR* genes in *C. milii* are similar to their counterparts in other cartilaginous fish. The close linkages of IgH and the TCR alpha/delta loci, and that of IgH and MHC (suggested by the linkage of *IgM* and *Trim69* genes) appear to be ancient features of gnathostome genomes. Furthermore, IgL genes are situated next to MHC paralogous genes, indicating that the IgL precursor was located near the primordial MHC; this ancestral linkage has been lost in bony vertebrates.

XI.2.1.1 Immunoglobulin genes

XI.2.1.1.1. IgH genes

Cartilaginous fish IgM genes are found in the so-called cluster organization, instead of the translocon arrangement seen in tetrapod species [186]. The apparent number of IgM loci in *C. milii* as revealed by genomic hybridization analysis using CH4 domain sequences as probes (**Supplementary Fig. XI.1**) is consistent with corresponding BLAST searches, in which 32 scaffolds gave positive hits (scaffolds 54, 121, 290, 325, 1166, 1253, 1290, 1436, 1549, 1564, 1570, 1816, 1861, 1873, 2013, 2158, 2427, 2476, 2666, 2709, 2843, 3307, 3412, 3547, 4561, 4836, 4891, 4913, 5858, 6954, 9077, 9876); these results suggest that the *C. milii* genome encodes about twice as many IgM loci than that of the nurse shark *G. cirratum*, which possesses about 15 IgM loci [187]. The distribution of IgM genes in the *C. milii* genome is not known, although two genes are found on scaffold_121. The *Trim69* gene of *C. milii* is found next to an IgM gene on scaffold_325. Interestingly, *Trim69* maps next to the gene encoding β2-microglobulin (the light chain of the MHC class I complex) in the human genome [188]; in nurse shark, the β2-microglobulin gene maps to the MHC locus [189].

Based on their VH sequences, three types of IgM genes were found that are quite similar to those of another holocephalin, *Hydrolagus*, which diverged from *C milii* ~170 million years ago[190]. (i) There are approximately 10 conventional IgM genes, with spacer sequence lenths in recombination signal sequences (RSS) identical to all other cartilaginous fish genes (V23, 12D23, 12D12, 23J). (ii) One unconventional IgM gene appears to encode a single-chain IgM, likely orthologous to the one first found in *Hydrolagus*. The VH domain lacks amino acids critical for association with VL domains, and the CH1 domain is similar in sequence to the CH2 domain, suggesting that the original CH1 domain exon was lost and replaced by a CH2 duplication. This putative single-chain IgM is found only in the holocephalins, and may have been superseded by the IgNAR class found in all elasmobranchs (see below). (ii) There are many 'orphan' VH genes, apparently lacking downstream CH exons; maintenance of open-reading frames in most of these genes suggests a novel function for these loci.

In addition to IgM genes, cartilaginous fish possess alternative immunoglobulin heavy chain genes [191]. Interestingly, no evidence for genes encoding the IgNAR and IgW isotypes could be found in the *C. milii* genome, despite the fact that V elements of the NAR-TCR isotype are present [44], linked to *IgM* genes. This may suggest that a single-domain V element evolved as a component of a TCR-like gene and was only subsequently transferred to an Ig-like gene. The NAR-TCR genes are linked to the TCR alpha/delta locus [44,192], and one VH fragment was found among a cluster of TCRalpha/delta V genes (scaffold_220); moreover, there is one CH domain downstream of this VH element. At the end of scaffold_220, another VH domain was found; this is part of a conventional IgM gene that contains at least VH-CH1-CH2-CH3-CH4 and is situated in opposite orientation relative to the TCRalpha/delta V cluster. This is compatible with close linkage of TCRalpha/delta locus to IgH or IgH elements in many non-placental mammals, birds, and *Xenopus* [193-195].

XI.2.1.1.2. IgL genes

Four isotypes of immunoglobulin light chain have so far been found in cartilaginous fishes; the kappa, lambda and sigma isotypes are present in all cold-blooded vertebrates; the sigma-prime, a so-called "dead-end" isotype has been found only in cartilaginous fishes [196]. Out of four isotypes, we identified three IgL types from the *C. milii* genome; IgL sigma was not detected.

We found ~20 IgL lambda genes, as judged by the presence of adjacent V and C domains (on scaffolds 215, 221, 1021, 1159, 2856, 3081, 3848, 5255, 20085); as in all cartilaginous fishes [197], lambda genes encode "germline-joined" V domains, possibly indicating that lambda genes might be related to the primordial V domain subject to RAG transposon insertion. Two IgL kappa genes were identified in scaffold_101; both V domains are not germline-joined and their RSS are similar to other vertebrate IgL kappa genes. Only a constant domain was identified for the IgL sigma-prime gene on scaffold_3225.

XI.2.1.2. T cell receptor genes

As expected, all four types of T cell receptors were found in the *C. milii* genome. With the exception of TCR gamma, they are found on small scaffolds. The TCR gamma gene (situated between nt. ~1870000 and 1900000 on scaffold_83) is flanked by the orthologous genes found on *H. sapiens* chromosome 7p14, indicating conserved synteny between *C. milii* and human, with the exception of a block inversion. As expected, the TCR gamma gene is in translocon organization, having a V cluster upstream of at least one J and one C domain; the recombination signal sequences (RSS) conform to the pattern found in all other vertebrate TCR gamma genes (23 bp spacer for V, 12 bp spacer for J elements)[198]. Similar to all other vertebrate species, the TCR alpha/delta genes appear to be adjacent in the *C. milii* genome, as evidenced by the interspersion of Valpha and Vdelta genes; however, the exact architecture of the TCR alpha/delta locus could not be determined, as the scaffolds did not contain the respective C domain genes.

XI.2.3. Structure of the major histocompatibility complex (MHC)

Key features | The presence of four MHC paralogous groups in the *C. milii* genome is compatible with two rounds of genome duplications in the ancestor of cartilaginous fishes. Polymorphic MHC class I and class II genes were identified, with the latter type of genes present in fewer numbers.

We first analysed the presence of MHC paralogous groups in the *C. milii* genome. The conceptually translated sequences of all known genes that map to the human MHC and three other MHC-paralogous regions were used as queries against the *C. milii* databases. For approximately half of human MHC-encoded genes (291 of 634 human sequences used as queries) orthologues were identified in the *C. milii* genome and confirmed by phylogenetic analyses (data not shown).

In the human genome, genes of the MHC paralogous group are distributed mainly on the specific regions of four sets of chromosomes (chromosomes 1, 6, 9 and 19), with some genes located on chromosomes 12, 15, and others [199]. Several *C. milii* scaffolds contained genes whose human orthologues map to one of the MHC paralogous regions (**Supplementary Fig. XI.2**). Most synteny seems to be conserved (except in inversions). As expected, some genes map to different paralogous regions in *C milii* compared to other vertebrates, perhaps due to the differential silencing after genome duplication. Unfortunately, we could not assemble the MHC in detail since these scaffolds were generally short.

Analyses of genome assembly and transcriptomes of *C. milii* revealed evidence for both alleles of a single MHC class Ia gene, and multiple (2~3) MHC class I-related genes, one of which is unlinked to the MHC (data not shown). The presence of a single class Ia is consistent with previous work with other shark species[200,201]. In contrast, only one canonical MHC class II alpha gene and one MHC class II beta gene were detected; consistent with studies in all fish species, no DM (or DO) loci were detected.

XI.2.4. Antigen presentation

Key features | Many elements of the antigen presentation pathways (**Supplementary Table XI.1**) through MHC class I and II molecules known from mammals are present in *C. milii*, compatible with the presence of polymorphic MHC class I and II genes.

All three immunoproteasome components for presentation of antigens via MHC class I molecules were found, including PSMB8 and PSMB10, located next to each other on scaffold_1084 [202]. The IFNγ-inducible PSMB9 proteasome subunit could not be identified in *C. milii* databases, but is present in the transcriptomes of *G. cirratum*; the PSMB11 subunit which is specifically expressed in mammalian cortical thymic epithelial cells appears to be a pseudogene in *C. milii*. Many genes relevant for the MHC Class II pathway were found, with the exception of Cathepsin S; by contrast, the number of Cathepsin L-like genes is greatly increased (13 genes), eight of which are located in tandem on scaffold_85.

XI.2.5. NK receptors

Key features | Despite their rapid evolution, some conserved representatives of NK receptor genes could be identified in *C. milii*.

Because NK receptors generally evolve rapidly, it is essentially impossible to establish orthology of these genes. The *C. milii* genome lacks genes containing C2-type immunoglobulin superfamily domains (IgSF) (similar to KIRs) or lectin-type (similar to Ly49). There was no activating receptor similar to NKG2D, NKp44, or NKp56; however, an NKp30/NCR3-like receptor (*H. sapiens,* ENSP00000365240) was identified (scaffold_5779; SINCAMP00000007648) [203]; NKp30 contains a V-type IgSF with germline- joined configuration, representing a primordial VJ-type of IgSF domains.

XI.2.6. Pattern recognition receptors

Key features | The TLR system of *C. milii* consists of 10 genes and is distinguished by the lack of identifiable TLR4 receptor components; important representatives of intracellular helicases and NOD-like proteins are present. A large number of NLRP3-like genes and several other upstream components of the inflammasome were identified; the AIM2 component appears to be missing.

XI.2.6.1. TLRs and signaling components

Nucleic acid sensing TLRs (TLR3, 7, 8, 9) are clearly recognizable in the *C. milii* genome (**Supplementary Table XI.2**). For the other TLRs, assignment of two TLR1/6/10 homologues is easily possible. For the related TLR2/5 genes, two homologues are found. Two genes are unassigned relative to human TLRs and probably are “fish-specific” TLRs. A clear TLR4 orthologue is missing; however, a pseudogene fragment related to TLR4 is located in a syntenic region on scaffold_45 between DBC1 and ASTN2 at ~ 3,886,000 bp. This suggests that a precursor of a complete TLR4 gene was lost from the *C. milii* genome. No expansion of TLR gene number relative to mammals was noted. The extracellular components (LBP, CD14, MD-2) of the TLR4-specific pathway could not be found, although several homologues of the LBP-related BPI encoding genes are present in the *C. milii* genome. The lack of all of these components is consistent with the failure to induce LPS responsiveness in elasmobranchs (Helen Dooley and Martin Flajnik, unpublished). Interestingly, a clear TRAM homolog could be found, indicating the loss of this component in teleosts (TRAM is not found in zebrafish and cod) is derived. This conclusion is supported by the analysis of the lamprey genome (http://www.ensembl.org/Petromyzon_marinus/Info/Index; version 7), in which two TRIF/TRAM encoding gene could be detected (presumptive lamprey TRIF, GL478822 [nt. 16796-17242]; presumptive lamprey TRAM, GL478822 [nt. 30868-31548]). Other key intracellular signaling components (MYD88 etc.) are present in *C. milii* (**Supplementary Table XI.2**).

XI.2.6.2. Other pathogen receptors

Important cytoplasmic sensing receptors, such as DHX58, IFIH1, DDX58 helicases are present, as are homologues of NOD1 and NOD2 (**Supplementary Table XI.3**).

XI.2.6.3. Inflammasome components

With respect to inflammasome components, downstream effectors, such as IL1B and IL18 are present in the *C. milii* genome (**Supplementary Table XI.3**). Several of the known upstream components, such as NLRP proteins appear to be present; however, owing to the fact that they are all large multidomain proteins sharing considerable sequence similarity, further work is required to determine the sizes of these gene families. Nonetheless, it is notable that we have identified at least 57 NLRP3-like genes in the genome of *C. milii*. For many of these sequences, full-length transcripts were obtained and related sequences are present in the transcriptomes of *G. cirratum* as well; interestingly, the conceptually translated full-length transcripts all lack the characteristic pyrin domain that mediates homotypic protein interactions between NLRP3 and ASC. No evidence for an AIM2 gene could be detected in the respective syntenic regions of the *C. milii* genome.

XI.2.7. Complement system

Key features | The complement system of *C. milii* is reminiscent of other cartilaginous fishes and appears to be essentially complete when compared to the situation in mammals (**Supplementary Table XI.4**).

As expected, genes belonging to the three complement pathways [204] are present in the genome of *C. milii*. There are additional loci for C1r (Genbank accession numbers JW865071 and JW873796) and C1s (SINCAMG00000010562 and Genbank accession number JW866600), which may be unique to this species. Similar to the situation in nurse shark [205], there are two factor B genes (Genbank accession numbers JW864721 and JW865377). Most complement regulatory genes were also found, suggesting that the complement system is similar to that of mammalian system.

XI.3. Communication, Coordination, Response

Key features | The evolutionary trajectory of the CXC class of chemokines and receptors is less dynamic than that of the CC class of chemokines/receptors, which are more numerous in *H. sapiens*. All of the basic components of the interferon type-I and type-II receptor/ligand system are present in the genome of the common ancestor of the cartilaginous and bony fish, but have undergone lineage-specific modifications and expansions. Despite the evolutionary distance between *H. sapiens* and *C. milii*, many components of the well-described human cytokine system could be identified in the *C. milii* genome; a perplexing aspect of the *C. milii* genome is the near complete complement of IL2RG-family receptor components, but paucity of corresponding ligands, particularly those associated with T helper cell responses in mammals.

XI.3.1. Chemokines and their receptors

Phylogenic analyses for both chemokines and chemokine receptors suffer from low bootstrapping values. In some cases, syntenic relationships help to resolve ambiguities in the assignment of orthology. Overall, the number of CC chemokines is 28 for *H. sapiens*, and 14 for *C. milii*; for two homeostatic chemokines (CCL19 and CCL25) and one dual-function chemokine (CCL20), orthology could be established (**Supplementary Table XI.5**). The genomes of *H. sapiens* and *C. milii* both encode 17 CXC chemokines; orthology could be established for CXCL8, CXCL12 (two copies are found in the *C. milii* genome) and CXCL14 (**Supplementary Table XI.5**). Neither XC-like nor CX_3C chemokines were found.

Ten CCR-type receptors are encoded in the genome of *H. sapiens*, but only 4 in *C. milii*, those orthologous to CCR4, CCR6, CCR7, and CCR9 (**Supplementary Table XI.6**). The lower number of CCR receptors in the *C. milii* genome corresponds to the presence of fewer CC chemokines. For most of the 7 CXCR-type genes in *H. sapiens*, orthologues can be defined in *C. milii*; the same is true for the other six chemokine receptors that are encoded in the genomes of *H. sapiens* and *C. milii*. Our analysis is consistent with recent work on the phylogeny of chemokines/receptors[206].

XI.3.2. Interferons and their receptors

XI.3.2.1. Interferon receptor components

At the human interferon receptor locus on chromosome 21, four genes encoding components of the type-1 (IFNAR1 and IFNAR2) and type-II (IFNGR2) and type-III (IL10RB) interferon receptors are present, all facing in the same orientation (**Supplementary Fig. XI.3a**). At the corresponding *C. milii* locus (scaffold_152), we identified four interferon-receptor-like genes in the same orientation, together with a fifth gene in the opposite orientation; a similar locus containing five interferon receptor-like genes is also present in the genome of *T. rubripes* (**Supplementary Fig. XI.3a**). The human IFNGR1 gene is located on human chromosome 6q23, closely linked to the human IL22RA2 and IL20RA genes. This association is an ancestral feature, as it is also conserved in the *C. milii* genome on scaffold_26 (**Supplementary Fig. XI.3a**); however, additional receptor genes are present at the *C. milii* locus; namely two IFNGR1-like genes (SINCAMP00000024611 and SINCAMP00000024606), and two IL22RA2-like genes (see **Supplementary Tables XI.7a,b**). The human IL28RA gene is located on chromosome 1p36.11, closely linked to the IL22RA1 gene; a homologous gene cluster was identified in the *C. milii* genome (IL28RA, SINCAMP00000014734; for IL22RA1, see **Supplementary Tables XI.7a,b**) on scaffold_36 (**Supplementary Fig. XI.3a**). This receptor gene cluster includes the linked, but functionally unrelated gene, MYOM3. Overall, the striking syntenic conservation of the *C. milii* and human interferon receptor loci indicates that the genomic configuration of these receptors was set prior to the divergence of the cartilaginous and bony fish. This syntenic conservation may be more variable in teleosts, at least in the case of *D. rerio*, where genes for type-I (*crfb1* and *crfb5*) and type-II (*crfb2*, *crfb13* and *crfb17*) interferon receptor chains map to discrete loci with no evidence of receptor gene clustering.

XI.3.2.2. Interferons

XI.3.2.2.1. Type-I interferons

In humans, sixteen type-I IFN genes are found in a distinct cluster on human chromosome 9p22, which contains 16 protein coding IFN genes (13 IFNA genes, plus one IFNB, one IFNE and one IFNW gene) and 12 IFN pseudogenes; a simplified representation of this locus (excluding pseudogenes) is depicted in **Supplementary Fig. XI.3b**. Although we could identify a *C. milii* scaffold containing orthologues of the genes that flank the human type-I

IFN locus, we could not identify any linked IFN-like genes within the same scaffold, indicating that the *C. milii* genome lacks a type-I IFN locus that is homologous to the human 9q22 locus. Using the human IFNA2, IFNB1, IFNW1, IFNE and IFNK proteins as query sequences, three type-I IFN genes were identified in the *C. milii* genome assembly. Two of these genes are found closely linked on scaffold_185 (**Supplementary Fig. XI.3b**); based on the presence or absence of conserved cysteine residues [207], they can be classified as members of subgroups I and II, respectively. The location of these genes, next to the *C. milii* GH1 orthologue, is homologous to the *D. rerio* locus that encodes three type-I IFN genes, IFNphi1-IFNphi3 located on zebrafish chromosome 3. Interestingly, the genes flanking the *C. milii* (and to a lesser degree *D. rerio*) type-I IFN locus display some syntenic conservation with human chromosome 17q23.3, which contains the GH1 gene. The third *C. milii* type-I IFN gene mapped to a single gene scaffold (scaffold_3336; SINCAMP00000015870), precluding further investigation based on syntenic assignments. The structural similarities of growth hormone and interferons and the close proximity of their genes raise intriguing questions about their possible evolutionary relationship.

XI.3.2.2.2. Type-II interferons

The locus encoding IFN-γ appears to be well conserved amongst cartilaginous fish, teleosts and mammals. In humans, a single IFNG gene is located on chromosome 12q14.3-q15, next to the genes for IL26 and IL22. A homologous locus, which encodes IFNG and IL22 genes, is present in the *C. milii* genome, on scaffold_133 (**Supplementary Fig. XI.3b**). Several syntenic genes are conserved in this region, although no orthologue for IL26 was found in the *C. milii* scaffold. By contrast, the configuration on *D. rerio* chromosome 4 is similar to the human genome, although in this case two IFNG genes are present (**Supplementary Fig. XI.3b**).

XI.3.2.2.3. Type-III interferons

To date, type III interferons have only been identified in mammals and birds. Although we were able to find putative orthologues of both chains of the type-III IFN receptor (IL28RA and IL10RB), we were unable to identify any orthologues for type-III interferons in the *C. milii* genome.

XI.3.3. Interleukins and cytokines and their receptors (**Supplementary Tables XI.7a,b,c**)

XI.3.3.1. Common-γ-chain cytokines

In humans, six interleukins (IL2, IL4, IL7, IL9, IL15 and IL21) are known to signal via receptors that pair with the common-γ chain (Cgamma), encoded by the *IL2RG* gene. *IL2RG* is located on the X chromosome of all mammals investigated to date; two *IL2RG* genes are found clustered on scaffold_2 of the *C. milii* genome, which, based on synteny appears to be the *C. milii* homologue of the human X chromosome. The presence of two linked *IL2RG* genes in the *C. milii* genome is analogous to the case reported in several teleosts (zebrafish, trout, pufferfish), and suggests that a duplication of the IL2RG gene had taken place prior to the divergence of cartilaginous and bony fish, and that one copy was later lost at some point prior to the divergence of the mammalian lineage.

We were able to identify potential orthologues for all of the receptors known to pair with Cgamma, except for *IL2RA*. In humans, the genes encoding the *IL2RA* and *IL15RA* chains are located together on chromosome 10p15.1 (**Supplementary Fig. XI.4a**). BLASTp searches of the *C. milii* predicted protein database identified a candidate *IL15RA* gene on scaffold_17. Investigation of the surrounding genes revealed what appears to be a break in synteny centered upon the *IL15RA* gene in *C. milii*: four genes that map immediately upstream of the human *IL15RA* gene are conserved in *C. milii* on scaffold_17, while three genes immediately downstream map to scaffold_191 (**Supplementary Fig. XI.4a**). No *IL2RA*-like gene could be identified on either scaffold, and BLAST searches failed to reveal any alternative candidates in the *C. milii* genome. An *IL15* orthologue is situated on scaffold_57 in the *C. milii* genome; however, orthologues for *IL2* and *IL21* are not present at the predicted syntenic position on scaffold_208 (**Supplementary Fig. XI.4b**). In line with the results from *C. milii*, interrogation of the nurse shark thymus and spleen transcriptomes identified orthologues of *IL2RB*, *IL2RG*, *IL15RA* and *IL15*, but neither *IL2RA* nor *IL2*. Given that some teleost fish (such as trout and fugu) appear to have *IL2RA* and *IL2* genes [208], while others (e.g., zebrafish) lack both the receptor and ligand, it is possible that the IL2/IL2RA ligand-receptor combination may have evolved early in the teleost lineage, subsequent to the divergence of bony and cartilaginous fish. Our search for additional Cgamma cytokines led to the identification of the *IL7* gene on scaffold_48. The *C. milii IL7* gene eluded initial BLAST-based searches, but was finally identified by searching for an orthologue of ZC2HC1A, a gene linked to the human *IL7* locus, and subsequently interrogating the surrounding genes. A *C. milii IL7R* orthologue was identified in the transcriptome database; however, this gene is not present in the current genome assembly.

BLAST- and synteny-based searches failed to identify any other IL2RG-ligands, although in several cases syntenic regions lacking the expected cytokine genes were found. An example is shown for the case of the *IL4* gene, which appears to be absent from the *C. milii* genome (**Supplementary Fig. XI.4c**). In humans, *IL4*, *IL13* and *IL5* genes are found clustered together at chromosome 5q31.1. BLAST searches failed to identify candidate orthologues for any of these cytokine genes in the *C. milii* genome, however *RAD50* and *KIF3A*, which flank the human *IL13* and *IL4* genes, were identified on scaffold_92 in the same tail-to-tail arrangement, but with no intervening gene. This is in contrast to the case in teleosts, where *IL4* (linked to KIF3A) and *IL13* (linked to RAD50) candidates have been identified (**Supplementary Fig. XI.4c**). Our failure to identify *IL4* and *IL13* genes is surprising, because candidate orthologues for their unique receptors (*IL4R*, *IL13RA1* and *IL13RA2*) are clearly present in the *C. milii* genome. The human *IL4R* gene is located at chromosome 16p12.1, together with the *IL21R* gene (**Supplementary Fig. XI.4d**). BLAST searches identified candidate *IL4R* and *IL21R* genes in the *C. milii* genome, clustered together on scaffold_147. We noticed that two genes (*JMJD5* and *NSMCE1*) on one side of the IL4R/IL21R locus were conserved between humans, *D. rerio* and *C. milii*; however, on the other flank no syntenic genes could be identified (**Supplementary Fig. XI.4d**). Additionally, the number and composition of the receptor genes found at this locus differs between species: two in humans (*IL4R* and *IL21R*), two in zebrafish (*IL4R* and an unclassified receptor) and three in the case of *C. milii*. Of the three genes present at the *C. milii* locus, two could be identified as *IL4R*- and *IL21R*-like; however, the third receptor was difficult to assign, and may be the *C. milii* equivalent of *IL9R* based on sequence similarity.

IL13 shares some functional similarities with IL4. The genes encoding IL13-receptor components, *IL13RA1* and *IL13RA2* are present in the *C. milii* genome (**Supplementary Fig. XI.4e**); this contrasts with the apparent absence of an *IL13*-like gene.

XI.3.3.2. IL6ST-family cytokines

IL6ST (gp130) is the common signalling chain component of the receptor complex for several important cytokines, including IL6, IL11, IL27, IL31, LIF, OSM and CNT. In humans, the *IL6ST* gene is found together with the related *IL31RA* gene on chromosome 5q11.2. An analogous locus was identified in the *C. milii* genome on scaffold_22; this locus showed evidence of syntenic conservation, but contains three *IL6ST*-like genes (**Supplementary Fig. XI.4f**). Our analysis suggests that all three of the genes present at the *C. milii* locus are more closely related to *IL6ST* than to *IL31RA*, a result consistent with the fact that neither *IL31* nor *OSMR* (encoding the second component of the IL31-receptor complex) genes could be detected in the *C. milii* genome. We identified several genes encoding receptor chains that associate with IL6ST, including *IL11RA*, *IL27RA*, *LIFR* and *CNTFR*. Interestingly, although two potential orthologues of *IL6* clustered together on scaffold_4 of the *C. milii* genome, an obvious *IL6R* orthologue could not be identified. In addition to *IL6*, we also identified a *LIF* orthologue in *C. milii*, but failed to identify an *IL11* gene, despite the presence of an *IL11RA* gene. Two heterodimeric ligands, IL27 and CNT, signal via a receptor complex that includes IL6ST. In both cases we could identify the genes encoding their cognate receptor chains (*IL27RA* and *CNTFR*), but genes for only one component of each heterodimeric ligand; *EBI3* in the case of IL27 and *CRLF1* for CNT.

XI.3.3.3. Colony-stimulating factors

The colony stimulating factors CSF1, CSF2 and CSF3 play important roles in myelo- and granulopoiesis. We were able to identify genes for CSF1R, and the CSF1R-ligand IL34 in the *C. milii* genome, but failed to identify a *CSF1* gene - an interesting result given the importance of CSF1 in bone homeostasis in mammals (see **Supplementary Note X**). An orthologue of the human *CSF2RB* gene was identified on the *C. milii* scaffold_299. *CSF2RB* encodes the common chain of the CSF2, IL3 and IL5 receptor complexes in humans, and BLAST searches revealed a cytokine receptor gene cluster on scaffold_18 that may contain the *C. milii* orthologues of *CSF2RA*, *IL3RA* and *IL5RA* genes. Five cytokine-receptor genes are present at the scaffold_18 locus, and the surrounding genes displayed syntenic conservation with the human *CRLF2/CSF2RA/IL3RA* receptor gene cluster on the X-chromosome at p22.33 (**Supplementary Fig. XI.4g**). Although it is difficult to confidently assign all five *C. milii* genes to human orthologues, it is conceivable that this locus encodes the *C. milii* equivalents of the *IL3R*, *IL5R* and *CSF2RA* genes.

Although we were able to identify putative receptor chains for CSF2, IL3 and IL5, we failed to detect genes for any of the ligands themselves. We did however identify a *CSF3* orthologue on scaffold_251; and a *CSF3R* candidate on scaffold_64, supported by sequence similarity, but not syntenic conservation.

XI.3.3.4. IL12/IL23 family

Genes for both chains of the IL12-receptor (*IL12RB1* and *IL12RB2*) were identified in the *C. milii* genome, but two genes corresponding to one of the components of the heterodimeric IL12 ligand (p40), *IL12B*, could be identified. This suggests the intriguing possibility that IL12 is composed of a heterodimer of two p40 homologues instead of the mammalian p35/p40 heterodimer. No orthologue encoding IL23R (which pairs with IL12RB1 to form the IL23-receptor complex) or IL23A was detectable. Since Il23 is a potent inducer of Th17 cells in mammals, this finding has important implications for the diversity of T cell lineages in *C. milii* (see section XI.5).

XI.3.3.5. IL 10 family

In humans, the *IL10* gene is located on chromosome 1 at q32.1-q32.2, within a cluster that contains the genes of four related cytokines – IL10, IL19, IL20 and IL24. A homologous cluster is present in the *C. milii* genome on scaffold_70 (**Supplementary Fig. XI.4h**). The *C. milii IL10* cluster contains three genes – one of these genes appears to be an *IL10* orthologue, and while the remaining two genes clearly belong to the IL10 gene family, they cannot be confidently assigned as direct orthologues of particular mammalian IL10-family members. In addition to the *IL10*-family members found clustered on scaffold_70, we identified an *IL22* orthologue on scaffold_133, as part of the *C. milii IFNG* cytokine cluster. No *IL26* candidate was identified; the absence of *IL28A*, *IL28B* and *IL29* is discussed in section XI.3.2. Orthologues of genes encoding all of the components of the IL10, IL19, IL20 and IL24 receptors (*IL10RA*, *IL10RB*, *IL20RA*, *IL20RB*, *IL22RA1* and *IL28RA*) were found in the *C. milii* genome, although assignment of the putative *IL20RB* orthologue is based on sequence data alone, and was not supported by syntenic conservation.

XI.3.3.6. IL17 family

There is good correspondence between receptors and ligands. One interesting aspect is that the Th2 cytokine IL17E (also known as Il25) is absent (as is its receptor); this is in line with the absence of other Th2-type interleukins (see **Supplementary Tables XI.7a,b,c**).

XI.3.4. TNF and TNF receptors (**Supplementary Table XI.8**)

Key features | Many members of the TNFR gene family and its ligands, which are key regulators of the adaptive immune system, are present in the genome of *C. milii*; notable exceptions are the genes encoding LTRβ and its ligands LTα and LTβ, known to regulate secondary lymphoid tissue formation in mammals.

It is notable that orthologues of the key co-stimulatory molecules TNFRSF4 (OX40), TNFRSF5 (CD40) and TNFRSF8 (OX30) were confidently identified. The same is true for TNFRSF11A (RANK), and TNFSF11 (RANKL), suggesting that their role in formation of the thymic medulla is evolutionarily ancient. Because of the additional role of RANK/RANKL in bone formation and the presence of TNFRSF11B (OPG), it appears that these receptor/ligand pairs predated the evolution of calcified bones (see **Supplementary Note X**). Orthologues of the key molecules involved in B cell homeostasis, TNFRSF13B (TACI) and TNFRSF17 (BCMA) were identified.

Some notable differences to the mammalian gene complements are:

(a) The human TNFRSF members TNFRSF1A, TNFRSF3 and TNFRSF7 all map close to the CD4 locus; none were clearly identified in the *C. milii* genome, suggesting that this entire genomic region is of more recent evolutionary origin. (b) No orthologue of the TNFRSF7 ligand TNFSF7 was identified in the *C. milii* genome. (c) While orthologues of TNFRSF9 (FAS) and TNFSF6 (FASL) were identified, no clear orthologues of TNFRSF10A-D were identified, although TNFSF10 (TRAIL) is present. (d) The ligands for TNFRSF1A and TNFRSF3 are located in the MHC region of *H. sapiens*; whether they are in the MHC region of *C. milii* is unclear.

XI.4. Effector cells and lymphoid organs

Key features | The most conspicuous differences between the immunogenome of *C. milii* and mammals were identified in genes relevant for differentiation of T cell lineages; the implications of these findings are discussed in section XI.5 and the main text.

XI.4.1. Lymphocyte-related genes

The *C. milii* genome encodes all lineage regulators known in mammals to effect the development and differentiation of B and T lymphocyte subsets, with several notable exceptions (**Supplementary Table XI.9**). First, a RORC orthologue could not be identified in any of the genomic and transcriptomic databases of *C. milii* and other cartilaginous fishes, suggesting that LTi-like cells, important for secondary lymphoid organ development [209] are not present in cartilaginous fishes (see XI.4.2). Second, although a FOXP3-like gene could be identified, it lacks the critical amino acid residues in the DNA binding domain required for its immunosuppressive functions [210]; **Supplementary Fig. XI.5**). Whereas a ZBTB7B/ThPOK gene, encoding a critical regulator of the CD4+ T cell lineage [211] could not be identified in the genome assembly, transcripts encoding related sequences were found in the transcriptomes of several cartilaginous fish, such as *C. milii*, *S. canicula*, *R. erinacea* and *Ginglymostoma cirratum* (**Supplementary Fig. XI.6**).

Our analysis revealed the presence of characteristic representatives of signalling components (including LCK), co-stimulatory molecules and their receptors and genes whose mutation is known to lead to primary immunodeficiencies (**Supplementary Table XI.9**). Likewise, key elements of the apoptosis pathways are present in the *C. milii* genome (**Supplementary Table XI.10**).

XI.4.1.1 CD8 coreceptors

The search for T lineage-associated co-receptor genes resulted in the identification of bona fide CD8A and CD8B genes (**Supplementary Fig. XI.7**; **Supplementary Table XI.11**); the characteristic interaction motifs are conserved in the conceptually translated protein sequences of LCK and CD8A genes (**Supplementary Fig. XI.8a**). Note that teleost CD8A sequences possess a functionally equivalent CxH motif, instead of the canonical CxC motif found in tetrapod CD8A sequences [212].

XI.4.1.2 CD4 coreceptors

Our search for a CD4 gene in the genome of *C. milii* and the transcriptomes of *C. milii* and *G. cirratum* proved unsuccessful (see **Supplementary Figs. XI.8, XI.9, XI.10** and **Supplementary Table XI.12** for phylogenetic analyses of vertebrate CD4 proteins and those encoded by potential candidate genes from cartilaginous fishes, and **Supplementary Table XI.13** for a summary of their structural hallmarks). While it was possible to identify a gene encoding a protein whose extracellular domain structure resembles that of CD4 and LAG3 (**Supplementary Fig. XI.8c**), the predicted protein lacks the critical CxC motif (**Supplementary Fig. XI.8b**) in an otherwise unusually long intracytoplasmic domain; this LAG3-related protein is also found expressed in the nurse shark transcriptome libraries of thymus and spleen. We also determined the expression levels of this gene (designated CD4-

R/LAG3 in **Supplementary Fig. XI.8d**) from transcriptome sequence data using RSEM v1.2.3 [213] and found that it is expressed 2-3 orders of magnitude less than the corresponding co-receptor *CD8A* gene; furthermore, it appears to be expressed at higher levels in peripheral lymphoid tissues (RT-PCR data not shown). Collectively, these findings do not support the presence of a bona fide *CD4*-like gene in cartilaginous fishes.

XI.4.1.3 T Lymphocyte lineages

The fact that all relevant CD8-lineage determinants are present in the *C. milii* genome indicates that cytolytic T cell lineages, such as NK and CD8+ T cells are present in cartilaginous fishes. The present analysis does not support the presence of a canonical CD4+ T helper cell lineage, although the presence of an unconventional CD8- T helper cell lineage cannot be excluded (see main text for further discussion).

In analogy to the smaller range of potential helper T lineages, the genome analysis also suggests that the types of innate lymphoid cells [214] in cartilaginous fishes might also be much smaller, possibly only comprising so-called group 1 ILCs, characterized by their ability to produce IFNγ. Note that group 2 ILCs are defined as producers of classical Th cytokines such as IL4, IL5, and IL13, all of which are missing in cartilaginous fishes; likewise, because group 3 ILCs depend on *RORC* for their development, they are also likely absent.

XI.4.2. Genes related to the formation and function of lymphoid organs

Cartilaginous fishes possess a thymus, the primary lymphoid organ critical for T cell development [215]. Indeed, the gene encoding the FOXN1 transcription factor, known to be essential for the differentiation of thymic epithelial cells [216] and the gene encoding the autoimmune regulator AIRE [217] are present (**Supplementary Table XI.9**). Likewise, the *HOX11* gene, a key component of the genetic pathway required for spleen development [218] could be unambiguously identified (**Supplementary Table XI.9**).

However, cartilaginous fishes, like all ectothermic vertebrates, lack lymph nodes and recognizable germinal centres. Indeed, a number of genes known to be critical for their development in mammals are missing in the genome of *C. milii*. First, a RORC orthologue, which in mammals is important for the development of lymphoid tissue inducer cells (LTi) [214] could not be identified in any of the genomic and transcriptomic databases of cartilaginous fishes (**Supplementary Table XI.9**). This is in line with the lack of critical regulators of early steps of lymph node development, such as LTβR and its ligands LTα and LTβ (**Supplementary Table XI.8**), and a crucial chemokine, CXCL13 (**Supplementary Table XI.5**). Moreover, IL21, a cytokine critically involved in the development of the T_{FH} subset of T cells [219] is also absent (**Supplementary Tables XI.7a,b,c**).

XI.5 Summary of major findings

The presence of elaborate innate immune functions in *C. milii* is supported by an essentially modern form of the complement system, a diverse repertoire of pathogen receptors, such as TLR- and NOD-like receptors and intracytoplasmic helicases, upstream and downstream effectors of the inflammasomes, and the basic components of the interferon system; a notable exception is the apparent lack of a TLR4-related receptor and associated components of this signalling pathway among the ten identifiable TLR-like genes, which is consistent with LPS non-responsiveness in sharks (M. Flajnik et al., unpublished observations). The immunoglobulin (Ig) genes are in the cluster-type organization found in all chondrichthyans, and the entire Ig system is most similar to another holocephalan, the ratfish [190]. *TCR* genes are found in the typical translocon organization of gnathostomes, including close linkage of the TCR alpha and delta loci [220]. Ig heavy (H) genes are linked to TCR loci, likely an ancestral feature of antigen receptors. The presence of four MHC paralogous groups in the *C. milii* genome (**Supplementary Fig. XI.2**) is compatible with two rounds of genome duplication in the ancestor of gnathostomes. The analysis of TNF and TNFR gene families suggests, among others, the presence of TNFSF2 (TNFα), an important regulator of inflammatory responses [221] and that of the TNFRSF11A (RANK) and TNFSF11 (RANKL) receptor/ligand pair, essential morphogenetic regulators [221]. Consistent with the lack of lymph nodes and germinal centres as well as the relatively long lag-time required to generate humoral immunity [222], the genes encoding the mammalian regulators of secondary lymphoid tissue formation, TNFRSF3 (LTβR) and its ligands (TNFSF1 [LTα] and TNFSF3 [LTβ]) [221] are absent from the *C. milii* genome, as is a critical cytokine of follicular helper T cells, IL21 [219]. By contrast, key determinants of formation and function of spleen (such as HOX11) [218] and thymus (FOXN1) [216], regulating differentiation of thymic epithelial cells, and AIRE [217], a key regulator of central tolerance) are present.

All hallmarks of the cytotoxic CD8 lineage of T cells are present in the *C. milii* genome (**Fig. 6a**). Surprisingly, however, despite the presence of polymorphic MHC class II and invariant chain genes a bona fide *CD4* gene is absent, as are genes encoding transcription factors regulating the differentiation of several T helper lineages and several of their key effector cytokines (**Fig. 6b**). The gene encoding FOXP3, the essential regulator of the CD4+ regulatory T (T_{reg}) cells, while present, lacks the structural hallmarks of its mammalian orthologues (**Supplementary Fig. XI.5**) [210]; in support of the lack of bona fide T_{reg} cells, we note the absence of the gene encoding IL2, a key regulator of T_{reg} cells in mammals [223], and its specific receptor, IL2RA We also noted the conspicuous absence of genes encoding cytokines of the T helper lineage (IL4, IL9, IL13, IL17E/IL25, IL31) from an otherwise seemingly modern complement of interleukin genes, whereas the gene encoding IFNγ, a classical Th1 cytokine, is present in *C. milii* and nurse shark. Several members of the IL10 family of anti-inflammatory cytokines are encoded in the in *C. milii* genome, suggesting that the balance between pro- and anti-inflammatory functions can be achieved without a dedicated regulatory T cell subset. *RORC*, the gene encoding RORγt, and genes encoding IL23 and IL23RA, which potentiate Th17 responses, are not present in cartilaginous fishes, suggesting that T_H17 cells [224] are not present; thus, in cartilaginous fish IL17 and IL22 might be furnished by non-lymphoid cells. The lack of the RORγt transcription factor also impacts the composition of the different types of innate lymphoid cells (ILCs) [214]. Cartilaginous fishes might only possess group 1 ILCs, characterized by their expression of IFNγ, but possibly neither group 2 ILCs, because the relevant effector cytokines (e.g. IL-4) are missing, nor group 3 ILCs (including lymphoid tissue inducer cells), because of their dependence on RORγt.

References

38 Miller, J. R., Koren, S. & Sutton, G. Assembly algorithms for next-generation sequencing data. *Genomics* **95**, 315-327, doi:10.1016/j.ygeno.2010.03.001 (2010).

39 Consortium, I. C. G. S. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature* **432**, 695-716, doi:10.1038/nature03154 (2004).

40 Dalloul, R. A. *et al.* Multi-platform next-generation sequencing of the domestic turkey (Meleagris gallopavo): genome assembly and analysis. *PLoS Biol* **8**, doi:10.1371/journal.pbio.1000475 (2010).

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

42 Trapnell, C., Pachter, L. & Salzberg, S. L. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**, 1105-1111, doi:10.1093/bioinformatics/btp120 (2009).

43 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:10.1038/nbt.1621 (2010).

44 Venkatesh, B. *et al.* Survey sequencing and comparative analysis of the elephant shark *(Callorhinchus milii)* genome. *PLoS Biol.* **5**, e101, doi:10.1371/journal.pbio.0050101 (2007).

45 Kasahara, M. *et al.* The medaka draft genome and insights into vertebrate genome evolution. *Nature* **447**, 714-719, doi:10.1038/nature05846 (2007).

46 Star, B. *et al.* The genome sequence of Atlantic cod reveals a unique immune system. *Nature* **477**, 207-210, doi:10.1038/nature10342 (2011).

47 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).

48 Bailey, J. A. *et al.* Recent segmental duplications in the human genome. *Science* **297**, 1003-1007, doi:10.1126/science.1072047 (2002).

49 Olshen, A. B., Venkatraman, E. S., Lucito, R. & Wigler, M. Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics* **5**, 557-572, doi:10.1093/biostatistics/kxh008 (2004).

50 Locke, D. P. *et al.* Comparative and demographic analysis of orang-utan genomes. *Nature* **469**, 529-533, doi:10.1038/nature09687 (2011).

51 Scally, A. *et al.* Insights into hominid evolution from the gorilla genome sequence. *Nature* **483**, 169-175, doi:10.1038/nature10842 (2012).

52 Maddock, M. B. & Schwartz, F. J. Elasmobranch cytogenetics: methods and sex chromosomes. *Bull. Mar. Sci.* **58**, 147-155 (1996).

53 Fujita, M. K., Edwards, S. V. & Ponting, C. P. The Anolis lizard genome: an amniote genome without isochores. *Genome Biol Evol* **3**, 974-984, doi:10.1093/gbe/evr072 (2011).

54 Fearnhead, P. & Vasileiou, D. Bayesian analysis of isochores. *Journal of the American Statistical Association* **104**, 132-141, doi:10.1198/jasa.2009.0009 (2009).

55 Costantini, M., Cammarano, R. & Bernardi, G. The evolution of isochore patterns in vertebrate genomes. *BMC Genomics* **10**, 146, doi:10.1186/1471-2164-10-146 (2009).

56 Costantini, M., Clay, O., Auletta, F. & Bernardi, G. An isochore map of human chromosomes. *Genome Res* **16**, 536-541, doi:10.1101/gr.4910606 (2006).

57 Costantini, M., Auletta, F. & Bernardi, G. Isochore patterns and gene distributions in fish genomes. *Genomics* **90**, 364-371, doi:10.1016/j.ygeno.2007.05.006 (2007).

58 Friedlander, M. R. *et al.* Discovering microRNAs from deep sequencing data using miRDeep. *Nat Biotechnol* **26**, 407-415, doi:10.1038/nbt1394 (2008).

59 Kozomara, A. & Griffiths-Jones, S. miRBase: integrating microRNA annotation and deep-sequencing data. *Nucleic Acids Res* **39**, D152-157, doi:10.1093/nar/gkq1027 (2011).

60 Heimberg, A. M., Cowper-Sal-lari, R., Semon, M., Donoghue, P. C. & Peterson, K. J. microRNAs reveal the interrelationships of hagfish, lampreys, and gnathostomes and the nature of the ancestral vertebrate. *Proc Natl Acad Sci U S A* **107**, 19379-19383, doi:10.1073/pnas.1010350107 (2010).

61 Peterson, K. J., Dietrich, M. R. & McPeek, M. A. MicroRNAs and metazoan macroevolution: insights into canalization, complexity, and the Cambrian explosion. *Bioessays* **31**, 736-747, doi:10.1002/bies.200900033 (2009).

62 Berezikov, E. Evolution of microRNA diversity and regulation in animals. *Nat Rev Genet* **12**, 846-860, doi:10.1038/nrg3079 (2011).

63 Heimberg, A. M., Sempere, L. F., Moy, V. N., Donoghue, P. C. & Peterson, K. J. MicroRNAs and the advent of vertebrate morphological complexity. *Proc Natl Acad Sci U S A* **105**, 2946-2950, doi:10.1073/pnas.0712259105 (2008).

64 Zhou, B., Wang, S., Mayr, C., Bartel, D. P. & Lodish, H. F. miR-150, a microRNA expressed in mature B and T cells, blocks early B cell development when expressed prematurely. *Proc Natl Acad Sci U S A* **104**, 7080-7085, doi:10.1073/pnas.0702409104 (2007).

65 Xiao, C. *et al.* MiR-150 controls B cell differentiation by targeting the transcription factor c-Myb. *Cell* **131**, 146-159, doi:10.1016/j.cell.2007.07.021 (2007).

66 Rayner, K. J. *et al.* MiR-33 contributes to the regulation of cholesterol homeostasis. *Science* **328**, 1570-1573, doi:10.1126/science.1189862 (2010).

67 Wijesekara, N. *et al.* miR-33a modulates ABCA1 expression, cholesterol accumulation, and insulin secretion in pancreatic islets. *Diabetes* **61**, 653-658, doi:10.2337/db11-0944 (2012).

68 Harris, R. S. Improved pariwise alignment of genomic DNA. *PhD. Thesis, The Pennsylvania State University* (2007).

69 Blanchette, M. *et al.* Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res* **14**, 708-715, doi:10.1101/gr.1933104 (2004).

70 Siepel, A. & Haussler, D. Phylogenetic estimation of context-dependent substitution rates by maximum likelihood. *Mol Biol Evol* **21**, 468-488, doi:10.1093/molbev/msh039 (2004).

71 Siepel, A. *et al.* Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* **15**, 1034-1050, doi:10.1101/gr.3715005 (2005).

72 Visel, A. *et al.* ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature* **457**, 854-858, doi:10.1038/nature07730 (2009).

73 Visel, A., Minovitsky, S., Dubchak, I. & Pennacchio, L. A. VISTA Enhancer Browser--a database of tissue-specific human enhancers. *Nucleic Acids Res* **35**, D88-92, doi:10.1093/nar/gkl822 (2007).

74 Lindblad-Toh, K. *et al.* A high-resolution map of human evolutionary constraint using 29 mammals. *Nature* **478**, 476-482, doi:10.1038/nature10530 (2011).

75 Pennacchio, L. A. *et al.* In vivo enhancer analysis of human conserved non-coding sequences. *Nature* **444**, 499-502, doi:10.1038/nature05295 (2006).

76 Shin, J. T. *et al.* Human-zebrafish non-coding conserved elements act in vivo to regulate transcription. *Nucleic Acids Res* **33**, 5437-5445, doi:10.1093/nar/gki853 (2005).

77 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).

78 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:10.1093/molbev/msq304 (2011).

79 Brunet, F. G. *et al.* Gene loss and evolutionary rates following whole-genome duplication in teleost fishes. *Mol Biol Evol* **23**, 1808-1816, doi:10.1093/molbev/msl049 (2006).

80 Jaillon, O. *et al.* Genome duplication in the teleost fish *Tetraodon nigroviridis* reveals the early vertebrate proto-karyotype. *Nature* **431**, 946-957, doi:10.1038/nature03025 (2004).

81 Guo, B., Zou, M. & Wagner, A. Pervasive indels and their evolutionary dynamics after the fish-specific genome duplication. *Mol Biol Evol* **29**, 3005-3022, doi:10.1093/molbev/mss108 (2012).

82 Chan, Y. F. *et al.* Adaptive evolution of pelvic reduction in sticklebacks by recurrent deletion of a *Pitx1* enhancer. *Science* **327**, 302-305, doi:10.1126/science.1182213 (2010).

83 McLean, C. Y. *et al.* Human-specific loss of regulatory DNA and the evolution of human-specific traits. *Nature* **471**, 216-219, doi:10.1038/nature09774 (2011).

84 Remm, M., Storm, C. E. & Sonnhammer, E. L. Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J Mol Biol* **314**, 1041-1052, doi:10.1006/jmbi.2000.5197 (2001).

85 Alexeyenko, A., Tamas, I., Liu, G. & Sonnhammer, E. L. Automatic clustering of orthologs and inparalogs shared by multiple proteomes. *Bioinformatics* **22**, e9-15, doi:10.1093/bioinformatics/btl213 (2006).

86 Castresana, J. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol Biol Evol* **17**, 540-552 (2000).

87 Stamatakis, A. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* **22**, 2688-2690, doi:10.1093/bioinformatics/btl446 (2006).

88 Ronquist, F. *et al.* MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Syst Biol* **61**, 539-542, doi:10.1093/sysbio/sys029 (2012).

89 Keane, T. M., Creevey, C. J., Pentony, M. M., Naughton, T. J. & McLnerney, J. O. Assessment of methods for amino acid matrix selection and their use on empirical data shows that ad hoc assumptions for choice of matrix are not justified. *BMC Evol Biol* **6**, 29, doi:10.1186/1471-2148-6-29 (2006).

90 Jones, D. T., Taylor, W. R. & Thornton, J. M. The rapid generation of mutation data matrices from protein sequences. *Comput Appl Biosci* **8**, 275-282, doi:10.1093/bioinformatics/8.3.275 (1992).

91 Altekar, G., Dwarkadas, S., Huelsenbeck, J. P. & Ronquist, F. Parallel Metropolis coupled Markov chain Monte Carlo for Bayesian phylogenetic inference. *Bioinformatics* **20**, 407-415, doi:10.1093/bioinformatics/btg427 (2004).

92 Shimodaira, H. & Hasegawa, M. CONSEL: for assessing the confidence of phylogenetic tree selection. *Bioinformatics* **17**, 1246-1247, doi:10.1093/bioinformatics/17.12.1246 (2001).

93 Blair, J. E. & Hedges, S. B. Molecular phylogeny and divergence times of deuterostome animals. *Mol Biol Evol* **22**, 2275-2284, doi:10.1093/molbev/msi225 (2005).

94 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:10.1093/molbev/msn278 (2009).

95 Arnason, U., Gullberg, A., Janke, A., Joss, J. & Elmerot, C. Mitogenomic analyses of deep gnathostome divergences: a fish is a fish. *Gene* **333**, 61-70, doi:10.1016/j.gene.2004.02.014 (2004).

96 Janvier, P. *Early vertebrates*. (Oxford University Press, 1996).

97 Arnason, U., Gullberg, A. & Janke, A. Molecular phylogenetics of gnathostomous (jawed) fishes: old bones, new cartilage. *Zool Scr* **30**, 249-255, doi:10.1046/j.1463-6409.2001.00067.x (2001).

98 Rasmussen, A. S. & Arnason, U. Molecular studies suggest that cartilaginous fishes have a terminal position in the piscine tree. *Proc Natl Acad Sci U S A* **96**, 2177-2182, doi:10.1073/pnas.96.5.2177 (1999).

99 Takezaki, N., Figueroa, F., Zaleska-Rutczynska, Z. & Klein, J. Molecular phylogeny of early vertebrates: monophyly of the agnathans as revealed by sequences of 35 genes. *Mol Biol Evol* **20**, 287-292, doi:10.1093/molbev/msg040 (2003).

100 Kikugawa, K. *et al.* Basal jawed vertebrate phylogeny inferred from multiple nuclear DNA-coded genes. *BMC Biol* **2**, 3, doi:10.1186/1741-7007-2-3 (2004).

101 Hallstrom, B. M. & Janke, A. Gnathostome phylogenomics utilizing lungfish EST sequences. *Mol Biol Evol* **26**, 463-471, doi:10.1093/molbev/msn271 (2009).

102 Roy, S. W. & Gilbert, W. Resolution of a deep animal divergence by the pattern of intron conservation. *Proc Natl Acad Sci U S A* **102**, 4403-4408, doi:10.1073/pnas.0409891102 (2005).

103 Roy, S. W. & Irimia, M. Rare genomic characters do not support Coelomata: intron loss/gain. *Mol Biol Evol* **25**, 620-623, doi:10.1093/molbev/msn035 (2008).

104 Takezaki, N., Rzhetsky, A. & Nei, M. Phylogenetic test of the molecular clock and linearized trees. *Mol Biol Evol* **12**, 823-833 (1995).

105 Nei, M. & Kumar, S. *Molecular evolution and phylogenetics*. (Oxford University Press, 2000).

106 Tamura, K. *et al.* MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol Biol Evol* **28**, 2731-2739, doi:10.1093/molbev/msr121 (2011).

107 Venditti, C., Meade, A. & Pagel, M. Detecting the node-density artifact in phylogeny reconstruction. *Syst Biol* **55**, 637-643, doi:10.1080/10635150600865567 (2006).

108 Suyama, M., Torrents, D. & Bork, P. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res* **34**, W609-612, doi:10.1093/nar/gkl315 (2006).

109 Hubisz, M. J., Pollard, K. S. & Siepel, A. PHAST and RPHAST: phylogenetic analysis with space/time models. *Brief Bioinform* **12**, 41-51, doi:10.1093/bib/bbq072 (2011).

110 Lanave, C., Preparata, G., Saccone, C. & Serio, G. A new method for calculating evolutionary substitution rates. *J Mol Evol* **20**, 86-93, doi:10.1007/BF02101990 (1984).

111 Paradis, E., Claude, J. & Strimmer, K. APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics* **20**, 289-290, doi:10.1093/bioinformatics/btg412 (2004).

112 Martin, A. P., Naylor, G. J. & Palumbi, S. R. Rates of mitochondrial DNA evolution in sharks are slow compared with mammals. *Nature* **357**, 153-155, doi:10.1038/357153a0 (1992).

113 Martin, A. P. Substitution rates of organelle and nuclear genes in sharks: implicating metabolic rate (again). *Mol Biol Evol* **16**, 996-1002 (1999).

114 Abramyan, J. *et al.* The western painted turtle genome, a model for the evolution of extreme physiological adaptations in a slowly evolving lineage. *Genome Biol* **14**, R28, doi:gb-2013-14-3-r28 [pii]10.1186/ (2013).

115 Britten, R. J. Rates of DNA sequence evolution differ between taxonomic groups. *Science* **231**, 1393-1398, doi:10.1126/science.3082006 (1986).

116 Martin, A. P. & Palumbi, S. R. Body size, metabolic rate, generation time, and the molecular clock. *Proc Natl Acad Sci U S A* **90**, 4087-4091, doi:10.1073/pnas.90.9.4087 (1993).

117 Schlotterer, C., Amos, B. & Tautz, D. Conservation of polymorphic simple sequence loci in cetacean species. *Nature* **354**, 63-65, doi:10.1038/354063a0 (1991).

118 Last, P. R. & Stevens, J. D. *Sharks and rays of Australia*. (CSIRO Australia, 1994).

119 Sullivan, K. J. Age and growth of the elephant fish *Callorhinchus milii* (Elasmobranchii: Callorhynchidae). *N. Z. J. Mar. Freshwater Res.* **11**, 745-753, doi:10.1080/00288330.1977.9515710 (1978).

120 Gorman, T. B. S. Biological and economic aspects of the elephant fish, *Callorhynchus milii* Bory, in Pegasus Bay and the Canterbury Bight. 53 (1963).

121 Parsons, G. R. Metabolism and swimming efficiency of the bonnethead shark Sphyrna tiburo. *Mar Biol* **104**, 363-367, doi:10.1007/BF01314338 (1990).

122 Carrier, J. C., Musick, J. A. & Heithaus, M. R. *Biology of sharks and their relatives*. (CRC Press, 2004).

123 Richter, C., Park, J. W. & Ames, B. N. Normal oxidative damage to mitochondrial and nuclear DNA is extensive. *Proc Natl Acad Sci U S A* **85**, 6465-6467 (1988).

124 Wagner, J. R., Hu, C. C. & Ames, B. N. Endogenous oxidative damage of deoxycytidine in DNA. *Proc Natl Acad Sci U S A* **89**, 3380-3384, doi:10.1073/pnas.89.8.3380 (1992).

125 Coulombe-Huntington, J. & Majewski, J. Characterization of intron loss events in mammals. *Genome Res* **17**, 23-32, doi:10.1101/gr.5703406 (2007).

126 Loh, Y. H., Brenner, S. & Venkatesh, B. Investigation of loss and gain of introns in the compact genomes of pufferfishes (Fugu and Tetraodon). *Mol Biol Evol* **25**, 526-535, doi:10.1093/molbev/msm278 (2008).

127 Roy, S. W., Fedorov, A. & Gilbert, W. Large-scale comparison of intron positions in mammalian genes shows intron loss but no gain. *Proc Natl Acad Sci U S A* **100**, 7158-7162, doi:10.1073/pnas.1232297100 (2003).

128 Proost, S. *et al.* i-ADHoRe 3.0--fast and sensitive detection of genomic homology in extremely large data sets. *Nucleic acids research* **40**, e11, doi:10.1093/nar/gkr955 (2012).

129 Kohn, M. *et al.* Reconstruction of a 450-My-old ancestral vertebrate protokaryotype. *Trends in genetics : TIG* **22**, 203-210, doi:10.1016/j.tig.2006.02.008 (2006).

130 Nakatani, Y., Takeda, H., Kohara, Y. & Morishita, S. Reconstruction of the vertebrate ancestral genome reveals dynamic genome reorganization in early vertebrates. *Genome Res* **17**, 1254-1265, doi:10.1101/gr.6316407 (2007).

131 Christoffels, A. *et al.* Fugu genome analysis provides evidence for a whole-genome duplication early during the evolution of ray-finned fishes. *Mol Biol Evol* **21**, 1146-1151, doi:10.1093/molbev/msh114 (2004).

132 Ellegren, H. Evolutionary stasis: the stable chromosomes of birds. *Trends Ecol Evol* **25**, 283-291, doi:10.1016/j.tree.2009.12.004 (2010).

133 Burt, D. W. Origin and evolution of avian microchromosomes. *Cytogenet Genome Res* **96**, 97-112, doi:10.1159/000063018 (2002).

134 Spitz, F., Gonzalez, F. & Duboule, D. A global control region defines a chromosomal regulatory landscape containing the HoxD cluster. *Cell* **113**, 405-417, doi:10.1016/S0092-8674(03)00310-6 (2003).

135 Zakany, J., Kmita, M. & Duboule, D. A dual role for Hox genes in limb anterior-posterior asymmetry. *Science* **304**, 1669-1672, doi:10.1126/science.1096049 (2004).

136 Montavon, T. *et al.* A regulatory archipelago controls Hox genes transcription in digits. *Cell* **147**, 1132-1145, doi:10.1016/j.cell.2011.10.023 (2011).

137 Kikuta, H. *et al.* Genomic regulatory blocks encompass multiple neighboring genes and maintain conserved synteny in vertebrates. *Genome Res* **17**, 545-555, doi:10.1101/gr.6086307 (2007).

138 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:10.1101/gr.116491.110 (2011).

139 Ohno, S. *et al.* Microchromosomes in holocephalian, chondrostean and holostean fishes. *Chromosoma* **26**, 35-40 (1969).

140 Hanna, R. N. *et al.* Characterization and expression of the nuclear progestin receptor in zebrafish gonads and brain. *Biol Reprod* **82**, 112-122, doi:10.1095/biolreprod.109.078527 (2010).

141 Tsang, P. & Callard, I. P. Luteal progesterone production and regulation in the viviparous dogfish, *Squalus acanthias*. *J Exp Zool* **241**, 377-382, doi:10.1002/jez.1402410313 (1987).

142 Pinter, J. & Thomas, P. The ovarian progestogen receptor in the spotted seatrout, Cynoscion nebulosus, demonstrates steroid specificity different from progesterone receptors in other vertebrates. *J Steroid Biochem Mol Biol* **60**, 113-119 (1997).

143 Alegre-Cebollada, J., Onaderra, M., Gavilanes, J. G. & del Pozo, A. M. Sea anemone actinoporins: the transition from a folded soluble state to a functionally active membrane-bound oligomeric pore. *Curr Protein Pept Sci* **8**, 558-572, doi:10.2174/138920307783018686 (2007).

144 van der Aa, L. M. *et al.* A large new subset of TRIM genes highly diversified by duplication and positive selection in teleost fish. *BMC Biol* **7**, 7, doi:10.1186/1741-7007-7-7 (2009).

145 Boudinot, P. *et al.* Origin and evolution of TRIM proteins: new insights from the complete TRIM repertoire of zebrafish and pufferfish. *PLoS One* **6**, e22022, doi:10.1371/journal.pone.0022022 (2011).

146 Zhang, J. *et al.* Loss of fish actinotrichia proteins and the fin-to-limb transition. *Nature* **466**, 234-237, doi:10.1038/nature09137 (2010).

147 Nechiporuk, A. & Raible, D. W. FGF-dependent mechanosensory organ patterning in zebrafish. *Science* **320**, 1774-1777, doi:10.1126/science.1156547 (2008).

148 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).

149 Jayatilake, G. S., Huddleston, J. A. & Abraham, E. P. Conversion of isopenicillin N into penicillin N in cell-free extracts of Cephalosporium acremonium. *Biochem J* **194**, 645-647 (1981).

150 Ueda, A., Nagai, H., Ishida, M., Nagashima, Y. & Shiomi, K. Purification and molecular cloning of SE-cephalotoxin, a novel proteinaceous toxin from the posterior salivary gland of cuttlefish Sepia esculenta. *Toxicon* **52**, 574-581, doi:10.1016/j.toxicon.2008.07.007 (2008).

151 Abrantes, K. G. & Barnett, A. Intrapopulation variations in diet and habitat use in a marine apex predator, the broadnose sevengill shark Notorynchus cepedianus. *Mar Ecol Prog Ser* **442**, 133-148, doi:10.3354/meps09395 (2011).

152 Niimura, Y. & Nei, M. Extensive gains and losses of olfactory receptor genes in mammalian evolution. *PLoS One* **2**, e708, doi:10.1371/journal.pone.0000708 (2007).

153 Schultz, J., Milpetz, F., Bork, P. & Ponting, C. P. SMART, a simple modular architecture research tool: identification of signaling domains. *Proc Natl Acad Sci U S A* **95**, 5857-5864 (1998).

154 Katoh, K., Misawa, K., Kuma, K. & Miyata, T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res* **30**, 3059-3066 (2002).

155 Lisney, T. J. A review of the sensory biology of chimaeroid fishes. *Rev Fish Biol Fisheries* **20**, 571-590, doi:10.1007/s11160-010-9162-x (2010).

156 Warren, W. C. *et al.* Genome analysis of the platypus reveals unique signatures of evolution. *Nature* **453**, 175-183, doi:10.1038/nature06936 (2008).

157 Grus, W. E. & Zhang, J. Origin and evolution of the vertebrate vomeronasal system viewed through system-specific genes. *Bioessays* **28**, 709-718, doi:10.1002/bies.20432 (2006).

158 Naito, T. *et al.* Putative pheromone receptors related to the Ca2+-sensing receptor in Fugu. *Proc Natl Acad Sci U S A* **95**, 5178-5181 (1998).

159 Pfister, P. & Rodriguez, I. Olfactory expression of a single and highly variable V1r pheromone receptor-like gene in fish species. *Proc Natl Acad Sci U S A* **102**, 5489-5494, doi:10.1073/pnas.0402581102 (2005).

160 Grus, W. E. & Zhang, J. Origin of the genetic components of the vomeronasal system in the common ancestor of all extant vertebrates. *Mol Biol Evol* **26**, 407-419, doi:10.1093/molbev/msn262 (2009).

161 Alioto, T. S. & Ngai, J. The repertoire of olfactory C family G protein-coupled receptors in zebrafish: candidate chemosensory receptors for amino acids. *BMC Genomics* **7**, 309, doi:10.1186/1471-2164-7-309 (2006).

162 Saraiva, L. R. & Korsching, S. I. A novel olfactory receptor gene family in teleost fish. *Genome Res* **17**, 1448-1457, doi:10.1101/gr.6553207 (2007).

163 Shi, P. & Zhang, J. Comparative genomic analysis identifies an evolutionary shift of vomeronasal receptor gene repertoires in the vertebrate transition from water to land. *Genome Res* **17**, 166-174, doi:10.1101/gr.6040007 (2007).

164 Zhang, J. & Webb, D. M. Evolutionary deterioration of the vomeronasal pheromone transduction pathway in catarrhine primates. *Proc Natl Acad Sci U S A* **100**, 8337-8341, doi:10.1073/pnas.1331721100 (2003).

165 Johnstone, K. A. *et al.* Genomic organization and evolution of the vomeronasal type 2 receptor-like (OlfC) gene clusters in Atlantic salmon, *Salmo salar*. *Mol Biol Evol* **26**, 1117-1125, doi:10.1093/molbev/msp027 (2009).

166 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:10.1093/nar/25.17.3389 (1997).

167 Vortkamp, A. *et al.* Regulation of rate of cartilage differentiation by Indian hedgehog and PTH-related protein. *Science* **273**, 613-622, doi:10.1126/science.273.5275.613 (1996).

168 St-Jacques, B., Hammerschmidt, M. & McMahon, A. P. Indian hedgehog signaling regulates proliferation and differentiation of chondrocytes and is essential for bone formation. *Genes Dev* **13**, 2072-2086 (1999).

169 Ingham, P. W., Nakano, Y. & Seger, C. Mechanisms and functions of Hedgehog signalling across the metazoa. *Nat Rev Genet* **12**, 393-406, doi:10.1038/nrg2984 (2011).

170 Karsenty, G., Kronenberg, H. M. & Settembre, C. Genetic control of bone formation. *Annu Rev Cell Dev Biol* **25**, 629-648, doi:10.1146/annurev.cellbio.042308.113308 (2009).

171 Liu, Y. *et al.* Parathyroid hormone gene family in a cartilaginous fish, the elephant shark (Callorhinchus milii). *J Bone Miner Res* **25**, 2613-2623, doi:10.1002/jbmr.178 (2010).

172 Kawasaki, K., Buchanan, A. V. & Weiss, K. M. Gene duplication and the evolution of vertebrate skeletal mineralization. *Cells Tissues Organs* **186**, 7-24, doi:10.1159/000102678 (2007).

173 Kawasaki, K. The SCPP gene family and the complexity of hard tissues in vertebrates. *Cells Tissues Organs* **194**, 108-112, doi:10.1159/000324225 (2011).

174 Kawasaki, K. The SCPP gene repertoire in bony vertebrates and graded differences in mineralized tissues. *Dev Genes Evol* **219**, 147-157, doi:10.1007/s00427-009-0276-x (2009).

175 King, B. L., Gillis, J. A., Carlisle, H. R. & Dahn, R. D. A natural deletion of the *HoxC* cluster in elasmobranch fishes. *Science* **334**, 1517, doi:10.1126/science.1210912 (2011).

176 Staines, K. A., MacRae, V. E. & Farquharson, C. The importance of the SIBLING family of proteins on skeletal mineralisation and bone remodelling. *J Endocrinol* **214**, 241-255, doi:10.1530/JOE-12-0143 (2012).

177 Kawasaki, K., Buchanan, A. V. & Weiss, K. M. Biomineralization in humans: making the hard choices in life. *Annu Rev Genet* **43**, 119-142, doi:10.1146/annurev-genet-102108-134242 (2009).

178 Styrkarsdottir, U. *et al.* New sequence variants associated with bone mineral density. *Nat Genet* **41**, 15-17, doi:10.1038/ng.284 (2009).

179 Duncan, E. L. *et al.* Genome-wide association study using extreme truncate selection identifies novel genes affecting bone mineral density and fracture risk. *PLoS Genet* **7**, e1001372, doi:10.1371/journal.pgen.1001372 (2011).

180 Koller, D. L. *et al.* Genome-wide association study of bone mineral density in premenopausal European-American women and replication in African-American women. *J Clin Endocrinol Metab* **95**, 1802-1809, doi:10.1210/jc.2009-1903 (2010).

181 Hwang, W. Y. *et al.* Efficient genome editing in zebrafish using a CRISPR-Cas system. *Nat Biotechnol* **31**, 227-229, doi:10.1038/nbt.2501 (2013).

182 DeLaurier, A. *et al.* Zebrafish sp7:EGFP: a transgenic for studying otic vesicle formation, skeletogenesis, and bone regeneration. *Genesis* **48**, 505-511, doi:10.1002/dvg.20639 (2010).

183 Laue, K., Jänicke, M., Plaster, N., Sonntag, C. & Hammerschmidt, M. Restriction of retinoic acid activity by Cyp26b1 is required for proper timing and patterning of osteogenesis during zebrafish development. *Development*, doi:10.1242/dev.021238 (2008).

184 Boskey, A. L., Spevak, L., Paschalis, E., Doty, S. B. & McKee, M. D. Osteopontin deficiency increases mineral content and mineral crystallinity in mouse bone. *Calcif Tissue Int* **71**, 145-154, doi:10.1007/s00223-001-1121-z (2002).

185 Malaval, L. *et al.* Bone sialoprotein plays a functional role in bone formation and osteoclastogenesis. *J Exp Med* **205**, 1145-1153, doi:10.1084/jem.20071294 (2008).

186 Hinds, K. R. & Litman, G. W. Major reorganization of immunoglobulin V_H segmental elements during vertebrate evolution. *Nature* **320**, 546-549, doi:10.1038/320546a0 (1986).

187 Lee, V. *et al.* The evolution of multiple isotypic IgM heavy chain genes in the shark. *J. Immunol.* **180**, 7461-7470 (2008).

188 Du Pasquier, L. Fish 'n' TRIMs. *J Biol* **8**, 50, doi:10.1186/jbiol150 (2009).

189 Ohta, Y. *et al.* Primordial linkage of *β2-Microglobulin* to the MHC. *J. Immunol.* **186**, 3563-3571, doi:10.4049/jimmunol.1003933 (2011).

190 Rast, J. P., Amemiya, C. T., Litman, R. T., Strong, S. J. & Litman, G. W. Distinct patterns of *IgH* structure and organization in a divergent lineage of chrondrichthyan fishes. *Immunogenetics* **47**, 234-245 (1998).

191 Dooley, H. & Flajnik, M. F. Antibody repertoire development in cartilaginous fish. *Dev Comp Immunol* **30**, 43-56, doi:10.1016/j.dci.2005.06.022 (2006).

192 Criscitiello, M. F., Saltis, M. & Flajnik, M. F. An evolutionarily mobile antigen receptor variable region gene: Doubly rearranging NAR-TcR genes in sharks. *Proc. Natl. Acad. Sci. USA* **103**, 5036-5041, doi:10.1073/pnas.0507074103 (2006).

193 Parra, Z. E., Ohta, Y., Criscitiello, M. F., Flajnik, M. F. & Miller, R. D. The dynamic TCRδ: TCRδ chains in the amphibian *Xenopus tropicalis* utilize antibody-like V genes. *Eur. J. Immunol.* **40**, 2319-2329, doi:10.1002/eji.201040515 (2010).

194 Parra, Z. E., Mitchell, K., Dalloul, R. A. & Miller, R. D. A second TCRδ locus in Galliformes uses antibody-like V domains: insight into the evolution of TCRδ and TCRμ genes in tetrapods. *J. Immunol.* **188**, 3912-3919, doi:10.4049/jimmunol.1103521 (2012).

195 Parra, Z. E., Lillie, M. & Miller, R. D. A model for the evolution of the mammalian t-cell receptor αδ and μ loci based on evidence from the duckbill platypus. *Mol. Biol. Evol.* **29**, 3205-3214, doi:10.1093/molbev/mss128 (2012).

196 Criscitiello, M. F. & Flajnik, M. F. Four primordial immunoglobulin light chain isotypes, including λ and κ, identified in the most primitive living jawed vertebrates. *Eur. J. Immunol.* **37**, 2683-2694, doi:10.1002/eji.200737263 (2007).

197 Anderson, M. K., Shamblott, M. J., Litman, R. T. & Litman, G. W. Generation of immunoglobulin light chain gene diversity in *Raja erinacea* is not associated with somatic rearrangement, an exception to a central paradigm of B cell immunity. *J. Exp. Med.* **182**, 109-119, doi:10.1084/jem.182.1.109 (1995).

198 Chen, H. *et al.* Characterization of arrangement and expression of the T cell receptor gamma locus in the sandbar shark. *Proc Natl Acad Sci U S A* **106**, 8591-8596, doi:10.1073/pnas.0811283106 (2009).

199 Flajnik, M. F. & Kasahara, M. Origin and evolution of the adaptive immune system: genetic events and selective pressures. *Nat. Rev. Genet.* **11**, 47-59, doi:10.1038/nrg2703 (2010).

200 Bartl, S., Baish, M. A., Flajnik, M. F. & Ohta, Y. Identification of class I genes in cartilaginous fish, the most ancient group of vertebrates displaying an adaptive immune response. *J Immunol* **159**, 6097-6104 (1997).

201 Ohta, Y. *et al.* Primitive synteny of vertebrate major histocompatibility complex class I and class II genes. *Proc Natl Acad Sci U S A* **97**, 4712-4717, doi:10.1073/pnas.97.9.4712 (2000).

202 Ohta, Y., McKinney, E. C., Criscitiello, M. F. & Flajnik, M. F. Proteasome, transporter associated with antigen processing, and class I genes in the nurse shark *Ginglymostoma cirratum*: evidence for a stable class I region and MHC haplotype lineages. *J. Immunol.* **168**, 771-781 (2002).

203 Flajnik, M. F., Tlapakova, T., Criscitiello, M. F., Krylov, V. & Ohta, Y. Evolution of the B7 family: co-evolution of B7H6 and NKp30, identification of a new B7 family member, B7H7, and of B7's historical relationship with the MHC. *Immunogenetics* **64**, 571-590, doi:10.1007/s00251-012-0616-2 (2012).

204 Nonaka, M. & Kimura, A. Genomic view of the evolution of the complement system. *Immunogenetics* **58**, 701-713, doi:10.1007/s00251-006-0142-1 (2006).

205 Shin, D.-H., Webb, B., Nakao, M. & Smith, S. L. Molecular cloning, structural analysis and expression of complement component Bf/C2 genes in the nurse shark, *Ginglymostoma cirratum*. *Dev Comp Immunol* **31**, 1168-1182, doi:10.1016/j.dci.2007.03.001 (2007).

206 Nomiyama, H., Osada, N. & Yoshie, O. Systematic classification of vertebrate chemokines based on conserved synteny and evolutionary history. *Genes Cells* **18**, 1-16, doi:10.1111/gtc.12013 (2013).

207 Robertsen, B. The interferon system of teleost fish. *Fish Shellfish Immunol* **20**, 172-191, doi:10.1016/j.fsi.2005.01.010 (2006).

208 Wang, T., Huang, W., Costa, M. M. & Secombes, C. J. The gamma-chain cytokine/receptor system in fish: More ligands and receptors. *Fish Shellfish Immunol* **31**, 673-687, doi:10.1016/j.fsi.2011.05.016 (2011).

209 Cherrier, M. & Eberl, G. The development of LTi cells. *Curr Opin Immunol* **24**, 178-183, doi:10.1016/j.coi.2012.02.003 (2012).

210 Andersen, K. G., Nissen, J. K. & Betz, A. G. Comparative genomics reveals key gain-of-function events in Foxp3 during regulatory T cell evolution. *Front Immunol* **3**, 113, doi:10.3389/fimmu.2012.00113 (2012).

211 Taniuchi, I. & Ellmeier, W. Transcriptional and epigenetic regulation of CD4/CD8 lineage choice. *Adv Immunol* **110**, 71-110, doi:10.1016/B978-0-12-387663-8.00003-X (2011).

212 Shaw, A. S. *et al.* Short related sequences in the cytoplasmic domains of CD4 and CD8 mediate binding to the amino-terminal domain of the $p56^{lck}$ tyrosine protein kinase. *Mol Cell Biol* **10**, 1853-1862, doi:10.1128/MCB.10.5.1853 (1990).

213 Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12**, 323, doi:10.1186/1471-2105-12-323 (2011).

214 Spits, H. *et al.* Innate lymphoid cells - a proposal for uniform nomenclature. *Nat Rev Immunol* **13**, 145-149, doi:10.1038/nri3365 (2013).

215 Criscitiello, M. F., Ohta, Y., Saltis, M., McKinney, E. C. & Flajnik, M. F. Evolutionarily conserved TCR binding sites, identification of T cells in primary lymphoid tissues, and surprising trans-rearrangements in nurse shark. *J. Immunol.* **184**, 6950-6960, doi:10.4049/jimmunol.0902774 (2010).

216 Nehls, M. *et al.* Two genetically separable steps in the differentiation of thymic epithelium. *Science* **272**, 886-889 (1996).

217 Anderson, M. S. *et al.* Projection of an immunological self shadow within the thymus by the aire protein. *Science* **298**, 1395-1401, doi:10.1126/science.1075958 (2002).

218 Roberts, C. W. M., Shutter, J. R. & Korsmeyer, S. J. *Hox11* controls the genesis of the spleen. *Nature* **368**, 747-749, doi:10.1038/368747a0 (1994).

219 Spolski, R. & Leonard, W. J. IL-21 and T follicular helper cells. *Int Immunol* **22**, 7-12, doi:10.1093/intimm/dxp112 (2010).

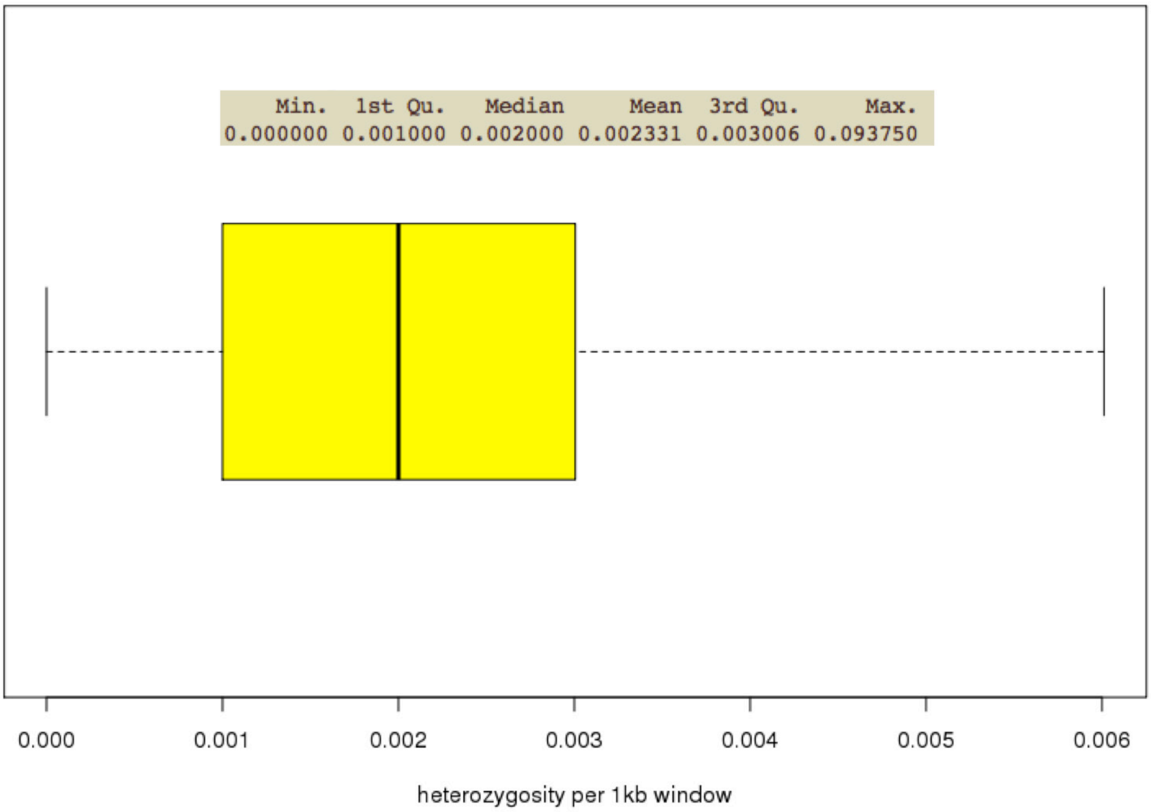
220 Chien, Y.-H., Iwashima, M., Kaplan, K. B., Elliott, J. F. & Davis, M. M. A new T-cell receptor gene located within the alpha locus and expressed early in T-cell differentiation. *Nature* **327**, 677-682, doi:10.1038/327677a0 (1987).

221 Locksley, R. M., Killeen, N. & Lenardo, M. J. The TNF and TNF receptor superfamilies: integrating mammalian biology. *Cell* **104**, 487-501, doi:10.1016/S0092-8674(01)00237-9 (2001).

222 Dooley, H. & Flajnik, M. F. Shark immunity bites back: affinity maturation and memory response in the nurse shark, *Ginglymostoma cirratum*. *Eur. J. Immunol.* **35**, 936-945, doi:10.1002/eji.200425760 (2005).

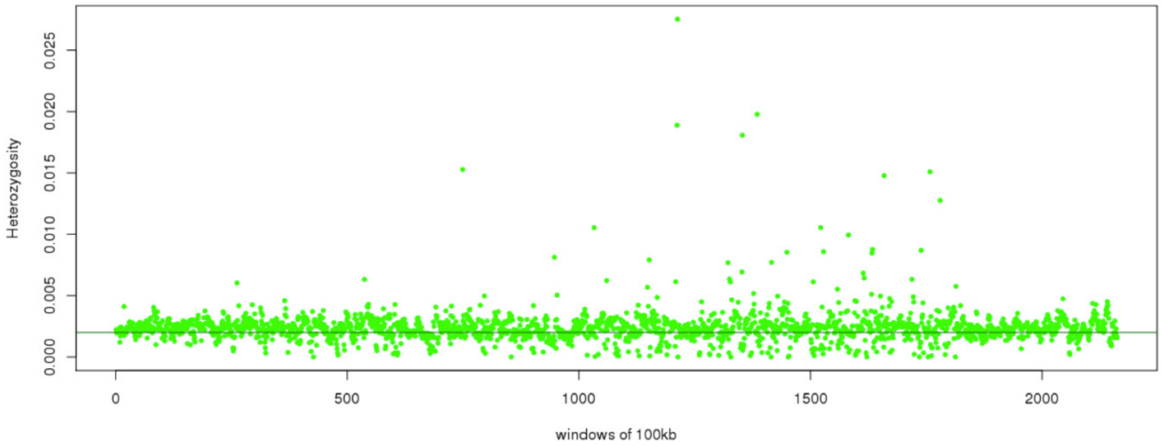
223 Fontenot, J. D., Rasmussen, J. P., Gavin, M. A. & Rudensky, A. Y. A function for interleukin 2 in Foxp3-expressing regulatory T cells. *Nat Immunol* **6**, 1142-1151, doi:10.1038/ni1263 (2005).

224 Ivanov, II, Zhou, L. & Littman, D. R. Transcriptional regulation of Th17 cell differentiation. *Semin. Immunol.* **19**, 409-417, doi:10.1016/j.smim.2007.10.011 (2007).

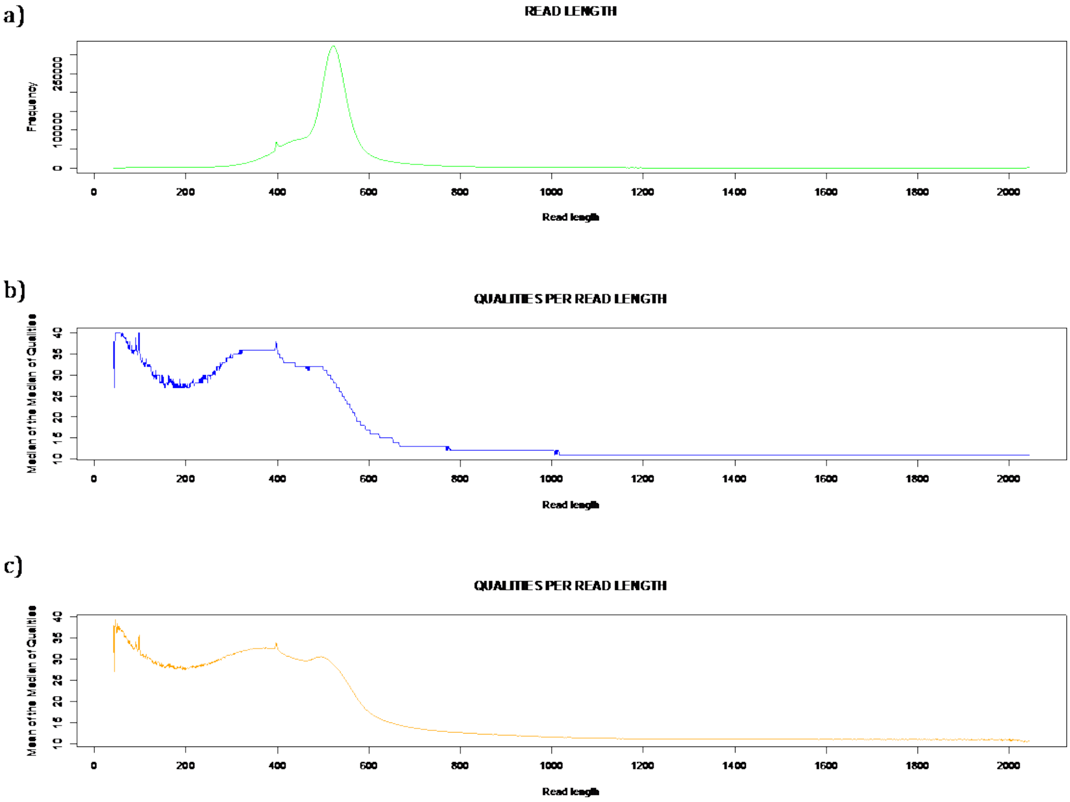


```
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
0.000000 0.001000 0.002000 0.002331 0.003006 0.093750
```

Supplementary Figure I.1 | Boxplot of the heterozygosity in 1-kb windows of callable *C. milii* genome.

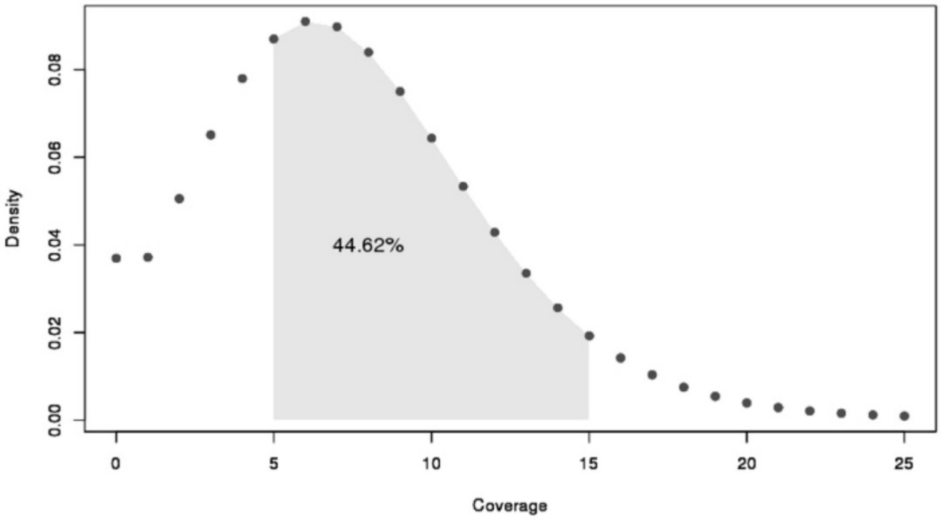


Supplementary Figure I.2 | Mean heterozygosity in 100 windows of 1-kb of callable *C. milii* genome.
The median for the windows is plotted as a horizontal line at 0.002.



Supplementary Figure I.3 | Quality assessment of the Roche 454 reads:

a) Read length distribution, b) median read quality per read length, c) mean read quality per read length.



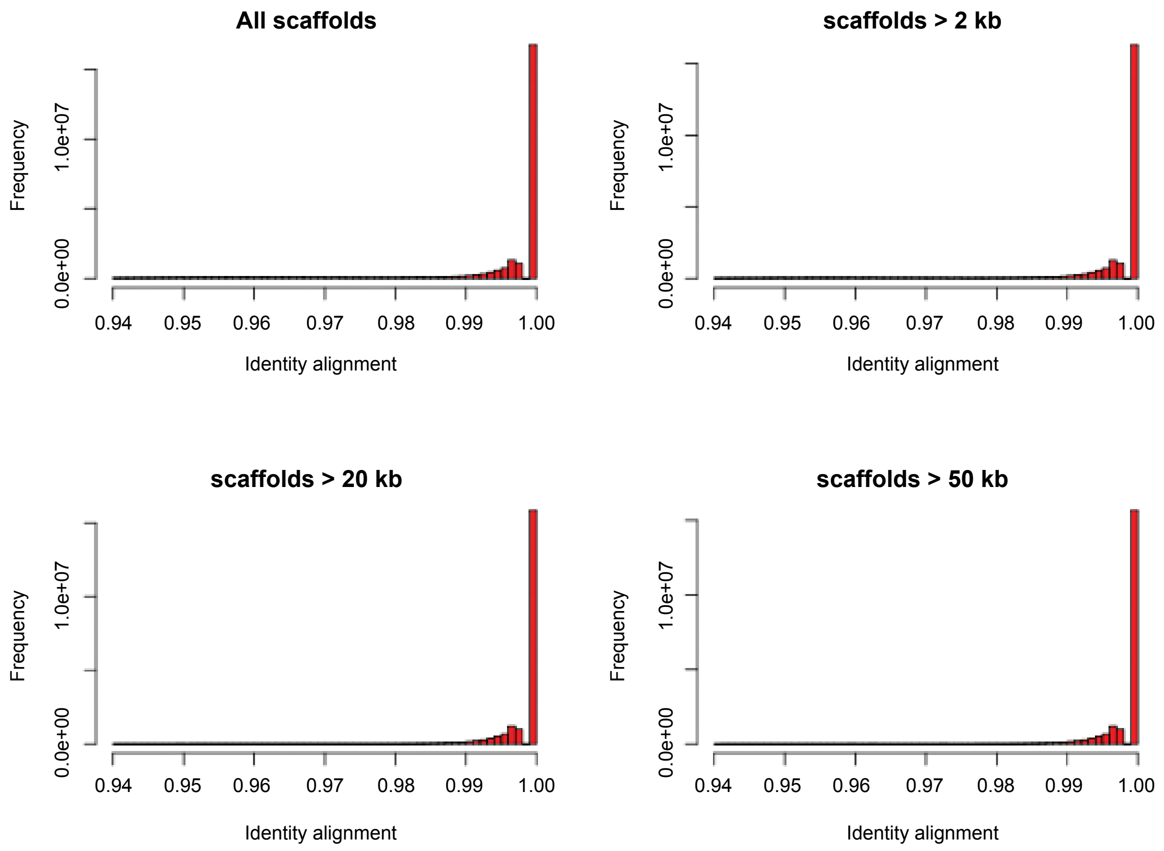
Supplementary Figure I.4 | Coverage distribution per non-repetitive base pair of the *C. milii* genome.

The figure inset represents the proportion of the genome with 5-15× coverage relative to the total length of the assembly.

a

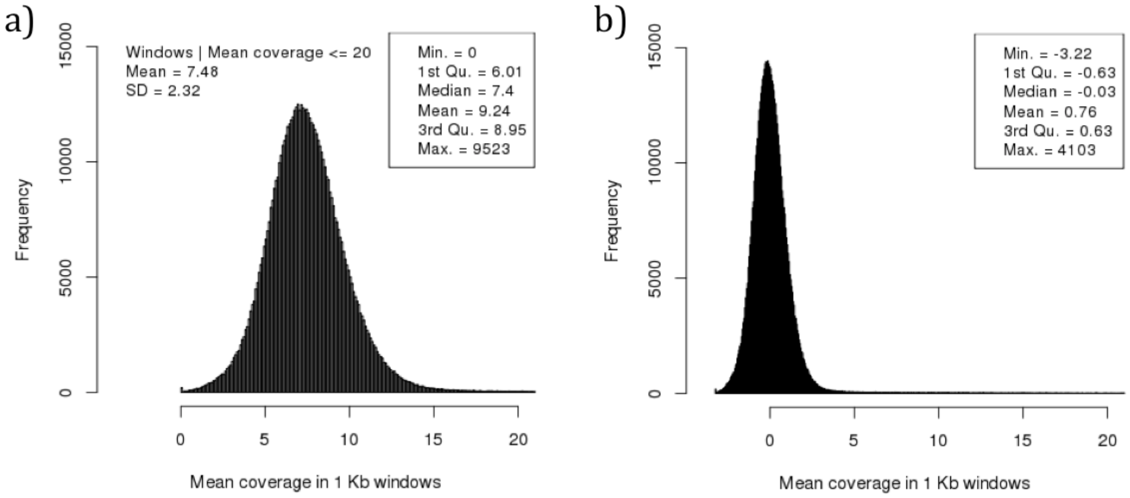
	all scaffolds	scaffolds > 2 kb	scaffolds > 20 kb	scaffolds > 50 kb
Mean	0.9938	0.9947	0.9973	0.9975
Median	1	1	1	1

b



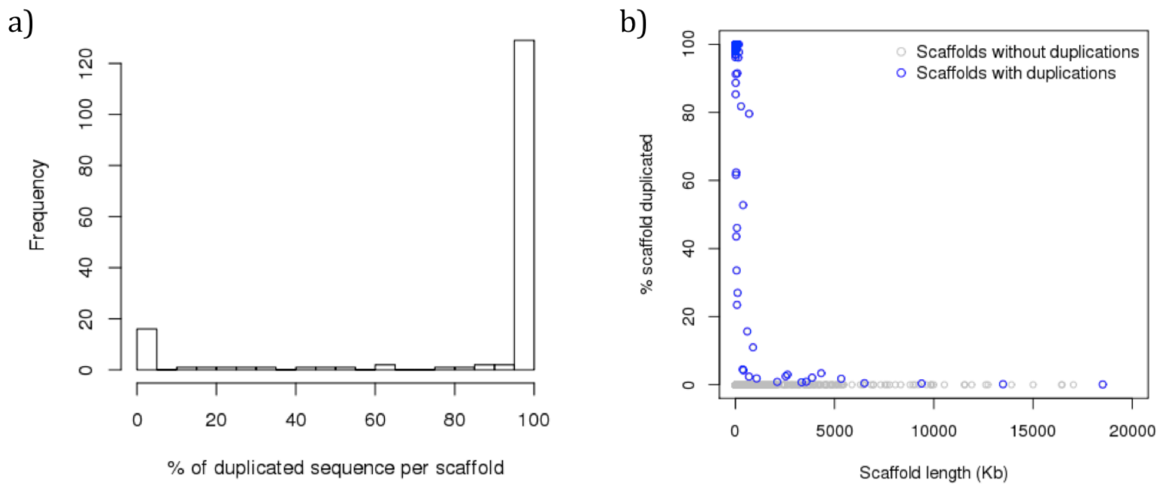
Supplementary Figure I.5 | Sequence identity of reads to reference genome according to the scaffold length.

a) Mean and median; b) distribution.

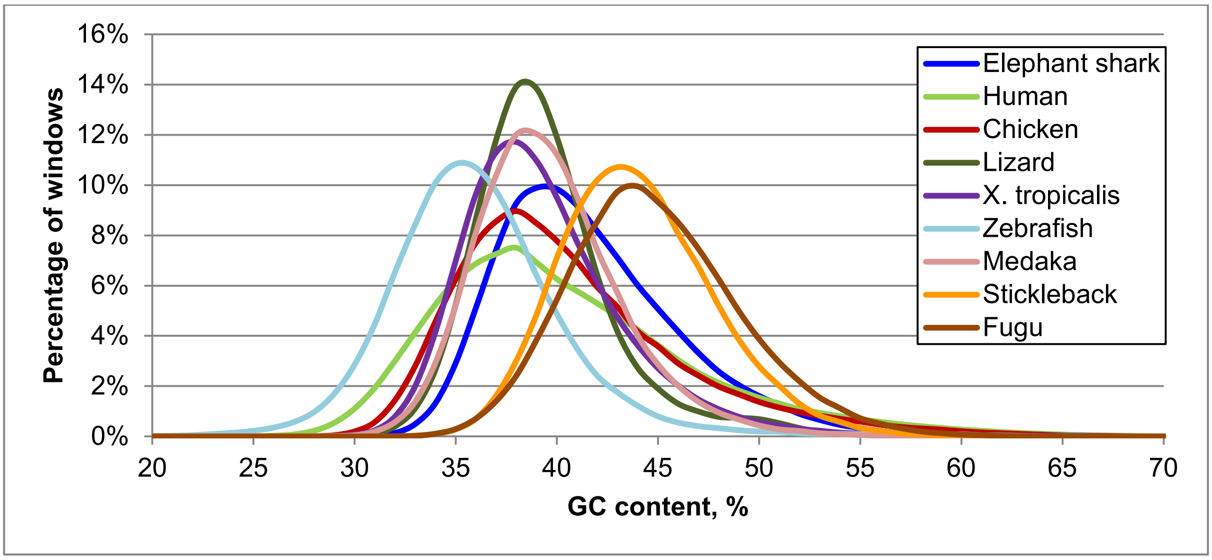


Supplementary Figure I.6 | Mean coverage of the *C. milii* genome before and after normalization.

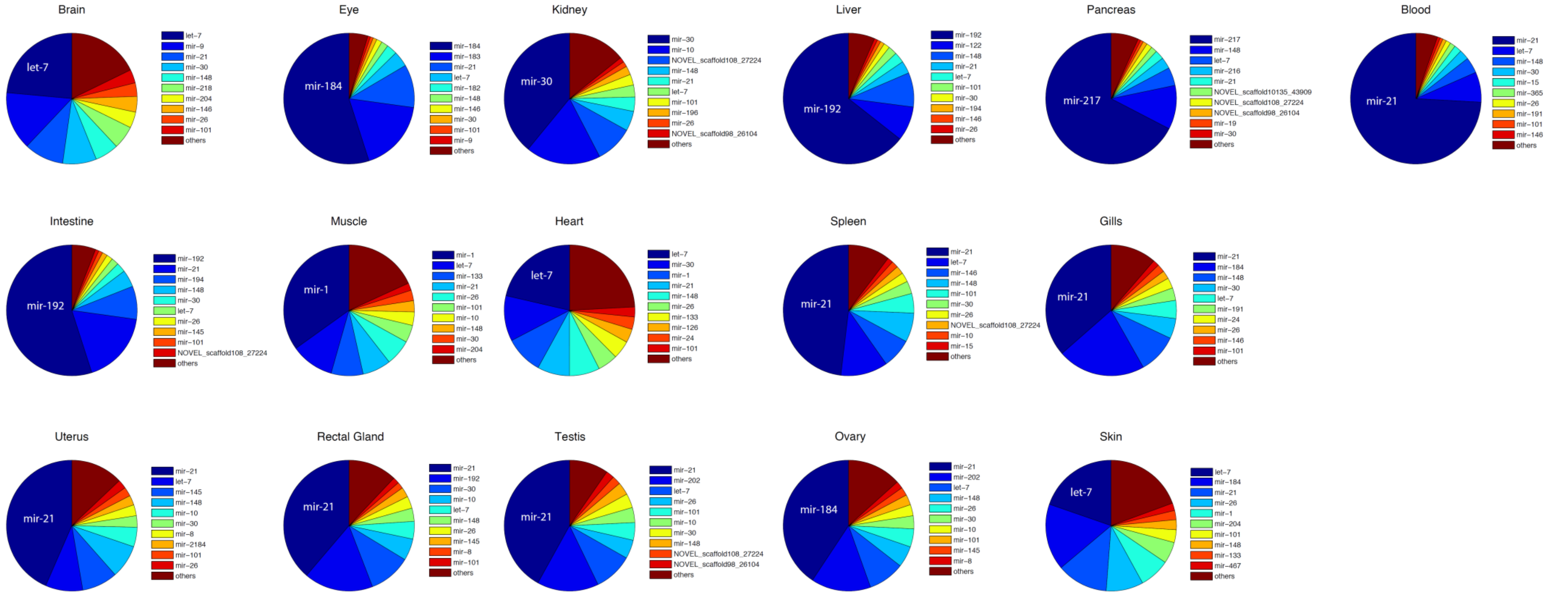
a) Distribution of the window mean coverage before normalization. Inset, the summary statistics of the entire distribution (boxed) and of only those windows with mean coverage lower than 20×. b) Distribution of window mean coverage after normalization and its summary statistics.



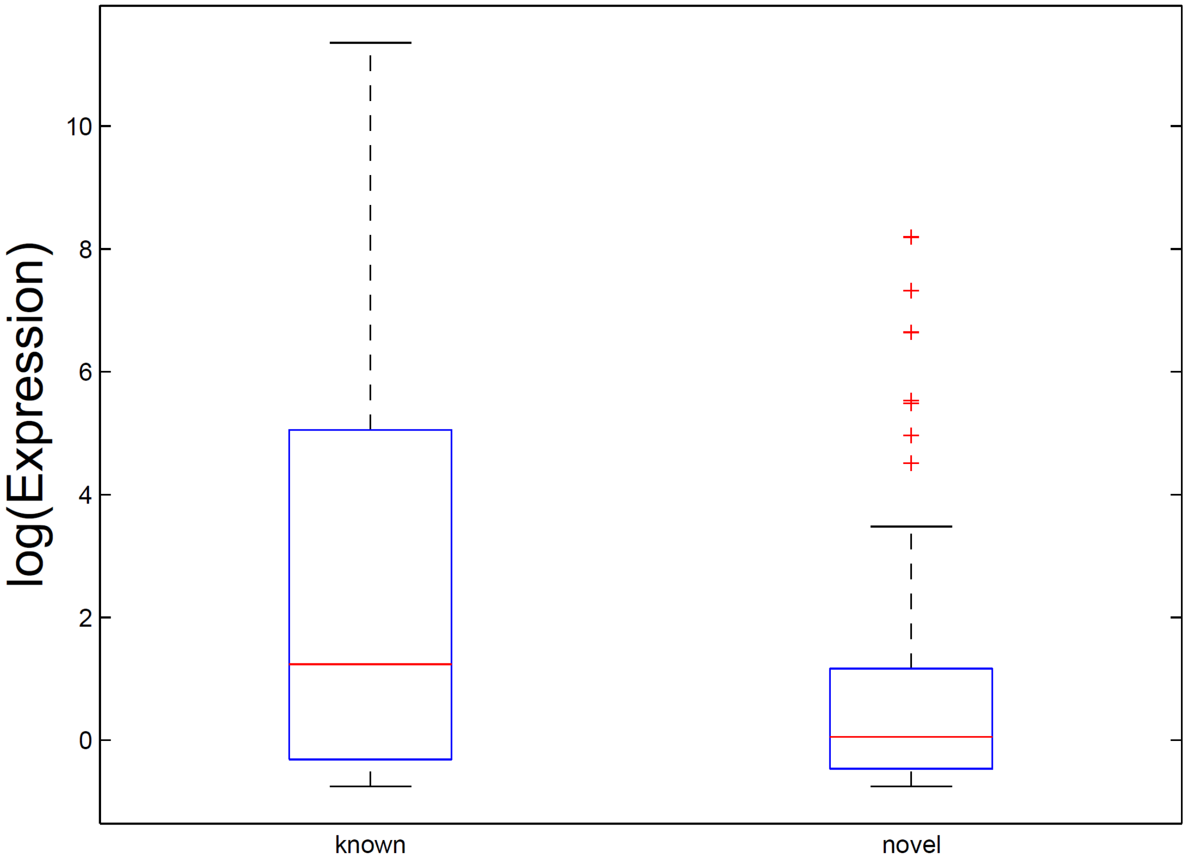
Supplementary Figure I.7 | Distribution of duplicated sequences across scaffolds. a) Distribution of the percentage of duplicated sequence per scaffold; b) scaffold length plotted against the percentage of scaffold sequence that is estimated as duplicated.



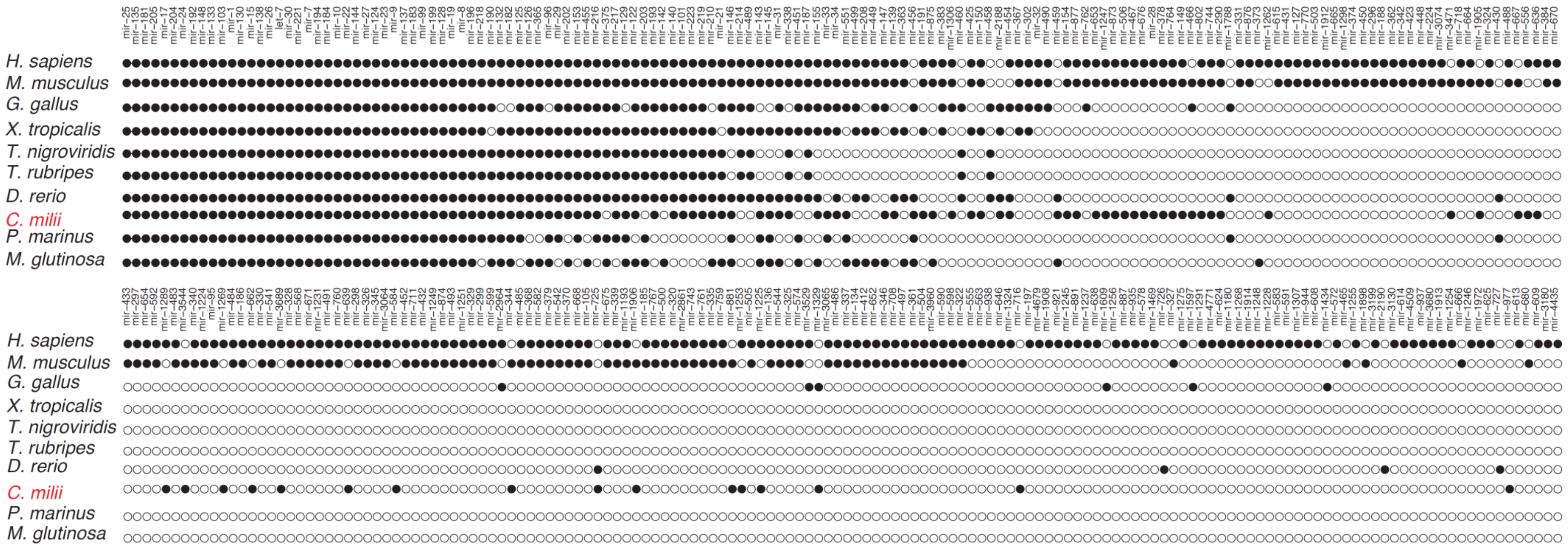
Supplementary Figure II.1 | Distribution of GC content in vertebrate genomes. The GC values of non-overlapping 3-kb windows were used. The windows were distributed into bins of 1% GC content.



Supplementary Figure III.1 | miRNA expression profile of the top 10 most highly expressed miRNAs in 16 tissues of *C. milii*.

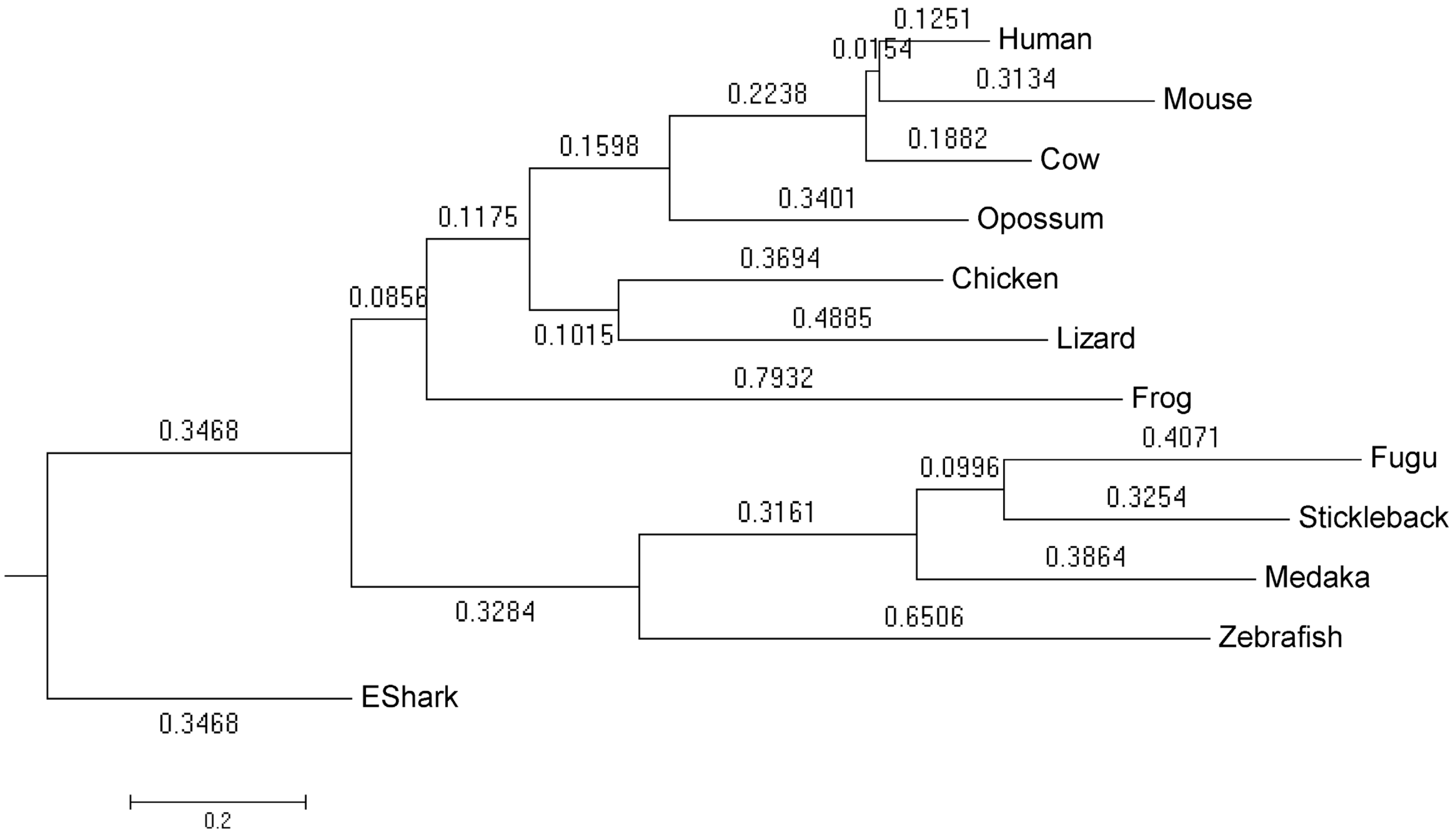


Supplementary Figure III.2 | Boxplot of expression levels for known and novel *C. milii* miRNAs identified in this study.



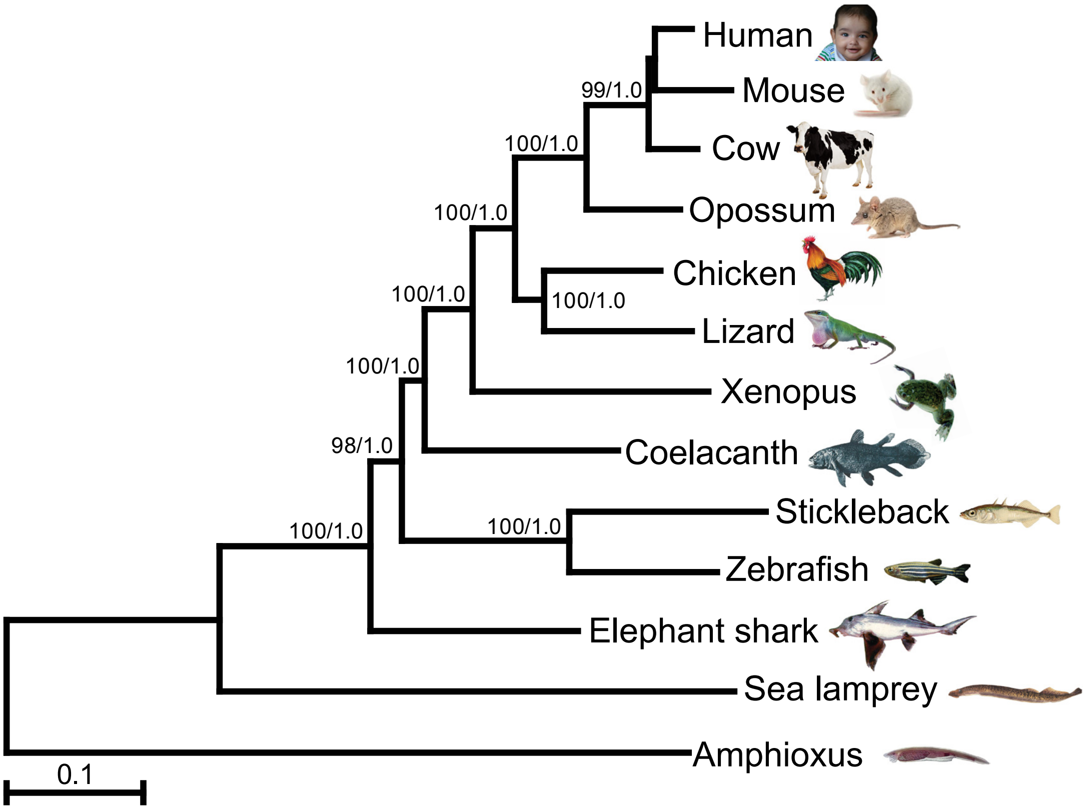
Supplementary Figure III.3 | Distribution of miRNA families across vertebrate lineages.

The presence of a miRNA family is indicated by a filled circle. The data were collated from miRBase release 19 (*Homo sapiens*, *Mus musculus*, *Gallus gallus*, *Xenopus tropicalis*, *Tetraodon nigroviridis*, *Takifugu rubripes*, *Danio rerio*), this study (*Callorhinchus milii*) and Heimberg et al. (2010) (sea lamprey - *Petromyzon marinus* and Atlantic hagfish - *Myxine glutionsa*).



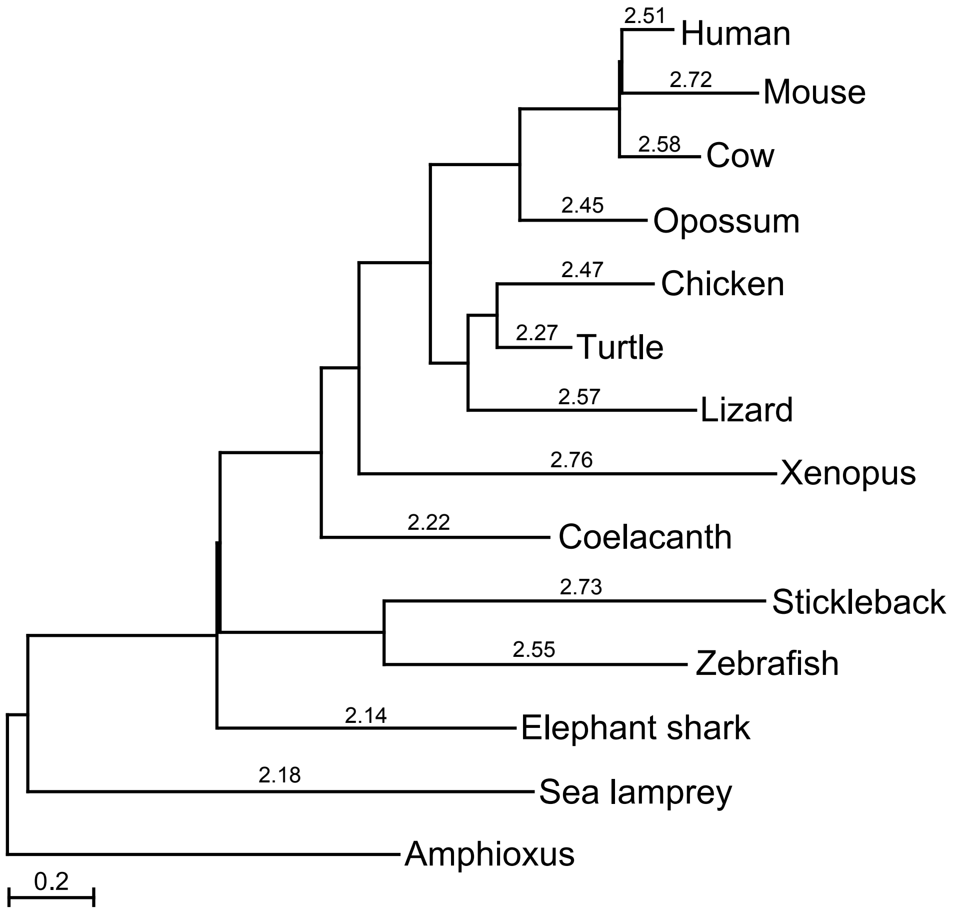
Supplementary Figure IV.1 | A neutral tree of 12 vertebrate genomes.

The tree was obtained from a multiple alignment of four-fold degenerate sites.

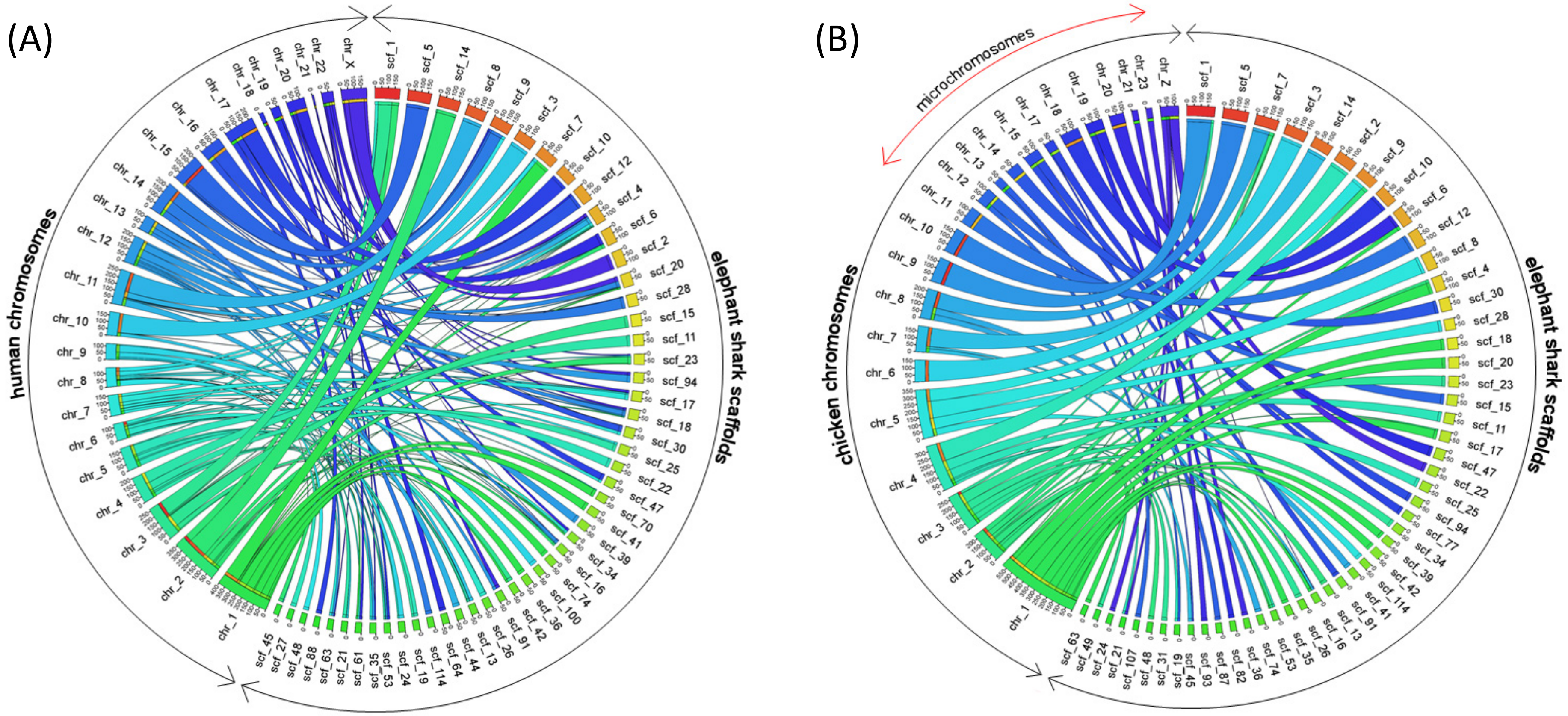


Supplementary Figure V.1 | The phylogenetic position of *C. milii*.

The tree is based on Maximum Likelihood (ML) and Bayesian Inference (BI) using a concatenated alignment of 699 one-to-one core orthologs (237,907 amino acid positions) from 13 chordates. ML bootstrap percentage (left) and BI posterior probability (right) are indicated at the nodes. Constrained nodes have no ML or BI value. Sea lamprey is the closest outgroup to gnathostomes. The tree has been rooted by specifying amphioxus as the outgroup. The scale bar represents 0.1 substitutions per site.

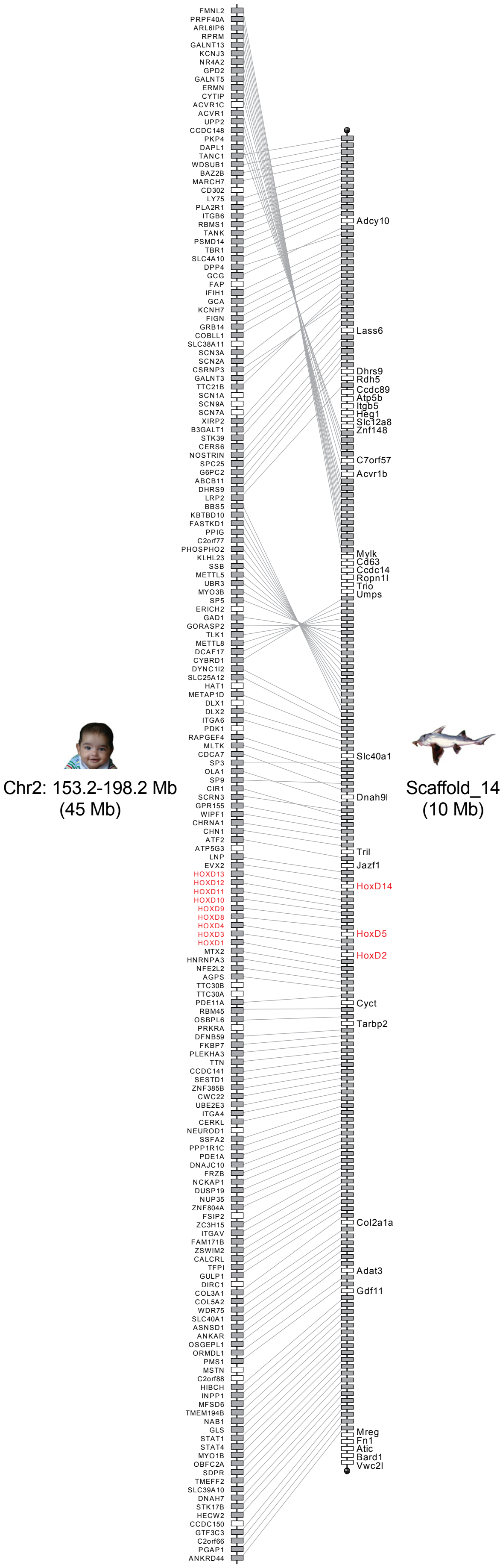


Supplementary Figure VI.1 | *C. milii* genome is evolving slower than turtle genome. A neutral tree of 14 chordates (including western painted turtle) based on four-fold degenerate (4D) sites suggests that *C. milii* possesses the slowest evolving vertebrate genome. 4D sites were extracted from the concatenated coding sequence alignment of 399 strict one-to-one core orthologs. Pairwise distances to amphioxus are shown for each species above their respective branches.



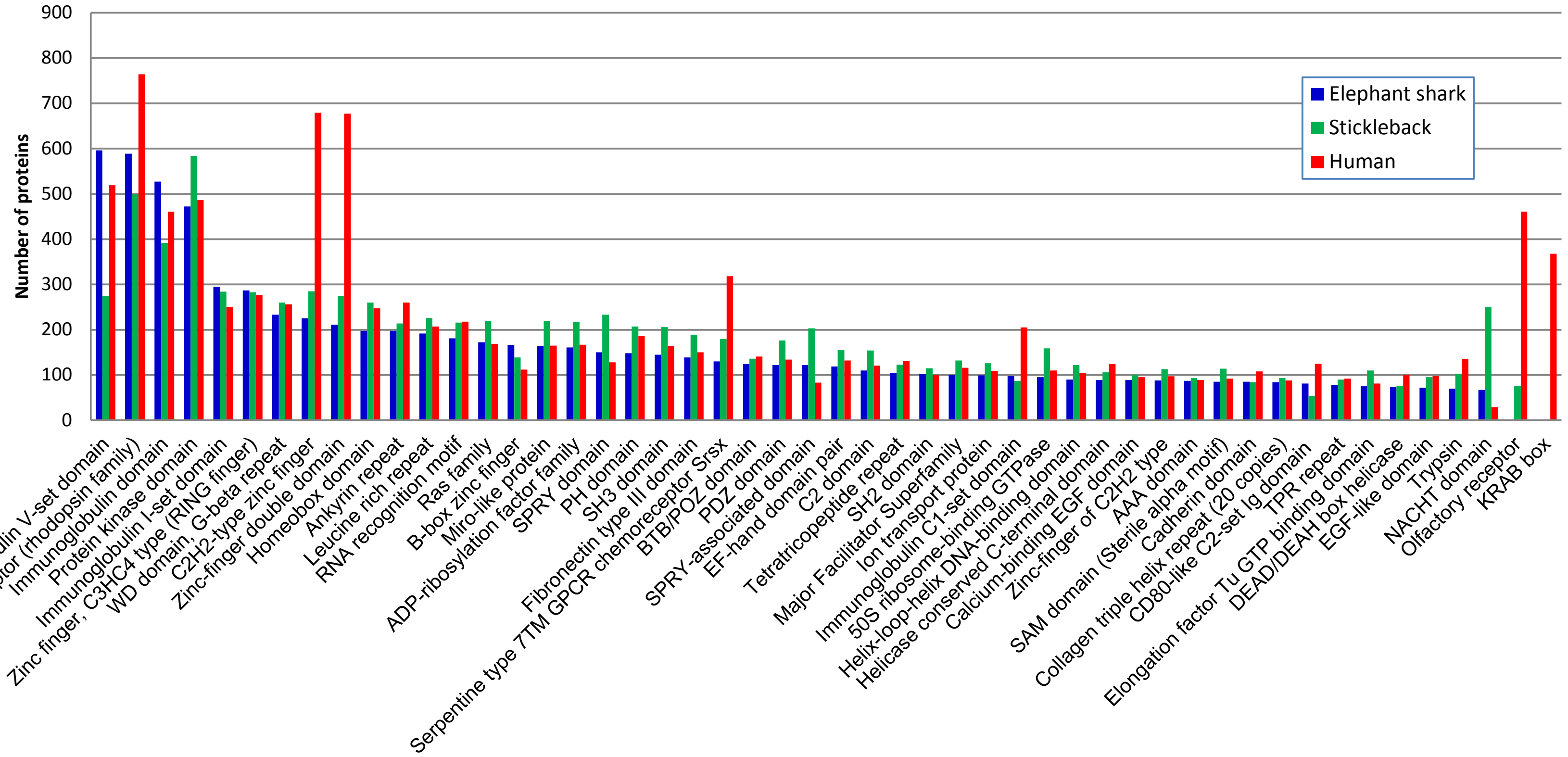
Supplementary Figure VIII.1 | Circos plot of the top 50 *C. milii* scaffolds syntenic to human and chicken chromosomes.

The number of syntenic genes between the top 50 *C. milii* scaffolds, and human (A) or chicken (B) chromosomes are illustrated using Circos v0.56 (Krzywinski et al. 2009). The width of ribbons linking the chromosomes of human and chicken to the *C. milii* scaffolds is proportional to the number of shared syntenic genes.


FMNL2
PRPF40A
ARL6IP6
RPRM
GALNT13
KCNJ3
NR4A2
GPD2
GALNT5
ERMN
CYTIP
ACVR1C
ACVR1
UPP2
CCDC148
PKP4
DAPL1
TANC1
WDSUB1
BAZ2B
MARCH7
CD302
LY75
PLA2R1
ITGB6
RBMS1
TANK
PSMD14
TBR1
SLC4A10
DPP4
GCG
FAP
IFIH1
GCA
KCNH7
FIGN
GRB14
COBLL1
SLC38A11
SCN3A
SCN2A
CSRNP3
GALNT3
TTC21B
SCN1A
SCN9A
SCN7A
XIRP2
B3GALT1
STK39
CERS6
NOSTRIN
SPC25
G6PC2
ABCB11
DHRS9
LRP2
BBS5
KBTBD10
FASTKD1
PPIG
C2orf77
PHOSPHO2
KLHL23
SSB
METTL5
UBR3
MYO3B
SP5
ERICH2
GAD1
GORASP2
TLK1
METTL8
DCAF17
CYBRD1
DYNC1I2
SLC25A12
HAT1
METAP1D
DLX1
DLX2
ITGA6
PDK1
RAPGEF4
MLTK
CDCA7
SP3
OLA1
SP9
CIR1
SCRN3
GPR155
WIPF1
CHRNA1
CHN1
ATF2
ATP5G3
LNP
EVX2
HOXD13
HOXD12
HOXD11
HOXD10
HOXD9
HOXD8
HOXD4
HOXD3
HOXD1
MTX2
HNRNPA3
NFE2L2
AGPS
TTC30B
TTC30A
PDE11A
RBM45
OSBPL6
PRKRA
DFNB59
FKBP7
PLEKHA3
TTN
CCDC141
SESTD1
ZNF385B
CWC22
UBE2E3
ITGA4
CERKL
NEUROD1
SSFA2
PPP1R1C
PDE1A
DNAJC10
FRZB
NCKAP1
DUSP19
NUP35
ZNF804A
FSIP2
ZC3H15
ITGAV
FAM171B
ZSWIM2
CALCRL
TFPI
GULP1
DIRC1
COL3A1
COL5A2
WDR75
SLC40A1
ASNSD1
ANKAR
OSGEPL1
ORMDL1
PMS1
MSTN
C2orf88
HIBCH
INPP1
MFSD6
TMEM194B
NAB1
GLS
STAT1
STAT4
MYO1B
OBFC2A
SDPR
TMEFF2
SLC39A10
DNAH7
STK17B
HECW2
CCDC150
GTF3C3
C2orf66
PGAP1
ANKRD44
Chr2: 153.2-198.2 Mb
(45 Mb)
Adcy10
Lass6
Dhrs9
Rdh5
Ccdc89
Atp5b
Itgb5
Heg1
Slc12a8
Znf148
C7orf57
Acvr1b
Mylk
Cd63
Ccdc14
Ropn1l
Trio
Umps
Slc40a1
Dnah9l
Tril
Jazf1
HoxD14
HoxD5
HoxD2
Cyct
Tarbp2
Col2a1a
Adat3
Gdf11
Mreg
Fn1
Atic
Bard1
Vwc2l
Scaffold_14
(10 Mb)

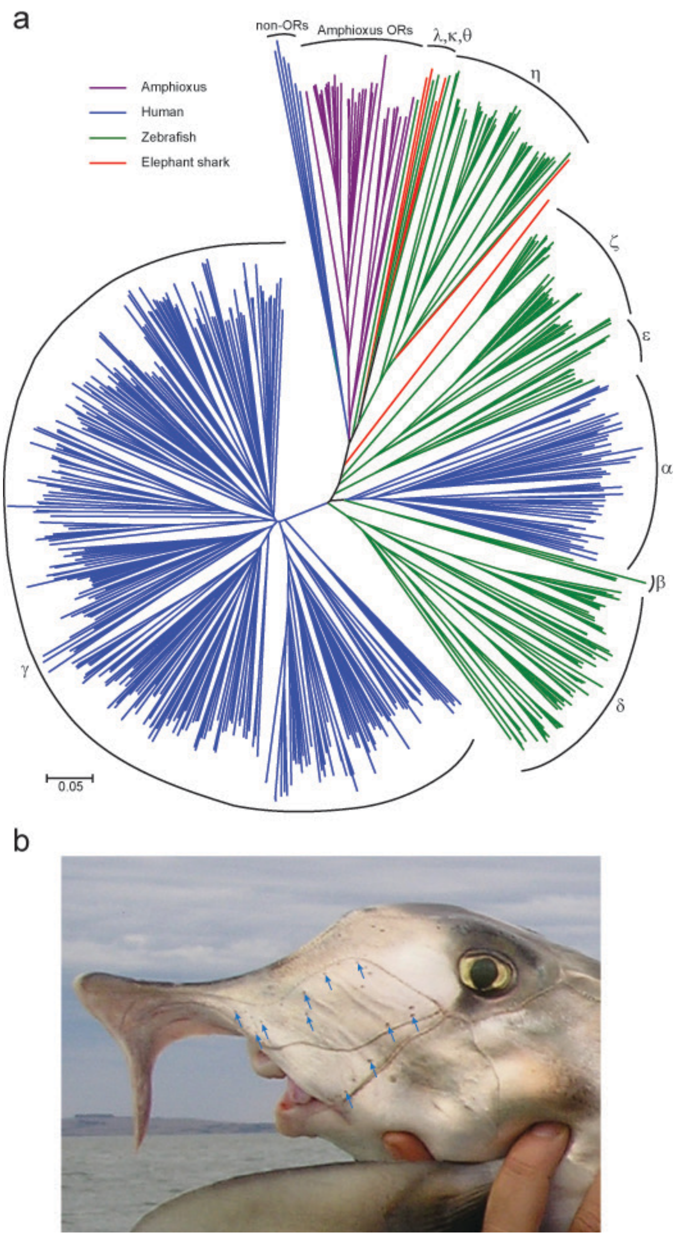
Supplementary Figure VIII.2 | An extensively conserved syntenic block in the *C. milii* and human genomes.

Genes are shown as rectangles and syntenic genes are colored grey. Scaffold ends are marked by grey circles. Synteny of 148 genes on *C. milii* scaffold_14 (10 Mb) is highly conserved in the human Chr_2q23.3 to 2q33.1 (45 Mb) encompassing the HOXD gene cluster (shown in red). In the *C. milii* locus, only genes that are not present in the human locus are labelled.



Supplementary Figure IX.1 | Top 50 protein domains (PFAM) in *C. milii*, stickleback and human proteomes

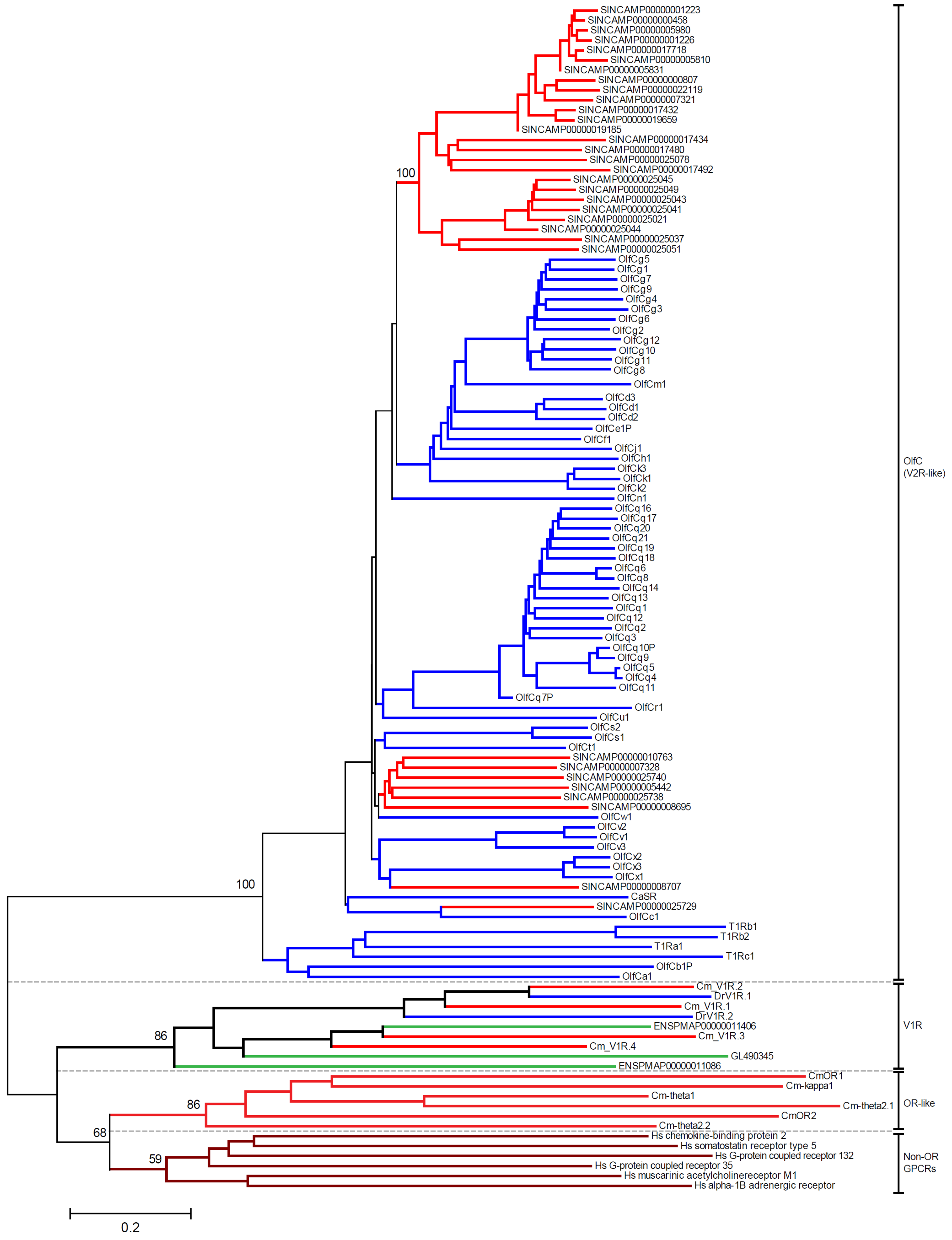
Protein domains are sorted by decreasing number of proteins in *C. milii*.



Supplementary Figure IX.2 | Neighbor joining tree of OR-like genes from *C. milii*, human, zebrafish and amphioxus.

(a) Neighbour joining tree of OR-like genes. A total of 589 protein sequences including 6 non-OR GPCR (outgroup) sequences were aligned using MAFFT version 6.864b (Katoh et al. 2002) and trimmed using GBlocks (Castresana 2000). A neighbor-joining tree was generated using ClustalW. Sequences used as outgroup were: human alpha-1B adrenergic receptor (NP_000670.1), human muscarinic acetylcholine receptor M1 (NP_000729.2), human somatostatin receptor type 5 (NP_001044.1), human chemokine-binding protein 2

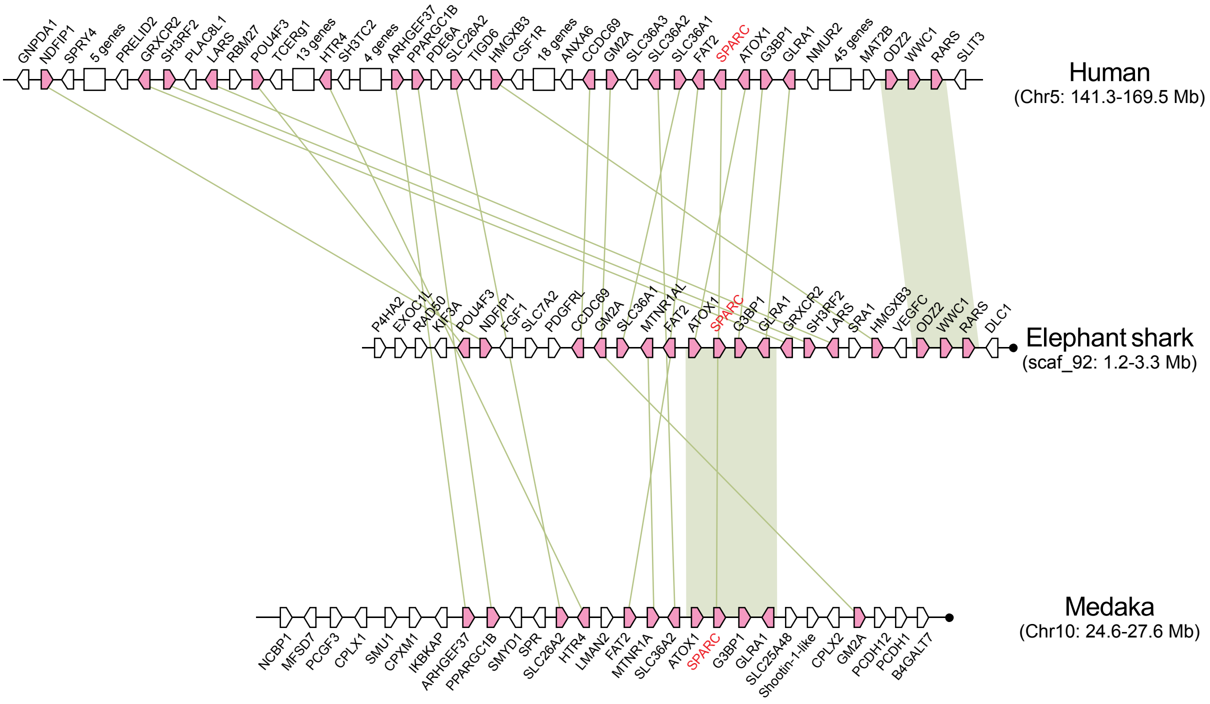
(NP_001287.2), human G-protein coupled receptor 35 isoform a (NP_005292.2), and human G-protein coupled receptor 132 (NP_037477.1). Different groups of ORs are shown in different colours. Bootstrap support percentages ≥50 are shown at the main nodes of the tree. The six elephant shark OR-like sequences are labeled. All non-elephant shark sequences used in the analysis were extracted from the datasets of a previous study (Niimura 2009). **(b)** *C. milii* snout showing the distribution of ampullae of Leorenzini (arrows indicate ampullary pores).



Supplementary Figure IX.3 | Neighbor joining tree of V1R-like and V2R-like (*OlfC*) genes from *C. milii*, zebrafish and lamprey.

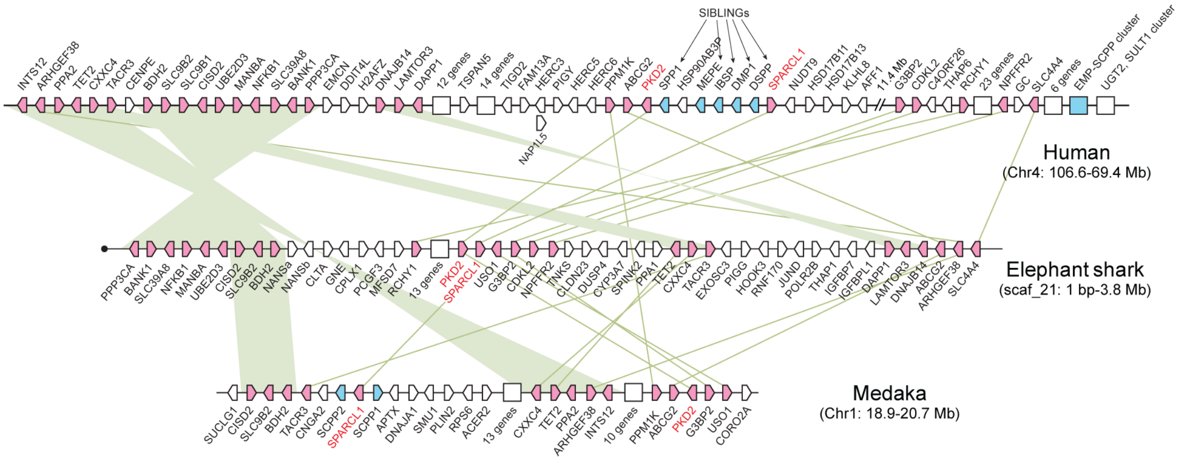
A total of 118 protein sequences including six elephant shark OR-like sequences and six non-OR GPCRs were aligned using MAFFT version 6.864b (Katoh et al. 2002). A neighbor-joining tree was generated using MEGA5 (Tamura et al. 2011) with Poisson correction and

1000 bootstrap replicates. Non-OR GPCR sequences used as outgroups were: human alpha-1B adrenergic receptor (NP_000670.1), human muscarinic acetylcholine receptor M1 (NP_000729.2), human somatostatin receptor type 5 (NP_001044.1), human chemokine-binding protein 2 (NP_001287.2), human G-protein coupled receptor 35 isoform a (NP_005292.2), and human G-protein coupled receptor 132 (NP_037477.1). Bootstrap support percentages ≥50 are shown at the main nodes of the tree. Red, blue and green branches denote *C. milii*, zebrafish and sea lamprey sequences respectively. Zebrafish V1R sequences were retrieved from GenBank whereas V2R sequences were obtained from (Alioto and Ngai 2006). Sea lamprey V2R-like sequences from (Grus and Zhang 2009) were checked against the Petromyzon_marinus_7.0 assembly (www.ensembl.org) to obtain the Ensembl IDs/scaffold number. Non-OR GPCRs are from (Niimura 2009). Brown branches represent human non-OR GPCR sequences.



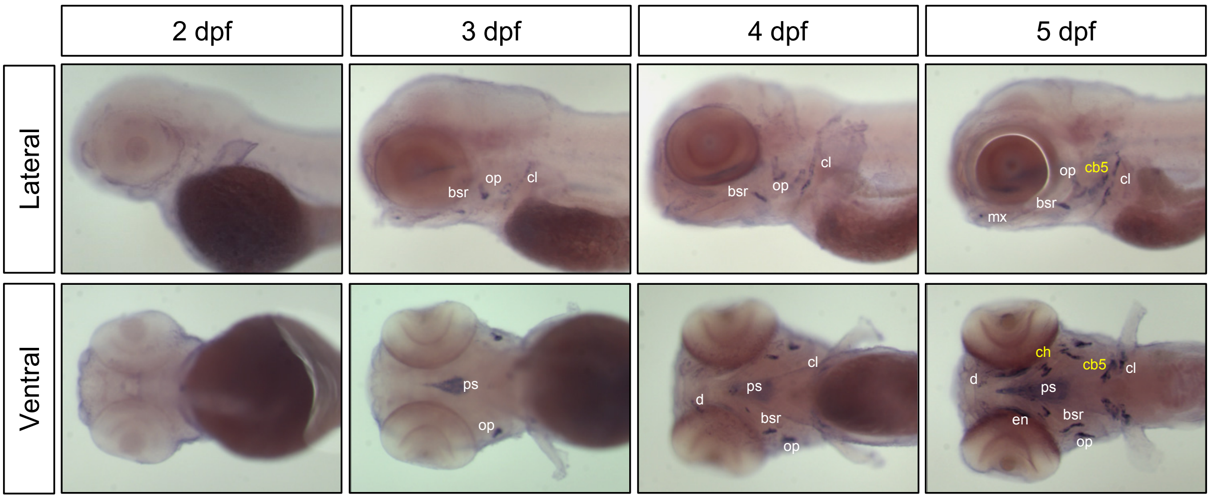
Supplementary Figure X.1 | The *Sparc* gene locus in human, elephant shark (*C. milii*) and medaka.

Genes are represented as block arrows. Syntenic genes are colored. For clarity, clusters of selected non-syntenic genes are grouped and represented as rectangles. Not drawn to scale.



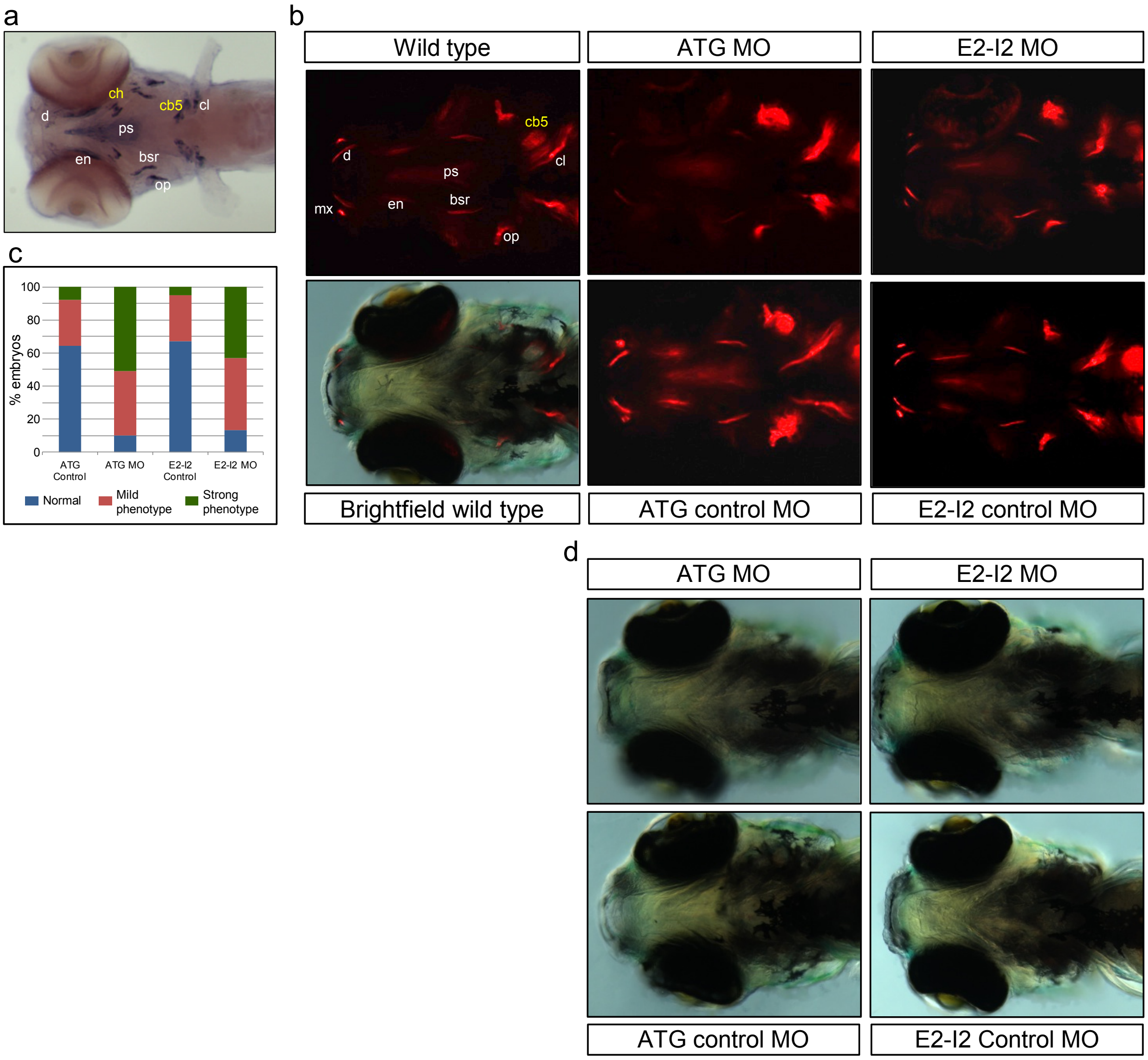
Supplementary Figure X.2 | The *Sparcl1* locus in human, elephant shark (*C. milii*) and medaka.

Genes are represented as block arrows. Syntenic genes are colored pink; SIBLING and other SCPP genes are colored blue. For clarity, clusters of some non-syntenic genes are grouped and shown as rectangles. Not drawn to scale. In zebrafish, *sparcl1* and *pkd2* are located 13.8 Mb apart on Chromosome 1 whereas *spp1* is located on Chromosome 10.

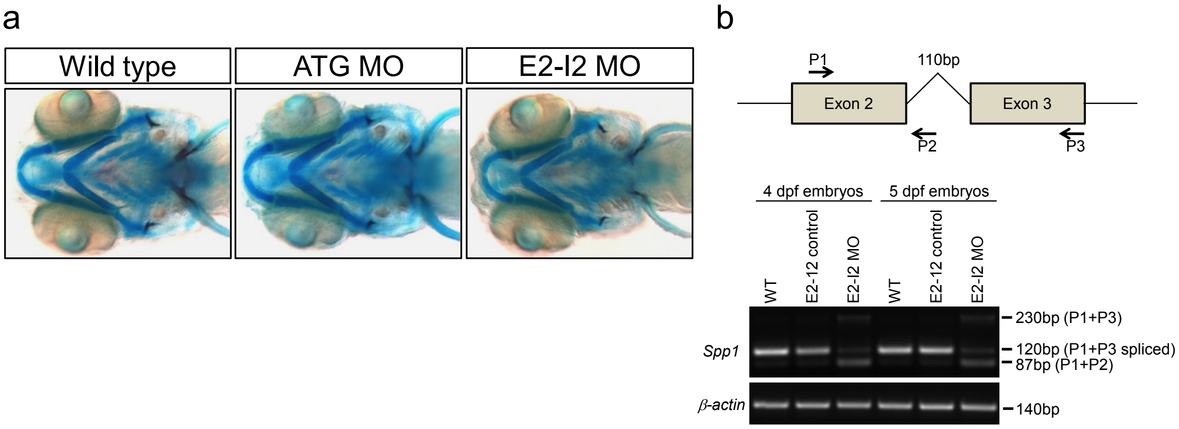


Supplementary Figure X.3 | Zebrafish spp1 is expressed from 2 dpf specifically in cells surrounding the bone matrix.

Lateral and ventral views of 2 to 5 dpf embryos hybridized with *spp1* RNA probe. There was no expression at 1 dpf. Endochondral bones (ch, ceratohyal; cb5, ceratobranchial 5) are highlighted in yellow whereas dermal bones (bsr, branchiostegal ray; cl, cleithrum; d, dentary; en, entopterygoid; mx, maxilla; op, operculum; ps, parasphenoid) are highlighted in white.



Supplementary Figure X.4 | Reduction of bone deposition in spp1 zebrafish morphants. (a) *spp1* is specifically expressed in cells surrounding the bone matrix. Ventral view of a 5 dpf embryo hybridized with *spp1* RNA probe. Endochondral bones (ch, ceratohyal; cb5, ceratobranchial 5) are highlighted in yellow whereas dermal bones (bsr, branchiostegal ray; cl, cleithrum; d, dentary; en, entopterygoid; op, operculum; ps, parasphenoid) are highlighted in white. (b) Ventral views of 5 dpf wild type, ATG MO, ATG control MO, E2-I2 MO and E2-I2 control MO embryos. Embryos were stained with alizarin red to reveal sites of bone deposition and imaged under red fluorescence. For one wild type embryo (bottom left), a merged bright field image and red fluorescence signal is shown to simultaneously visualize anatomical structures and bone deposition. (c) Proportion of embryos showing normal phenotype (resembling wild type), 'mild' bone-phenotype and 'strong' bone-phenotype (the latter showing the most reduction of bones). ATG MO and E2-I2 MO morphants show a significantly higher proportion of strong bone-phenotype ($p<0.01$, Fisher's exact test) compared to their respective controls. (d) Bright field images of embryos shown in panel (b) showing normal growth of morphant embryos.



Supplementary Figure X.5 | *spp1* gene knockdown in zebrafish.

(a) Five dpf zebrafish embryos stained with alcian blue for cartilage. Wild type, ATG morphant and E2-I2 morphant embryos are shown. Morphant embryos do not show any discernible change in cartilage matrix formation. **(b)** E2-I2 MO inhibits splicing of *spp1* in zebrafish. Upper panel shows the position of RT-PCR primers (P1, P2 and P3) in relation to *spp1* gene. Lower panel shows the RT-PCR fragments separated on a 3% agarose gel. The wild type spliced product is represented by a 120 bp fragment, whereas unspliced products are represented by the 87 bp and 230 bp fragments. Splicing of intron 2 is considerably inhibited in 4-dpf and 5-dpf embryos injected with E2-I2 MO, while normal splicing occurs in 4-dpf and 5-dpf embryos that are wild-type or injected with E2-I2 control MO.

a

```
                                   ── Exon 6 target site →
Wild Type   GCATTTACAGGAGGCAGATGAGGAATCTGAAACAGATGAGAAGGAAGAGGAGAATGTGAGTG

ex6_1.1     GCATTTACAGGAGGCAGATGAGGAATCTGAAACAGATgaaggaagagGAGAATGAAGAGGAGAATGTGAGTG +10

ex6_2.1     GCATTTACAGGAGGCAGATGAGGAATCTGAAACAG--------GAAGAGGAGAATGTGAGTG -8
```

b

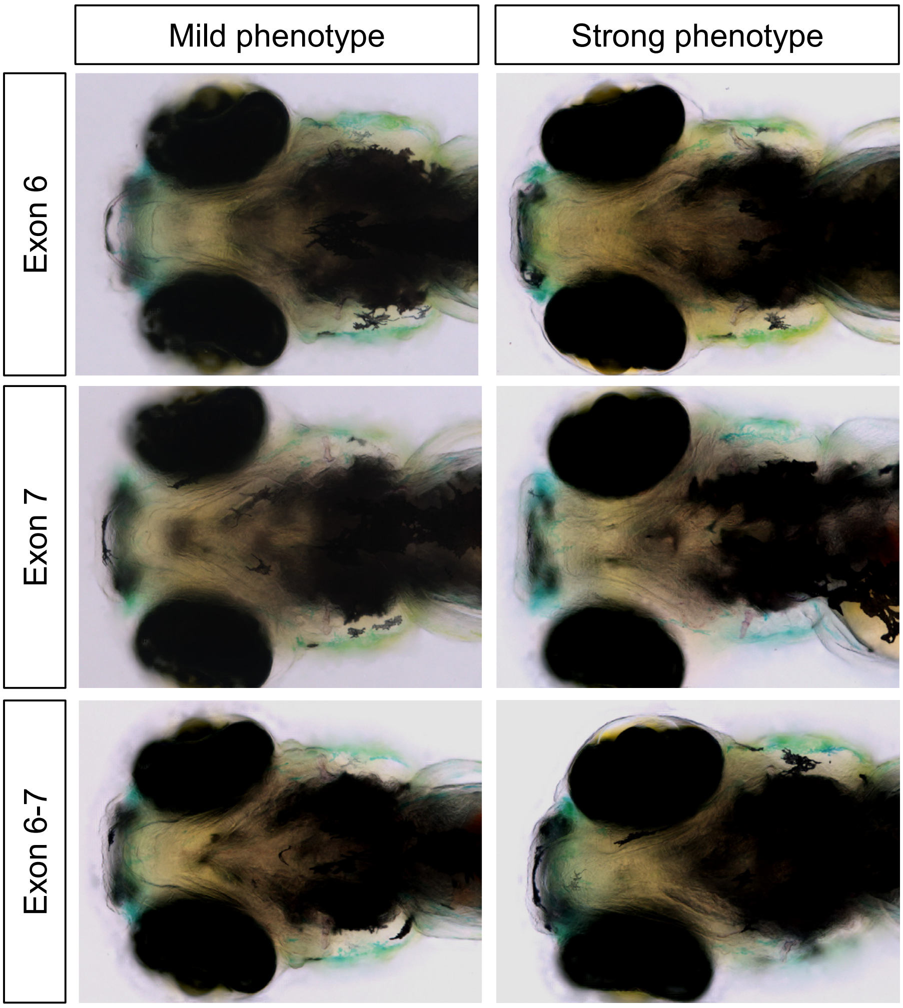
```
                                  ← Exon 7 target site ──
Wild Type CCCATTATCAACACAGGCCGGGGAGACAGTTTGGGCTACCCTAGCGACTACAAAAAATCC

ex7_1.1   CCCATTATCAACACAG---------ACAGTTTGGGCTACCCTAGCGACTACAAAAAATCC -9
ex7_1.2   CCCATTATCAAC-------------ACAGTTTGGGCTACCCTAGCGACTACAAAAAATCC -13
ex7_1.3   CCCATTATCAACACAGGCCGGGG------TTTGGGCTACCCTAGCGACTACAAAAAATCC -6

ex7_2.1   CCCATTATCAACACAGGCCGGG-------------CTACCCTAGCGACTACAAAAAATCC -13
ex7_2.2   CCCATTATCAAC-------------ACAGTTTGGGCTACCCTAGCGACTACAAAAAATCC -13
ex7_2.3   CCCATTATCAACACAGGCCGGGGctggGACAGTTTGGGCTACCCTAGCGACTACAAAAAATCC +3 (-1,+4)

ex7_3.1   CCCATTATCAACACAGGCC--------AGTTTGGGCTACCCTAGCGACTACAAAAAATCC -8
ex7_3.2   CCCATTATCAACACAGGCCattatcaacACAGTTTGGGCTACCCTAGCGACTACAAAAAATCC +3 (-6,+9)
ex7_3.3   CCCATTATCAAC-------------ACAGTTTGGGCTACCCTAGCGACTACAAAAAATCC -13
ex7_3.4   CCCATTATCAACACAGGCCGGGGctACAGTTTGGGCTACCCTAGCGACTACAAAAAATCC 0 (-2,+2)

ex7_4.1   CCCATTATCAACACAGGCCGGGGAtcacagGTTTGGGCTACCCTAGCGACTACAAAAAATCC +2 (-4,+6)
ex7_4.2   CCCATTATCAACACAGGCCGGGG---CAGTTTGGGCTACCCTAGCGACTACAAAAAATCC -3
ex7_4.3   CCCATTATCAAC-------------ACAGTTTGGGCTACCCTAGCGACTACAAAAAATCC -13
```

c

```
                                     ── Exon 6 target site →                       ← Exon 7 target site ──
Wild Type    GCATTTACAGGAGGCAGATGAGGAATCTGAAACAGATGAGAAGGAAG ←── 2.6 kb ──→ AGGCCGGGGAGACAGTTTGGGCTACCCTAGCGACTACAAAAAATCC

ex67_1.1     GCATTTACAGGAGGCAGATGAGTAATCTGAAACAtaattagagtcactcacattctcctcttccttc-----------TGGGCTACCCTAGCGACTACAAAAAATCC -2623 bp (-2656,+33)
ex67_1.2     GCATTTACAGGAGGCAGATGAGGAATCTGAAACAGATGAGA--------------------------------CAGTTTGGGCTACCCTAGCGACTACAAAAAATCC -2644 bp
ex67_1.3     GCATTTACAGGAGGCAGATGAGGAATCTGAAACAGAT--------------------------------------GTTTGGGCTACCCTAGCGACTACAAAAAATCC -2650 bp
ex67_1.4     GCATTTACAGGAGGCAGATGAGGAATCTGAAACC-------------------------------------------------TACCCTAGCGACTACAAAAAATCC -2661 bp
ex67_1.5     GCATTTACAGGAGGCAGATGAGGAATCTGAAACAGA-------------------------------------CAGTTTGGGCTACCCTAGCGACTACAAAAAATCC -2643 bp

ex67_2.1     GCATTTACAGGAGGCAGATGAGGAATCTGAAACAG-----------------------------------------TTTGGGCTACCCTAGCGACTACAAAAAATCC -2653 bp
ex67_2.2     GCATTTACAGGAGGCAGATGAGGAATCTGAAACAtaattagagtcactcacattctcctcttccttc-----------TGGGCTACCCTAGCGACTACAAAAAATCC -2623 bp (-2656,+33)
ex67_2.3     GCATTTACAGGAGGCAGATGAGGAATCTGAAAC-------------------------------------------------CTACCCTAGCGACTACAAAAAATCC -2661 bp

ex67_3.1     GCATTTACAGGAGGCAGATGAGGAATCTGAAACAtaattagagtcactcacattctcctcttccttc-----------TGGGCTACCCTAGCGACTACAAAAAATCC -2623 bp (-2656,+33)
ex67_3.2     GCATTTACAGGAGGCAGATGAGGAATCTGAAAC--------------------------------------GACAGTTTGGGCTACCCTAGCGACTACAGAAAATCC -2650 bp
ex67_3.3     GCATTTACAGGAGGCAGATGAGGAATCTGAAAC--------------------------------------------------------AGCGACTACAAAAAATCC -2668 bp

ex67_4.1     GCATTTACAGGAGGCAGATGAGGAATCTGAAACAGA-------------------------------------CAGTTTGGGCTACCCTAGCGACTACAAAAAATCC -2643 bp
ex67_4.2     GCATTTACAGGAGGCAGATGAGGAATCTGAAAC-------------------------------------------------CTACCCTAGCGACTACAAAAAATCC -2661 bp
ex67_4.3     GCATTTACAGGAGGCAGATGAGTAATCTGAAACAtaattagagtcactcacattctcctcttccttc-----------TGGGCTACCCTAGCGACTACAAAAAATCC -2623 bp (-2656,+33)

ex67_5.1     GCATTTACAGGAGGCAGATGAGGAATCTGAAAC--------------------------------------------------------AGCGACTGCAAAAAATCC -2668 bp
ex67_5.2     GCATTTACGGGAGGCAGATGAGGAATCTGAAACAG-----------------------------------------TTTGGGCTACCCTAGCGACTACAAAAAATCC -2653 bp
ex67_5.3     GCATTTACAGGAGGCAGATGAGGAATCTGAAACAGATGgaaggaaggaagaggagaatgtgagtgactctaatta-------------TAGCGACTACAAAAAATCC -2625 bp (-2662,+37)
ex67_5.4     GCATTTACAGGAGGCAGATGAGGAATCTGAAACAGtc-----------------------------------------TGGGCTACCCTAGCGACTACAAAAAATCC -2647 bp (-2649,+2)
ex67_5.5     GCATTTACAGGAGGCAGATGAGGAATCTGAAAC-------------------------------------------------CTACCCTAGCGACTACAAAAAATCC -2661 bp
ex67_5.6     GCATTTACAGGAGGCAGATGAGGAATCTGAAACAtaattagagtcactcacattctcctcttccttc-----------TGGGCTACCCTAGCGACTACAAAAAATCC -2623 bp (-2656,+33)
```

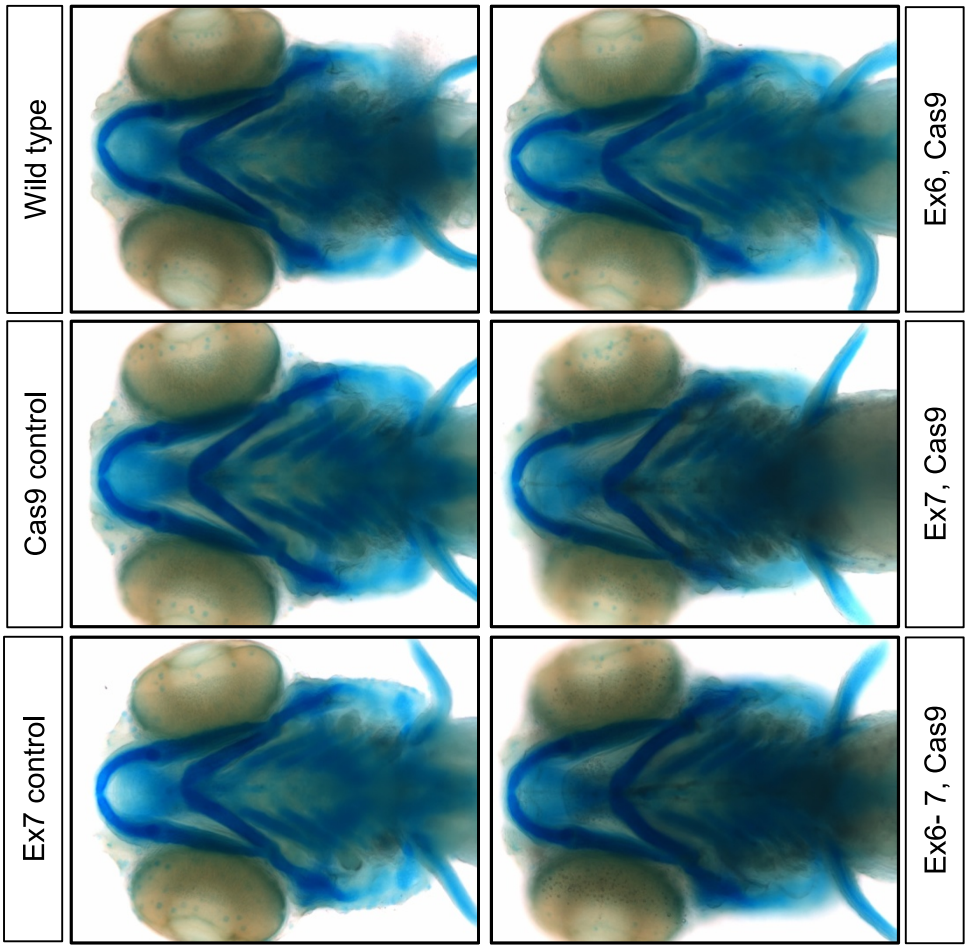
Supplementary Figure X.6 | Targeted indel mutations induced by sgRNA:Cas9 mRNA in the zebrafish spp1.

Either exon 6 (a), exon 7 (b) or both exons (c) were targeted. Target sequences of embryos were PCR amplified, cloned and sequenced. Allele sequences of representative embryos with indel mutations are shown. Wild type alleles are not shown. The wild type reference sequence is shown at the top with the target site highlighted in yellow and protospacer adjacent motif (PAM) sequence highlighted in red. Deletions are shown as dashes, and insertions as lower case letters highlighted in grey. The net change in the length caused by each indel mutation is shown to the right of the sequence (+, insertion; - deletion). In general, the strength of the phenotype seems to be related more to the extent of somatic chimaerism rather than to the type of indels.



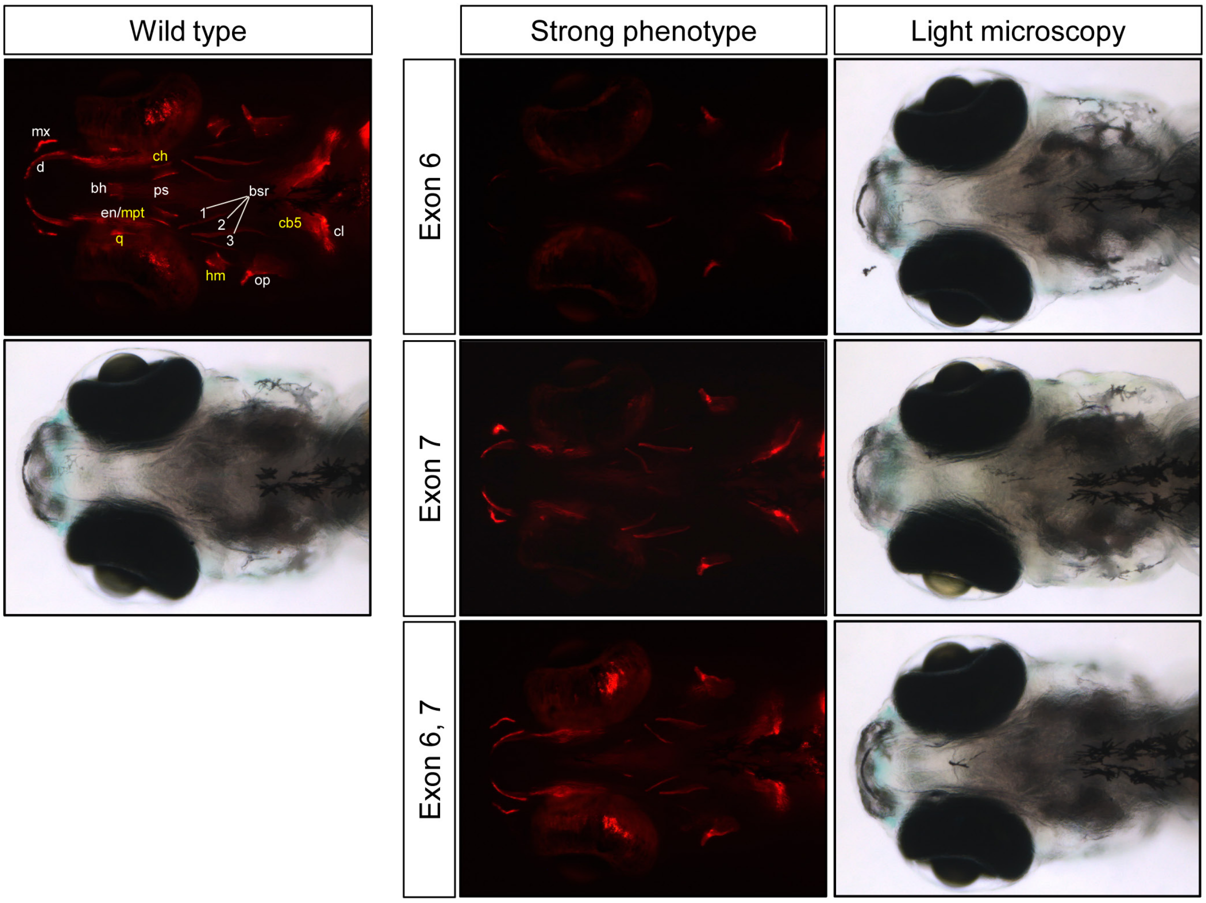
Supplementary Figure X.7 | Brightfield images of embryos shown in Fig. 4d.

Embryos injected with Cas9 mRNA + various sgRNA showed an overall growth comparable to wild type.



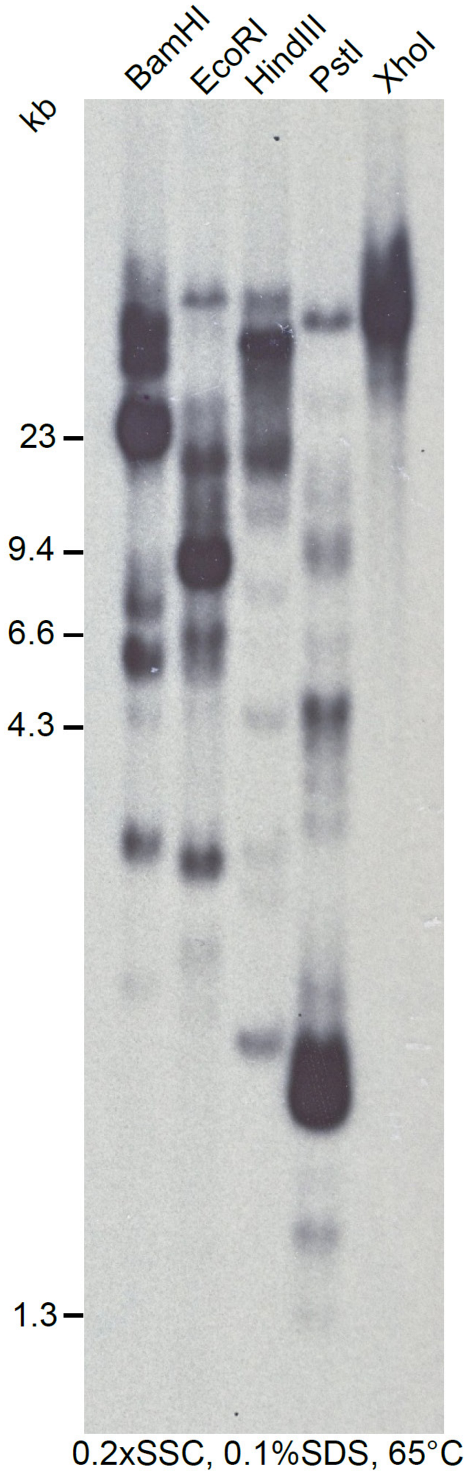
Supplementary Figure X.8 | Alcian blue staining of embryos injected with Cas9 mRNA and various sgRNAs.

Although bone formation was reduced in embryos treated with Cas9 mRNA + sgRNAs, the cartilage development was unaffected.



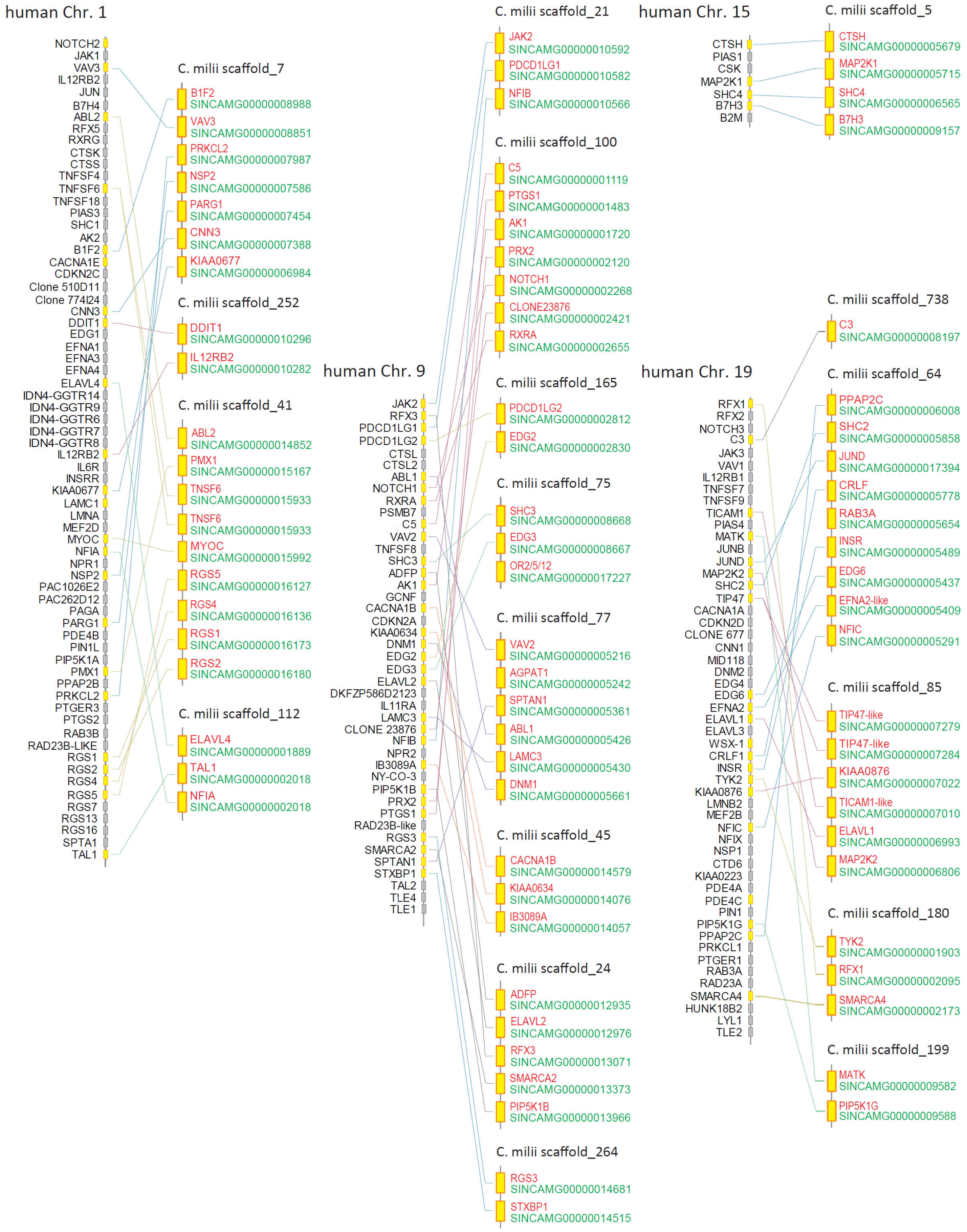
Supplementary Figure X.9 | Targeted mutagenesis of zebrafish *spp1* by sgRNA:Cas9 results in reduced bone formation in 15 dpf embryos.

Left panel: ventral view of a 15 dpf wild type embryo stained with Alizarin red to reveal sites of bone deposition (red fluorescence). Its light microscopy image is given below. Endochondral bones (cb5, ceratobranchial 5; ch, ceratohyal; hm, hyomandibula; mpt, metapterygoid; q, quadrate) are highlighted in yellow whereas dermal bones (bh, basihyal; bsr, branchiostegal ray; cl, cleithrum; d, dentary; en, entopterygoid; mx, maxilla; op, operculum; ps, parasphenoid) are highlighted in white. Entopterygoid (en) and metapterygoid (mpt) are fused and hence cannot be readily distinguished. However, both are affected in *spp1* mutants shown in the right panel. Right panel: ventral views of 15 dpf embryos injected with *Cas9* mRNA together with single guide RNA (sgRNA) targeting *spp1* exon 6, exon 7 or both exons. The 'strong' bone-reduction phenotype embryos are shown (stained with Alizarin red). Light microscopy images of the mutants show normal development of the embryos.



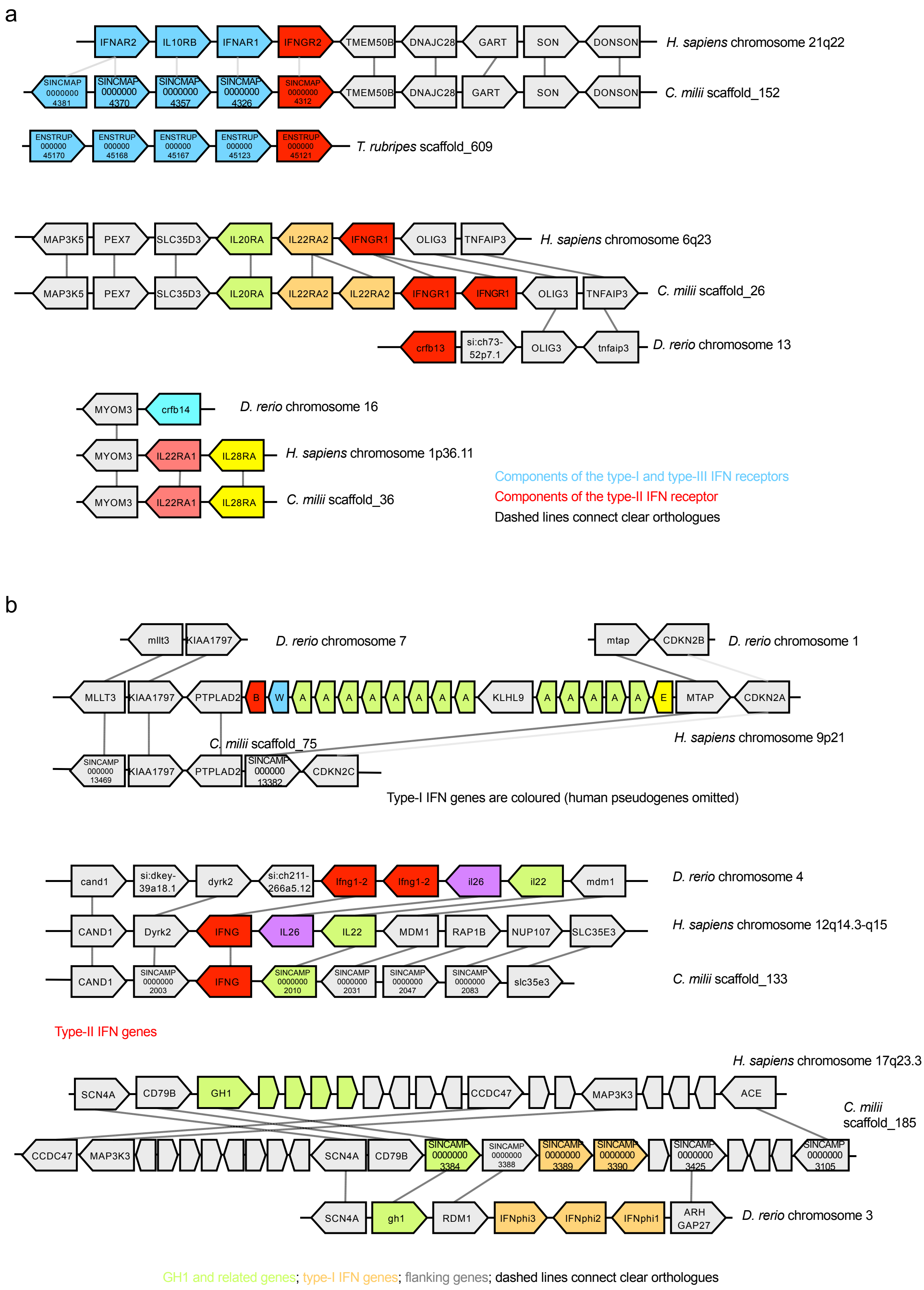
Supplementary Figure XI.1 | Multiple IgM genes in the genome of *C. milii*.

Sothern blot of genomic DNA of *C. milii* digested with the indicated restriction enzymes and hybridized with an *IgM* probe using the indicated hybridization conditions.

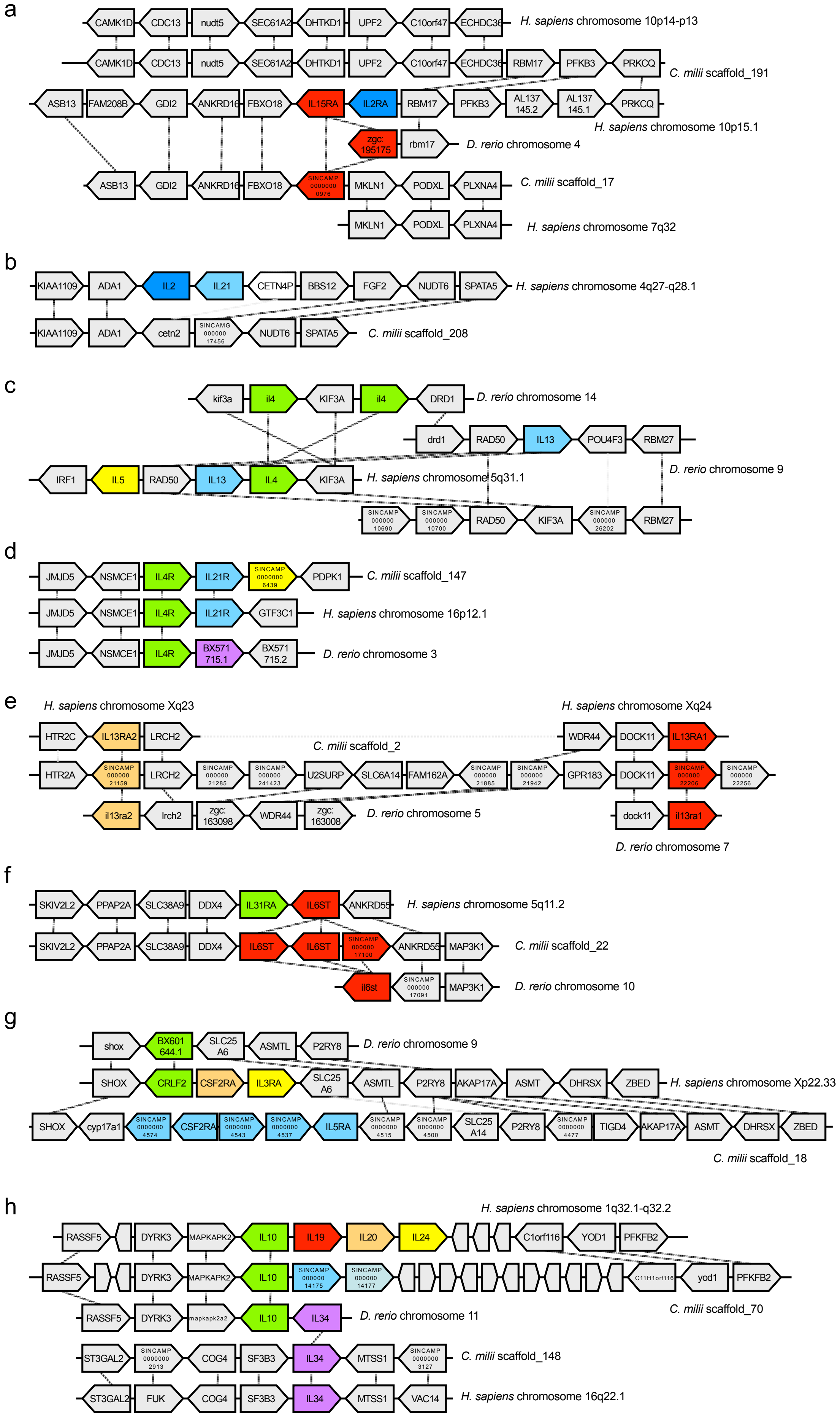


Supplementary Figure XI.2 | MHC paralogous groups in *H. sapiens* and *C. milii*.

The four paralogous human chromosomes are indicated and aligned with the presumptive *C. milii* counterparts.



Supplementary Figure XI.3 | Chromosomal organization of interferon receptor and interferon genes in vertebrate genomes.
a, Receptor genes. Species and chromosomal locations are indicated. **b**, Ligand genes.


a
CAMK1D
CDC13
nudt5
SEC61A2
DHTKD1
UPF2
C10orf47
ECHDC3
H. sapiens chromosome 10p14-p13
CAMK1D
CDC13
nudt5
SEC61A2
DHTKD1
UPF2
C10orf47
ECHDC3
RBM17
PFKB3
PRKCQ
C. milii scaffold_191
ASB13
FAM208B
GDI2
ANKRD16
FBXO18
IL15RA
IL2RA
RBM17
PFKB3
AL137 145.2
AL137 145.1
PRKCQ
H. sapiens chromosome 10p15.1
zgc: 195175
rbm17
D. rerio chromosome 4
ASB13
GDI2
ANKRD16
FBXO18
SINCAMP 0000000 0976
MKLN1
PODXL
PLXNA4
C. milii scaffold_17
MKLN1
PODXL
PLXNA4
H. sapiens chromosome 7q32
b
KIAA1109
ADA1
IL2
IL21
CETN4P
BBS12
FGF2
NUDT6
SPATA5
H. sapiens chromosome 4q27-q28.1
KIAA1109
ADA1
cetn2
SINCAMG 000000 17456
NUDT6
SPATA5
C. milii scaffold_208
c
kif3a
il4
KIF3A
il4
DRD1
D. rerio chromosome 14
drd1
RAD50
IL13
POU4F3
RBM27
IRF1
IL5
RAD50
IL13
IL4
KIF3A
H. sapiens chromosome 5q31.1
D. rerio chromosome 9
SINCAMP 000000 10690
SINCAMP 000000 10700
RAD50
KIF3A
SINCAMP 000000 26202
RBM27
d
JMJD5
NSMCE1
IL4R
IL21R
SINCAMP 0000000 6439
PDPK1
C. milii scaffold_147
JMJD5
NSMCE1
IL4R
IL21R
GTF3C1
H. sapiens chromosome 16p12.1
JMJD5
NSMCE1
IL4R
BX571 715.1
BX571 715.2
D. rerio chromosome 3
e
H. sapiens chromosome Xq23
H. sapiens chromosome Xq24
HTR2C
IL13RA2
LRCH2
WDR44
DOCK11
IL13RA1
C. milii scaffold_2
HTR2A
SINCAMP 000000 21159
LRCH2
SINCAMP 000000 21285
SINCAMP 000000 241423
U2SURP
SLC6A14
FAM162A
SINCAMP 000000 21885
SINCAMP 000000 21942
GPR183
DOCK11
SINCAMP 000000 22206
SINCAMP 000000 22256
il13ra2
lrch2
zgc: 163098
WDR44
zgc: 163008
D. rerio chromosome 5
dock11
il13ra1
D. rerio chromosome 7
f
SKIV2L2
PPAP2A
SLC38A9
DDX4
IL31RA
IL6ST
ANKRD55
H. sapiens chromosome 5q11.2
SKIV2L2
PPAP2A
SLC38A9
DDX4
IL6ST
IL6ST
SINCAMP 000000 17100
ANKRD55
MAP3K1
C. milii scaffold_22
il6st
SINCAMP 000000 17091
MAP3K1
D. rerio chromosome 10
g
shox
BX601 644.1
SLC25 A6
ASMTL
P2RY8
D. rerio chromosome 9
SHOX
CRLF2
CSF2RA
IL3RA
SLC25 A6
ASMTL
P2RY8
AKAP17A
ASMT
DHRSX
ZBED
H. sapiens chromosome Xp22.33
SHOX
cyp17a1
SINCAMP 0000000 4574
CSF2RA
SINCAMP 0000000 4543
SINCAMP 0000000 4537
IL5RA
SINCAMP 0000000 4515
SINCAMP 0000000 4500
SLC25 A14
P2RY8
SINCAMP 0000000 4477
TIGD4
AKAP17A
ASMT
DHRSX
ZBED
C. milii scaffold_18
h
H. sapiens chromosome 1q32.1-q32.2
RASSF5
DYRK3
MAPKAPK2
IL10
IL19
IL20
IL24
C1orf116
YOD1
PFKFB2
RASSF5
DYRK3
MAPKAPK2
IL10
SINCAMP 000000 14175
SINCAMP 000000 14177
C11H1orf116
yod1
PFKFB2
C. milii scaffold_70
RASSF5
DYRK3
mapkapk2a2
IL10
IL34
D. rerio chromosome 11
ST3GAL2
SINCAMP 0000000 2913
COG4
SF3B3
IL34
MTSS1
SINCAMP 0000000 3127
C. milii scaffold_148
ST3GAL2
FUK
COG4
SF3B3
IL34
MTSS1
VAC14
H. sapiens chromosome 16q22.1

Supplementary Figure XI.4 | Chromosomal organization of interleukin receptor and interleukin genes in vertebrate genomes.

Dashed lines connect orthologues.

```
                                 #          ***      *               #
Hs NMRPPFTYATLIRWAILEAPEKQRTLNEIYHWFTRMFAFFRNHPATWKNAIRHNLSLHKCFVRVESEKGAVWTVDELEFRKKR
Dr .........SM......KS....L..K...Q...SM.FY..HNT......V..............GR..S......E..LRRK
Cm .I.......A........TS...M......L....K......NT......V..............NMR.A......M..QRRK
Gc .V...L...A........T.DR.L......Q....T.....YNT......V..............NV..S....N....ERR.
```

Supplementary Figure XI.5 | Protein sequence comparison of DNA binding domains of *Foxp3*-like genes.

Hs, *Homo sapiens*; Dr, *Danio rerio*; Cm, *Callorhinchus milii*; Gc, *Ginglymostoma cirratum*. Signature residues predicted to be involved in NFAT interaction (#) or to affect DNA binding (*) are marked.

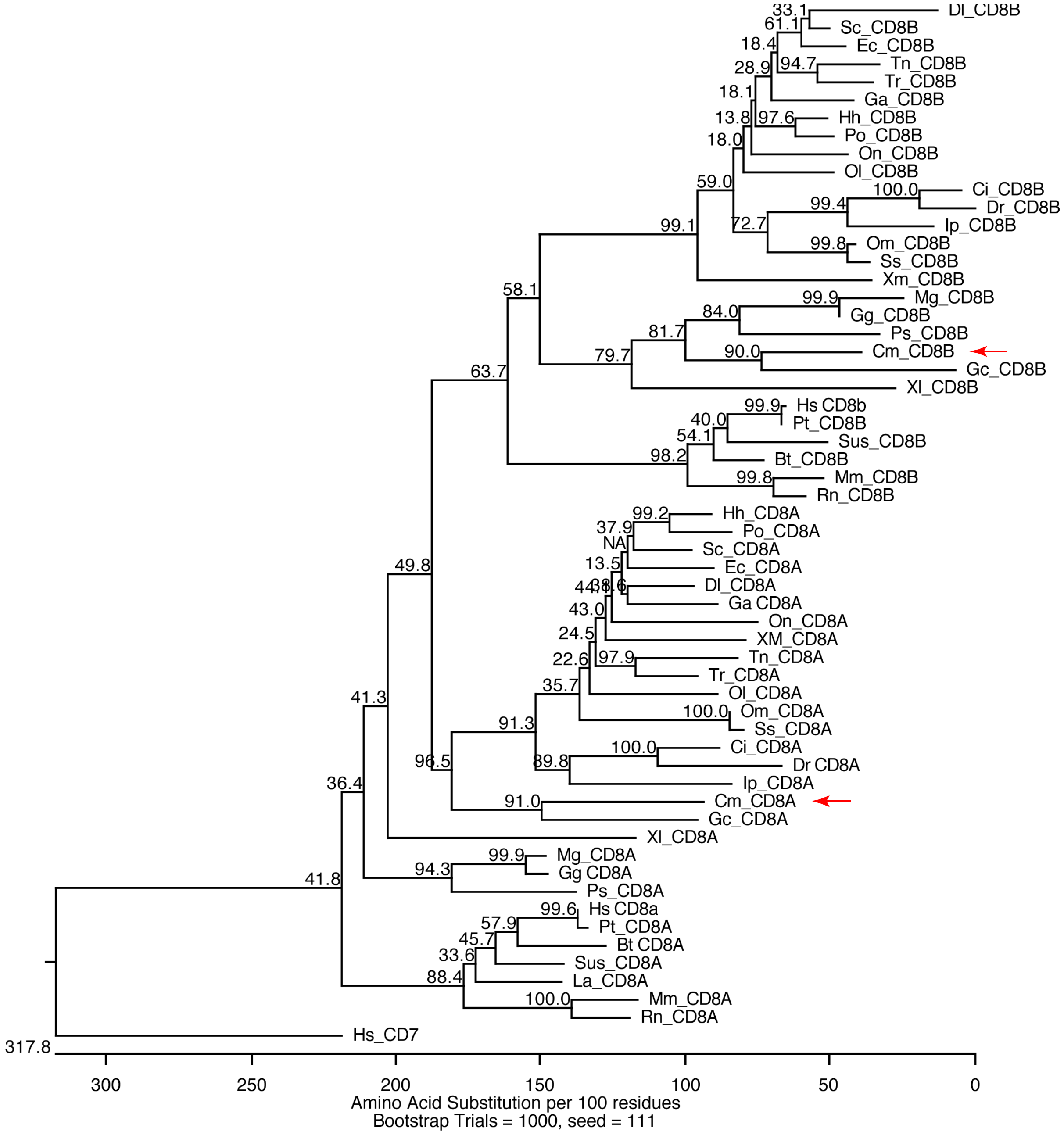
```
Hs_ MGSPEDDLIGIPFPDHSSELLSCLNEQRQLGHLCDLTIRTQGLEYRTHRAVLAACSHYFKKLFT-EGGGGAVMGAGGSGTATGGAGAG-VC
Lc_ MGTTEDGLIGIPFPEHSNELLSSLNEQRHKGLLCDVTIKTRGLEYRTHRAVLAACSQYFKKLFT----------------CGTMAGQQDVC
Xt_ MASSEDELIGIPFPEHSSELLSSLNEQRHRGVLCDITIKTRGLEYRTHRAVLAACSDYFRKMFT---------------GVPTRGKCPDVC
Gm_ MSPGEDGLIGIPFPEHSNELLSRLNDQRRAGLLCDLTLTSRGERYPTHRSVMAAVSLYFRRLFG------------RGEGGRGGGGGFSVC
Dr_
Cm_ MGPGPEEGLIGIPFPQHSSELLRGLNEQRRRGLLCDLTLVTQGLEYRTHRSVLAACSLYFRRLFGGGGGGGGGGGGGGGGRGGVGGHRNVC
Sc_ MGTAEDELIGIPFPEHSSELLNSLNEQRHKGLLCDVTIVTQGLEYRTHRAVLAGCSR>
Re_                                    VTIVTQGLEYRTHRAVLAACSHYFRKLFT----------------SKPYSGQRNVC
Gc_

Hs_ ELDFVGPEALGALLEFAYTATLTTSSANMPAVLQAARLLEIPCVIAACMEILQGSGLEAPSPDEDDCERARQYLEAFATATASGVPNGEDS
Lc_ ELDFVEPEVMGALLEFAYTATLTISSSNMREMLHAAQMLEIQCVMDACADILRSSGGCVSAPEPPVMAHGEQNSYEKNKQYLDCFAAVTNG
Xt_ QLDFLKPDALSALLDFAYTATLTISNANMRDVLRAARLLEIPCVVDACVEILQCNGHREEMGGDAEDLECFLRARQYLECYMAQENGESAA
Gm_ QLDCVAPDALDALLEFAYTATLTIRSSGMRDVLRAAQLLGIKCVADACRDILGEKEEVVVEEQGRKAEKERKLWVREMEKVSRTELLKPHP
Dr_                            <MRDVLKGAQLLGIQCVADACRDILGETGDAPTDAVEEAEPLPSRRKQDRCSVSPARPVCRRSE
Cm_ ELDFVPPEVLSALLEFAYTATLTISSSNMREVLRASRVLEIGCVAQACADILGQTEGHAKEAWAGGEGAEGGEGFGCPPHAKSAPGPGLGR
Sc_                                                                               <KEAMDSRGSTVG
Re_ ELDFIQPPVLAALLEFAYTATLTISGANMREVLRAARVLEIQCVADACLDILRCSGEGEQEEQLLQQE>
Gc_

Hs_ PPQVPLPPPPPPPPRPVARRSRKPRKAFLQTKGARANHLVPEVPTVPAHPLTYEEEEVAGRVGSSGGSGPGDSYSPPTGTASPPEGPQSYE
Lc_ ICNDDGYTRGVIIKKRQEGGGYKKRQKFLRNFRSAEKKLIPEGELPPAMLNDFPYPPKTEERYSPIANPRDADPASMPISENSNHRHQYLP
Xt_ LSPQADSPPPHPHNIPVPPKSVQIIPRRGRKKFLQVNPNRRNHNGSPFRVADDLLDRDGGHAEALSPASVPPGEPHLSYERYAADNGLGGQ
Gm_ HTPGLVLLPAARQHPLAALQSASQPGAGGPPQGPAERGLRRRPRAREGHRAQRRAAPTTDKLSESEDLGEDSGMREMMTPSLPASMEGGAA
Dr_ DIHGYEGHPPPAAGVLMNGGGTWLQEATPIPRRPEDTPS----------------------------------------------------
Cm_ EEEEEEEEERGHATRYQPHASRQPPGPHPALAPPPEAPRRGRKRPERRRKLKPPTPLSLRREEEEEEEEEGVRRGRGSLYRIVRPLRTCLR
Sc_ AAVVVVAAEEEEEAAGPVDVIQTSEVPKFRH--RKRPKKLPRVPLNHRRSLYRIIQPLRASMNNLISEHAAEQRADDSPVRQEDPYPTPLGD
Re_
Gc_

Hs_ PYE-GEEEEEELVYPPAYGL-AQGGGPPLSPEELGSDEDAIDPDLMAYLSSLHQDNLAPGLDSQDKLVRKRRSQMPQECPVCHKIIHGAGK
Lc_ DSL-YNEEQRPPLHYPQAS------------PEQPLLTDEETVEPGSYWGPTNDPEINPTLSNPDKLVRKRRSQMPQECPVCHKIIHGAGK
Xt_ TIFVPPSPPEEILSDEET----------------------SDMVFQNPYDPENPELVASGLDGADKLVRKRRSQLPQECPVCHKIIHGAGK
Gm_ ARA-------------------------------------------------------------GGGGRKRKSQTPQQCPVCQKIIHGAGK
Dr_ ---EEEESGLQGRATPHHSQSQLADGGGG--------------------------------GVTAVSGRKRKSQTPQQCPVCQKIIHGAGK
Cm_ RPG*NGGLSNGGGGGEEEEEEEEEGGEPRAPATAAGGGGEGGGGGGGGGGGGGGLAEAQQEPGGLPGTRRRKSQMPQSCPICQKVIHGAGK
Sc_ QRGGGEEEEAHKDVERNEELGYQPLLASPVDLRAVAEEEAAELDPASYLNSLSNGILSNSLGLPDKLVRRRKSQMPQECPICRKVIHGAGK
Re_
Gc_                            <PPLAGDEEGAELPASTYLSLVSNGLLGDGTLSSLDRAPGARRRKSQMPQECPICHKVIHGAGK

Hs_ LPRHMRTHTGEKPFACEVCGVRFTRNDKLKIHMRKHTGERPYSCPHCPARFLHSYDLKNHMHLHTGDRPYECHLCHKAFAKEDHLQRHLKG
Lc_ LPRHMRTHTGEKPFACEECGVRFTRNDKLKIHMRKHTGERPYSCDHCEARFLHSYDLKNHMYLHTGDRPFECTLCHKAFAREDHLQRHLKG
Xt_ LPRHMRTHTGEKPFVCQVCGVRFTRNDKLKIHMRKHTGERPYCCEHCSARFLHSYD>
Gm_ LPRHMRTHTGEKPFQCSACGVRFTLNDKLKIHMRKHTGERPYPCTHCPARFLHSYDLKNHLSLHSGARPFECPLCHKAFAREDHLQRHRKG
Dr_ LPRHMRTHTGEKPFQCTACGVRFTRNDKLKIHMRKHTGERPYPCPSCPARFLHSYDLKNHLSLHSGDRPFECPLCHKAFAREDHLQRHRKG
Cm_ LPRHMRTHTGEKPFACHVCGVRFTRNDKLKIHMRKHTGERPYACELCDARFLHGYDLKNHLRLHTGDRPFE>
Sc_ LPRHMRTHTGEKPFACEVCDVRFTRNDKLKIHMRKHTGERPYSCDCCDARFLHSYDLKNHARLHTGDRPFECSQCRKAFVRIDHLQRHLKG
Re_                        <RNDKLKIHMRKHTGERPYCCNCCDARFLHSYDLKNHARLHTGDRPFECSQCRKAFVRLDHLHRHLKG
Gc_ LPRHIRTHTGEKPFACQVCGVRFTRNDKLKIHMRKHTGERPYSCNCCDARFLHSYDLKNHARLHTGDRPFECSQCRKAFVRIDHLQRHLKG

Hs_ QNCLEVRTRRRRKDDAPPHYPPPSTAAASPAGLDLSNGHLDTFRLSLARFWEQSAPTGPPVSTPGPPDDDEEEGAPTTPQAEGAMESS
Lc_ QNCLEVRTRKRRRDDDFKEGDY>
Xt_ LKNH>
Gm_ HSCLEQRPRRPRR
Dr_ HSCLELRPRRPRRTPGPAHSSDSPPPESFSLHQQHAEHMAGLLEGPGLPLRIPYPELLWRAVAVAGPMGKDGGSSPHGGSTI (+ 20aa)
Cm_
Sc_ QNCLEFRTRRRKGDCRPLEEQEVNWEHLRRDGAFPEEYRITRESPGGESPFPEDYGIRRESPGGESPFPE>
Re_ QNCLEVRTRRRKDDDGGDGDGDD>
Gc_ QNCLEIRTRKRRGESQGLQEHPVNWERLAVDGAFPEEYGIERETPGIDCVFPEGYRIKRENPRKDGAFPEEYGIEKESPGI (+ 85aa)
```

*: PCGPLARAPQHPPPAPAEEEEEEEEEEGERGNAAANAVEEDEEEEEEEEEEEEAK

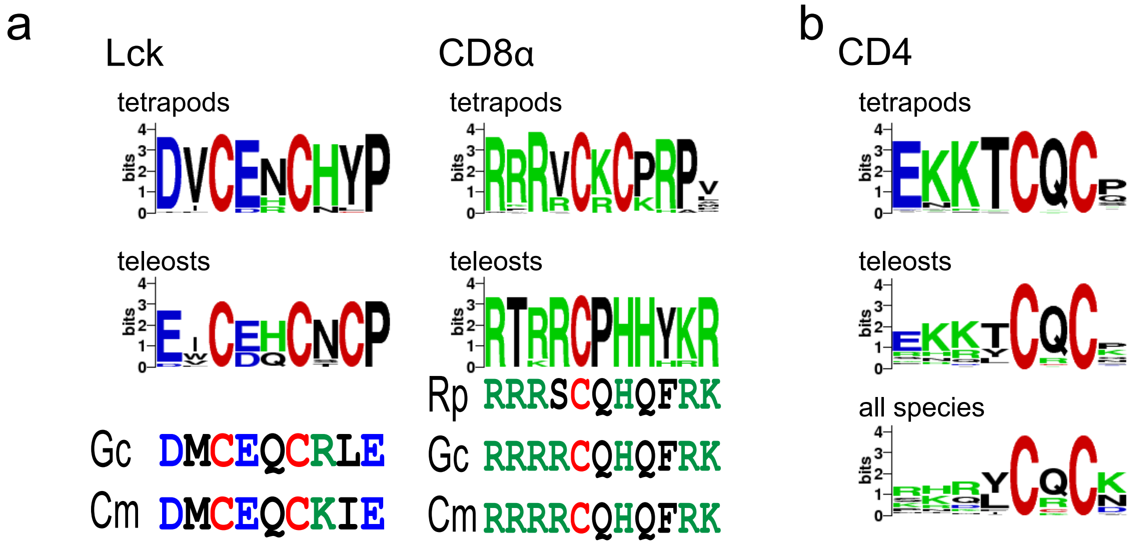
-: indicates gap introduced into contiguous sequence to maximize alignment

Hs, *Homo sapiens*; Lc, *Latimeria chalumnae* (coelecanth); Xt, *Xenopus tropicalis* (tropical clawed frog); Gm, *Gadus morhua (cod)*; Dr, *Danio rerio* (zebrafish); Cm, *Callorhinchus milii* (elephant shark); Sc, *Scyliorhinus canicula* (catshark); Re, *Raja erinacea* (little skate), Gc, *Ginglymostoma cirratum* (nurse shark; KC763332)

Supplementary Figure XI.6 | *Zbtb7B*/*ThPOK*-like genes in vertebrates.



Supplementary Figure XI.7 | Phylogenetic analysis of CD8-like proteins of vertebrates. The sequences used for this analysis are listed in Supplementary Table IX.11. Arrows indicate the positions of *C. milii* CD8A and CD8B proteins.



c

```
H. sapiens    MNRGVPFRHLLLVLQLALLPAATQG-----------------------------------
G. cirratum   MWACVCDLFLSLSLCVFISSLPLCCPLFLSSQRVVCGKVWKSFSAAALLGRLVAFGTDAE

H. sapiens    ------------------------------------------------------------
G. cirratum   LDTHPVPLPFSFHPETSQVIPGETEQGWLAVVELGPMDVPTSPLVRLCGIFLLLITALPP

H. sapiens    -----KKVVLGK-----KGDTVELTCTASQ--KKSIQ-----FHWKNSN----QIKILGN
G. cirratum   GSRVSPYVTEGDTVYACQGETITLLCQVPDVVSRPITSPPGFWKWTSADGTGTTLTILQY

H. sapiens    QGSFLTKGPSKLNDRA--DSRRSLWDQGNFPLIIKNLKIEDSDTYICEVE----DQKEEV
G. cirratum   LSSVRSNSFSKLSARSRISERRQF---GNFSLLISSLDRSDSGSYSCEFSFGRSQARATW

H. sapiens    QLLVFGLTANSDTHLLQGQSLTLTLESPPGSSPSVQCRSPRGKNIQGGKTLSVSQLELQD
G. cirratum   QLRVIEVKATMANPLIETQRVELTCE---GSVQNVSWSGPLGPA-GKGRTLALSNLAVQH

H. sapiens    SGTWTCTVLQNQKKVE--FKIDIVVLAFQKASSIVYKKEGEQVEFSFPLAFTVEKLTGSG
G. cirratum   QGDWVCTCWFPGGTVQSRYQLDVVGLNEPLDKPVFLPVSS---AFLLPCRLN-KALLPLK

H. sapiens    ELWWQAER--------------ASSSKSWITFDLKNKEVSVKRVTQDPKLQMGKKLPLHL
G. cirratum   AAWYQDGQELITLKADSVTKTWSKPQVPWVLFSSSQPITNLSVMVRAVTLAQGGTFECRV

H. sapiens    TLPQALPQYAGSGNLTLALEAKTGKLHQEVNLVVMRATQLQKNLTCEVWGPTSPKLMLSL
G. cirratum   TLKGVTIRRMVNVTLIEVRGSHPTPVPVGTNMSLVCNVSSHSGQTGIRW--RSPSTMEGL

H. sapiens    KLENKEAKVSKREKAVWVLNPEAGMWQCLLSDS------GQVLLESNIKVLPTWSTPVQP
G. cirratum   EDRRVRGEGSLLIRLIEVTQRHVGDWICEISQGDQLCGQGTYSLNITTLTLSEFGDP--P

H. sapiens     MALIVLGGVAGL-LLFIGLGIFFCV--RCRHRRRQAERMS-----QIKRLLSEKK----
G. cirratum    LVLIIAASVGAFVLLLLATVIAVCLSKRARRRRRALKRLRHPLCREHSYQLSNQPLCHS
C. milii      <LYYVTPGVLGVLLIVILVLSVCSIKH-RQTRRRR-LRRMKYPLCRVHSNQLSNQPLCGS

H. sapiens    --------------TCQCPHRFQKTCSPI---------------
G. cirratum   NDYTPGDRPLPPPPIPYCPHRQPRKGRPSHARGSRHARRSQFGP
C. milii      NDYVPSHRPLPTPPRMSCPHKQRTTRKPRAGAGGRVY
                               *
```

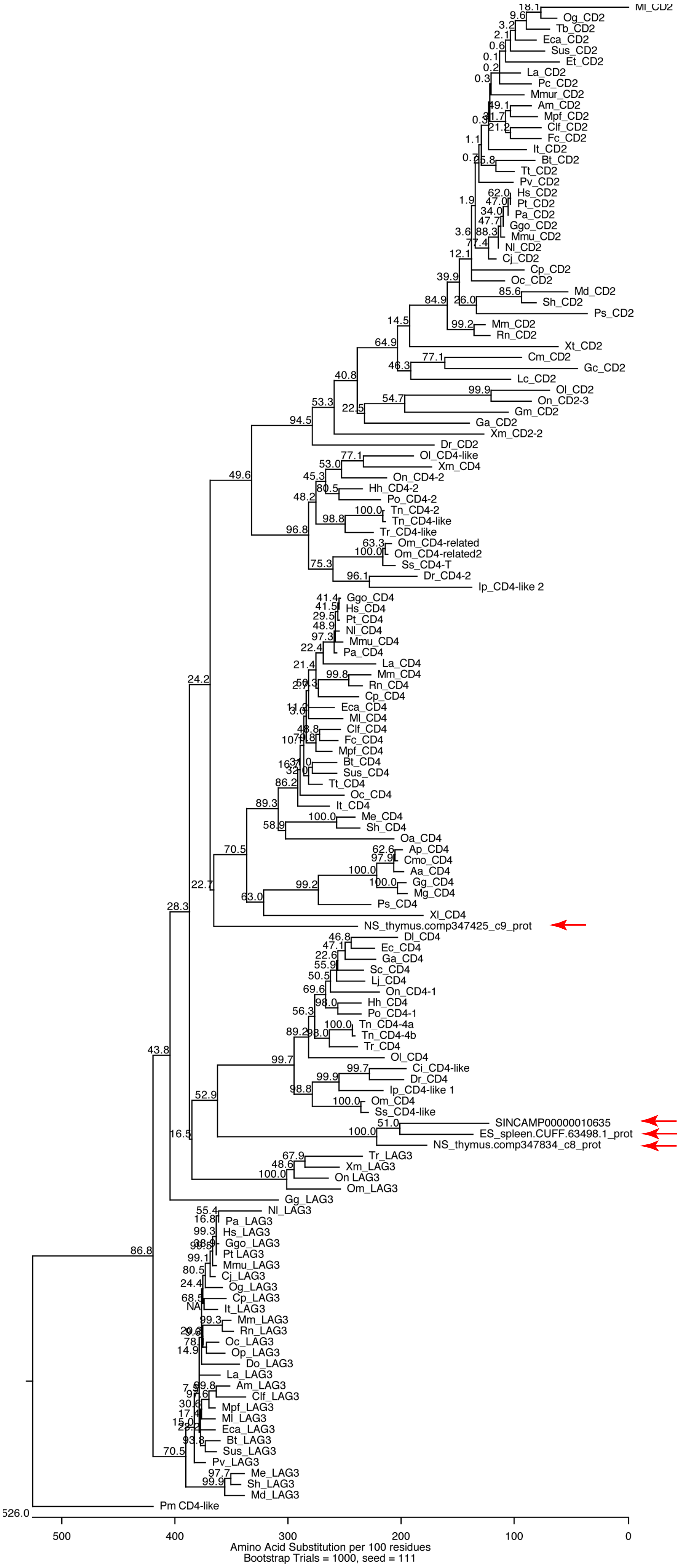
Signal peptide   IgSF domains   Transmembrane helix

d

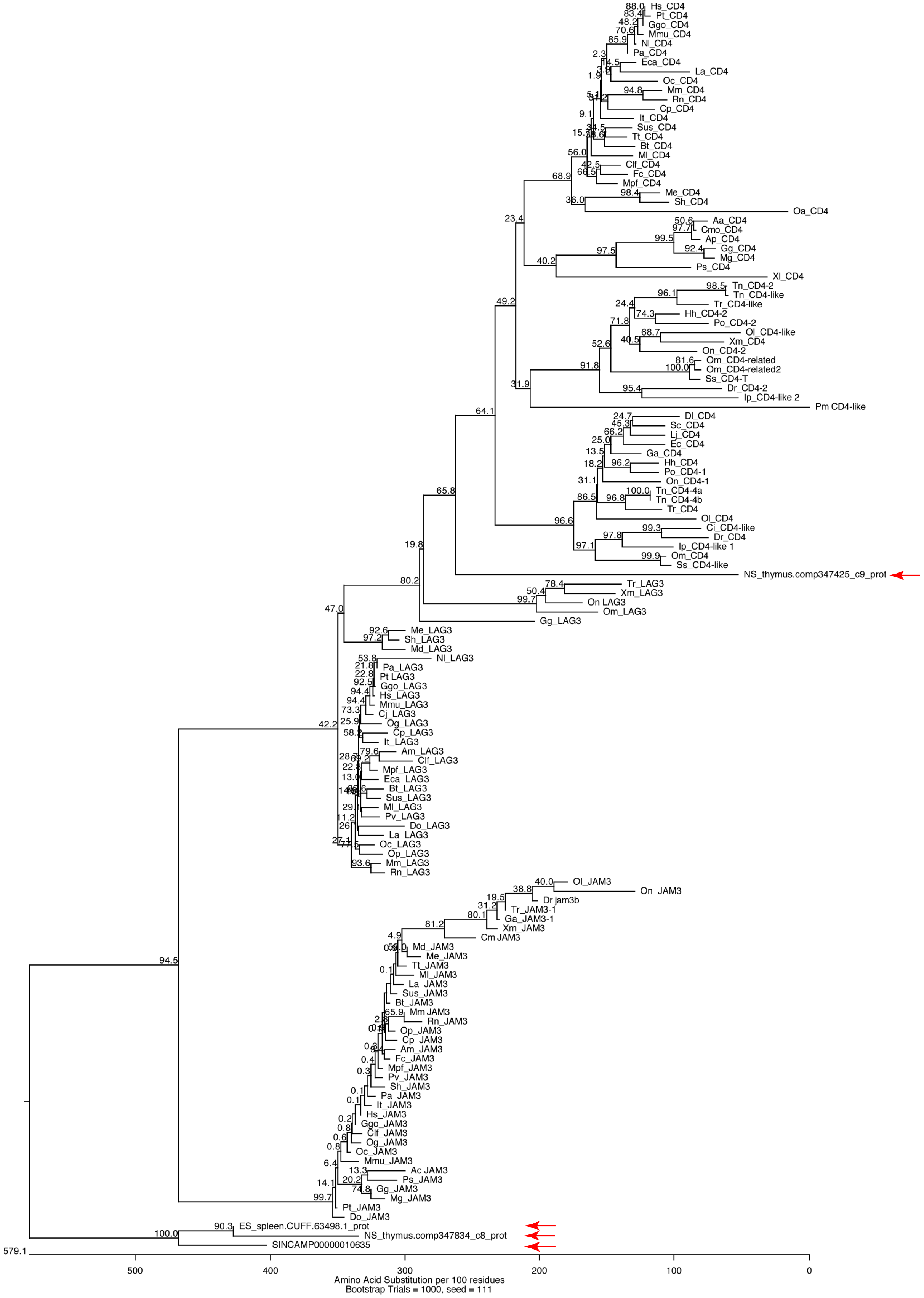
Gene	Transcript ID	Length (bp)	TPM	Ratio (x/LCK)
CD4-R/LAG3	NS_thymus KC707916	3,673	0.866 x 10^{-3}	3.4 x 10^{-6}
CD8A	NS_thymus KC707917	3,799	21.98	85 x 10^{-3}
LCK	NS_thymus KC707918	6,445	258.41	
CD4-R/LAG3	NS_spleen KC707916	1,894	2.83 x 10^{-3}	82 x 10^{-6}
CD8A	NS_spleen KC707917	2,622	2.64	77 x 10^{-3}
LCK	NS_spleen KC707918	4,186	34.41	

Supplementary Figure XI.8 | Protein signatures of LCK, and CD8 and CD4 co-receptors.

a, Sequence motifs of the CD8α and CD4 interaction domain of LCK in tetrapods and teleosts; the sequences in LCK proteins of cartilaginous fishes are similar (left panel). Signature of the interaction domain in CD8α proteins of tetrapods, teleosts and cartilaginous fishes; note that the CXH motif in teleosts is functionally equivalent to the CXC motif in tetrapods; the sequences in cartilagionous fishes conform to the fish signature. Rp, *Rhinobatus productus*; Gc, *Ginglymostoma cirratum*; Cm, *Callorhinchus milii*. **b**, The LCK interaction motif is identical in tetrapods and teleosts. **c**, Protein sequence alignment of human CD4 and the closest relative identified in databases of cartilaginous fishes. Note the identical domain structure of the extracellular IgSF and transmembrane domains; however, the intra-cytoplasmic domain is much longer in the proteins of cartilaginous fishes and lacks the characteristic CXC motif required for interaction with LCK. **d**, Expression analysis of relevant genes in the transcriptomes of thymus and spleen the nurse shark (NS) *G. cirratum*. Note the vastly different expression levels of *CD8A* and the *CD4*-related genes.


Ml_CD2
Og_CD2
Tb_CD2
Eca_CD2
Sus_CD2
Et_CD2
La_CD2
Pc_CD2
Mmur_CD2
Am_CD2
Mpf_CD2
Clf_CD2
Fc_CD2
It_CD2
Bt_CD2
Tt_CD2
Pv_CD2
Hs_CD2
Pt_CD2
Pa_CD2
Ggo_CD2
Mmu_CD2
Nl_CD2
Cj_CD2
Cp_CD2
Oc_CD2
Md_CD2
Sh_CD2
Ps_CD2
Mm_CD2
Rn_CD2
Xt_CD2
Cm_CD2
Gc_CD2
Lc_CD2
Ol_CD2
On_CD2-3
Gm_CD2
Ga_CD2
Xm_CD2-2
Dr_CD2
Ol_CD4-like
Xm_CD4
On_CD4-2
Hh_CD4-2
Po_CD4-2
Tn_CD4-2
Tn_CD4-like
Tr_CD4-like
Om_CD4-related
Om_CD4-related2
Ss_CD4-T
Dr_CD4-2
Ip_CD4-like 2
Ggo_CD4
Hs_CD4
Pt_CD4
Nl_CD4
Mmu_CD4
Pa_CD4
La_CD4
Mm_CD4
Rn_CD4
Cp_CD4
Eca_CD4
Ml_CD4
Clf_CD4
Fc_CD4
Mpf_CD4
Bt_CD4
Sus_CD4
Tt_CD4
Oc_CD4
It_CD4
Me_CD4
Sh_CD4
Oa_CD4
Ap_CD4
Cmo_CD4
Aa_CD4
Gg_CD4
Mg_CD4
Ps_CD4
Xl_CD4
NS_thymus.comp347425_c9_prot
Dl_CD4
Ec_CD4
Ga_CD4
Sc_CD4
Lj_CD4
On_CD4-1
Hh_CD4
Po_CD4-1
Tn_CD4-4a
Tn_CD4-4b
Tr_CD4
Ol_CD4
Ci_CD4-like
Dr_CD4
Ip_CD4-like 1
Om_CD4
Ss_CD4-like
SINCAMP00000010635
ES_spleen.CUFF.63498.1_prot
NS_thymus.comp347834_c8_prot
Tr_LAG3
Xm_LAG3
On_LAG3
Om_LAG3
Gg_LAG3
Nl_LAG3
Pa_LAG3
Hs_LAG3
Ggo_LAG3
Pt LAG3
Mmu_LAG3
Cj_LAG3
Og_LAG3
Cp_LAG3
It_LAG3
Mm_LAG3
Rn_LAG3
Oc_LAG3
Op_LAG3
Do_LAG3
La_LAG3
Am_LAG3
Clf_LAG3
Mpf_LAG3
Ml_LAG3
Eca_LAG3
Bt_LAG3
Sus_LAG3
Pv_LAG3
Me_LAG3
Sh_LAG3
Md_LAG3
Pm CD4-like
526.0
500
400
300
200
100
0
Amino Acid Substitution per 100 residues
Bootstrap Trials = 1000, seed = 111

Supplementary Figure XI.9 | Phylogenetic analysis of CD4-like proteins of vertebrates. In this analysis, LAG3 and CD2 sequences were included. The sequences used for this analysis, including species identifiers are listed in Supplementary Table XI.12. Arrows indicate the positions of relevant proteins of cartilaginous fishes (see Supplementary Table XI.13 for details).



Supplementary Figure XI.10 | Phylogenetic analysis of CD4-like proteins of vertebrates. Similar to Supplementary Figure XI.9, except for inclusion of LAG3 and JAM3 sequences.

References

Alioto TS, Ngai J. 2006. The repertoire of olfactory C family G protein-coupled receptors in zebrafish: candidate chemosensory receptors for amino acids. *BMC Genomics.* 7:309.

Castresana J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol Biol Evol.* 17:540-552.

Grus WE, Zhang J. 2009. Origin of the genetic components of the vomeronasal system in the common ancestor of all extant vertebrates. *Mol Biol Evol.* 26:407-419.

Heimberg AM, Cowper-Sal-lari R, Semon M, Donoghue PC, Peterson KJ. 2010. microRNAs reveal the interrelationships of hagfish, lampreys, and gnathostomes and the nature of the ancestral vertebrate. *Proc Natl Acad Sci U S A.* 107:19379-19383.

Katoh K, Misawa K, Kuma K, Miyata T. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* 30:3059-3066.

Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, Jones SJ, Marra MA. 2009. Circos: an information aesthetic for comparative genomics. *Genome research.* 19:1639-1645.

Niimura Y. 2009. On the origin and evolution of vertebrate olfactory receptor genes: comparative genome analysis among 23 chordate species. *Genome Biol Evol.* 1:34-44.

Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S. 2011. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol Biol Evol.* 28:2731-2739.

Supplementary Table I.1 | Assembly statistics for *Callorhinchus milii*-6.1.3

	Contigs	**Scaffolds**
Total numbers	67,425	21,208
Total bases	936,942,150	974,487,278
N50 bases	46,577	4,521,921
N50 number	5,393	60
Maximum bases	631,173	18,507,834

Supplementary Table I.2 | Types of interspersed repetitive sequences in the *C. milii* genome

Type of sequence	Number of copies	Total length (Mb)	Percentage of genome
LINE	~374,000	117.7	12.6%
- L2	~246,000	79.8	8.5%
- CR1	~124,000	37.0	4.0%
- Others	~4,000	0.9	0.1%
SINE	~753,000	123.2	13.1%
- tRNA-L2	~550,000	94.0	10.0%
- Deu	~180,000	27.4	2.9%
- Others	~23,000	1.8	0.2%
DNA transposon	~55,000	5.9	0.6%
LTR element	~24,000	3.5	0.4%
Others/Unclassified	~112,000	21.7	2.3%
Total (non-redundant)	-	264.0	28.2%

Supplementary Table I.3 | Proportion of callable genome, number of SNPs and Ti/Tv for the consecutive filters applied in the SNP calling.
These statistics were obtained for the CDS too.

Scaffolds with length > 50kb						
	bp mapable and not found in repeats or gaps	**without INDELS**	**w/o INDELs AND coverage between 5× and 15x**	**w/o INDELs AND coverage between 5× and 15× AND w/o DUPs**	**w/o INDELs AND coverage between 5× and 15× AND with interval of 3bp between SNPs**	**CDS**
Number Scaffolds	657	657	657	649	649	515
Callable bp	618,687,353	247,295,157	172,861,486	172,721,512	172,697,509	3,042,472
Percentage Genome	63.48	25.37	17.73	17.72	17.72	0.31
Number of SNPs	1,519,279	530,021	429,272	426,093	402,090	4,099
Ti/tv	1.618	1.802	1.817	1.828	1.868	3.027

Supplementary Table I.4 | Comparison of heterozygosity value with other fish species.

	# Kb in the callable portion of the assembly	# SNPs	Heterozygosity (per bp)
Medaka[1]	480,300	16,448,457	N/A
Atlantic cod[2]	500,614	1,047,875	0.00209
C. milii	172,698	402,090	0.00233

[1] Kasahara et al. 2007

[2] Star et al. 2011

Supplementary Table I.5 | Summary of duplication analysis in the *C. milii* genome.

	All segments	**Duplications (SD>3, >10windows)**
# Windows	661,750	3,117
# Segments	17,711	179
# Windows / segment	37	17
Mean coverage (SD)	9.24 (28.91)[a]	46.28 (40.00)[b]
# bp in duplications[c]	__	5,466,045
% genome[d]	__	0.56
Size range (Kb)	__	12–297
Median size (Kb)	__	31

[a] Calculated from all windows used for segmentation and before smoothing

[b] Calculated from only those windows defined as duplicated after the segmentation step

[c] Not corrected for copy-number

[d] Relative to the total length of the assembly

Supplementary Table I.6 | List of *C. milii* scaffolds with the number of duplications and the percentage of their sequence considered as duplicated.

Scaffold	Scaffold length (bp)	Number of duplications	Number of base pairs in duplications	Percentage of base pairs in duplications
scaffold_1	18,507,834	1	16,736	0.09
scaffold_7	13,485,325	1	21,409	0.16
scaffold_19	9,386,079	1	34,492	0.37
scaffold_33	6,506,902	1	29,764	0.46
scaffold_41	5,334,932	1	93,782	1.76
scaffold_64	4,326,226	1	145,814	3.37
scaffold_76	3,867,380	1	79,759	2.06
scaffold_86	3,567,987	1	31,118	0.87
scaffold_92	3,337,817	1	24,567	0.74
scaffold_109	2,628,759	1	78,201	2.97
scaffold_111	2,535,139	1	62,208	2.45
scaffold_122	2,111,713	1	17,613	0.83
scaffold_180	1,072,290	1	19,825	1.85
scaffold_199	890,228	1	97,682	10.97
scaffold_220	694,188	7	552,592	79.60
scaffold_221	690,336	1	16,398	2.38
scaffold_232	601,629	1	94,168	15.65
scaffold_285	404,662	1	17,057	4.22
scaffold_290	389,319	6	205,434	52.77
scaffold_296	371,076	1	16,767	4.52
scaffold_325	287,866	4	235,431	81.78
scaffold_371	182,215	2	178,030	97.70
scaffold_375	175,851	1	175,850	100.00
scaffold_396	147,573	2	141,802	96.09
scaffold_408	129,947	2	129,771	99.86
scaffold_430	112,960	1	30,493	26.99
scaffold_449	99,835	2	91,371	91.52
scaffold_491	83,118	1	19,492	23.45
scaffold_494	81,585	1	37,572	46.05
scaffold_535	69,203	1	69,202	100.00
scaffold_567	63,060	1	21,168	33.57
scaffold_608	55,744	1	55,743	100.00
scaffold_620	54,570	1	54,569	100.00
scaffold_635	52,906	1	52,905	100.00
scaffold_650	50,623	1	50,576	99.91
scaffold_659	49,577	1	49,288	99.42
scaffold_663	49,074	1	49,073	100.00
scaffold_670	47,875	1	47,814	99.87
scaffold_694	44,534	1	44,533	100.00
scaffold_695	44,304	1	19,285	43.53

scaffold_729	39,587	1	24,698	62.39
scaffold_746	37,355	1	37,354	100.00
scaffold_759	35,906	1	35,905	100.00
scaffold_793	32,335	1	31,801	98.35
scaffold_804	31,283	1	19,285	61.65
scaffold_814	30,750	1	30,749	100.00
scaffold_829	30,080	1	30,079	100.00
scaffold_848	29,199	1	29,002	99.33
scaffold_858	28,618	1	28,617	100.00
scaffold_859	28,572	1	28,482	99.69
scaffold_903	26,462	1	24,140	91.23
scaffold_915	25,850	1	25,849	100.00
scaffold_923	25,455	1	25,454	100.00
scaffold_930	24,950	1	24,949	100.00
scaffold_937	24,818	1	24,817	100.00
scaffold_942	24,653	1	24,652	100.00
scaffold_951	24,376	1	24,054	98.68
scaffold_958	23,756	1	23,569	99.21
scaffold_960	23,720	1	23,719	100.00
scaffold_971	23,461	1	23,168	98.75
scaffold_982	23,323	1	23,322	100.00
scaffold_980	23,185	1	23,184	100.00
scaffold_993	22,861	1	22,624	98.96
scaffold_1001	22,762	1	22,761	100.00
scaffold_1002	22,484	1	22,483	100.00
scaffold_1022	21,827	1	21,826	100.00
scaffold_1025	21,773	1	21,772	100.00
scaffold_1032	21,503	1	21,502	100.00
scaffold_1055	21,082	1	21,081	100.00
scaffold_1058	21,023	1	20,674	98.34
scaffold_1054	21,005	1	21,004	100.00
scaffold_1077	20,780	1	20,150	96.97
scaffold_1073	20,654	1	20,653	100.00
scaffold_1080	20,502	1	20,501	100.00
scaffold_1092	20,296	1	20,295	100.00
scaffold_1088	20,193	1	20,192	100.00
scaffold_1121	19,627	1	19,626	99.99
scaffold_1118	19,600	1	19,329	98.62
scaffold_1124	19,597	1	19,452	99.26
scaffold_1126	19,499	1	19,498	99.99
scaffold_1149	19,476	1	19,475	99.99
scaffold_1132	19,373	1	19,372	99.99
scaffold_1138	19,314	1	19,173	99.27
scaffold_1141	19,283	1	19,282	99.99
scaffold_1142	19,279	1	18,927	98.17
scaffold_1158	18,968	1	18,860	99.43

scaffold_1155	18,935	1	18,663	98.56
scaffold_1161	18,855	1	18,854	99.99
scaffold_1170	18,674	1	18,673	99.99
scaffold_1166	18,648	1	18,647	99.99
scaffold_1169	18,597	1	18,596	99.99
scaffold_1179	18,541	1	18,476	99.65
scaffold_1175	18,535	1	18,534	99.99
scaffold_1195	18,298	1	18,297	99.99
scaffold_1210	18,067	1	18,066	99.99
scaffold_1207	18,032	1	18,031	99.99
scaffold_1227	17,904	1	17,903	99.99
scaffold_1218	17,771	1	17,770	99.99
scaffold_1226	17,679	1	17,678	99.99
scaffold_1224	17,673	1	17,672	99.99
scaffold_1230	17,643	1	17,596	99.73
scaffold_1228	17,573	1	17,321	98.57
scaffold_1234	17,566	1	17,565	99.99
scaffold_1231	17,540	1	15,555	88.68
scaffold_1246	17,428	1	17,077	97.99
scaffold_1260	17,368	1	17,367	99.99
scaffold_1251	17,287	1	17,286	99.99
scaffold_1253	17,264	1	17,172	99.47
scaffold_1258	17,236	1	17,235	99.99
scaffold_1254	17,226	1	17,128	99.43
scaffold_1256	17,183	1	17,182	99.99
scaffold_1268	17,038	1	17,037	99.99
scaffold_1275	16,985	1	16,984	99.99
scaffold_1288	16,859	1	16,858	99.99
scaffold_1290	16,841	1	16,788	99.69
scaffold_1292	16,750	1	16,749	99.99
scaffold_1293	16,722	1	16,721	99.99
scaffold_1309	16,689	1	16,688	99.99
scaffold_1311	16,513	1	14,089	85.32
scaffold_1327	16,203	1	16,202	99.99
scaffold_1334	16,035	1	16,034	99.99
scaffold_1346	15,965	1	15,867	99.39
scaffold_1344	15,908	1	15,838	99.56
scaffold_1352	15,832	1	15,831	99.99
scaffold_1354	15,811	1	15,508	98.08
scaffold_1377	15,678	1	15,677	99.99
scaffold_1392	15,640	1	15,589	99.67
scaffold_1367	15,619	1	15,421	98.73
scaffold_1382	15,566	1	15,565	99.99
scaffold_1380	15,508	1	15,507	99.99
scaffold_1381	15,499	1	15,498	99.99
scaffold_1386	15,436	1	15,435	99.99

scaffold_1387	15,423	1	15,422	99.99
scaffold_1389	15,412	1	15,411	99.99
scaffold_1401	15,308	1	14,719	96.15
scaffold_1436	14,998	1	14,997	99.99
scaffold_1453	14,716	1	14,715	99.99
scaffold_1470	14,577	1	14,517	99.59
scaffold_1479	14,495	1	14,494	99.99
scaffold_1481	14,424	1	14,423	99.99
scaffold_1522	14,205	1	14,204	99.99
scaffold_1511	14,092	1	14,091	99.99
scaffold_1531	13,923	1	13,922	99.99
scaffold_1558	13,805	1	13,804	99.99
scaffold_1549	13,800	1	13,735	99.53
scaffold_1559	13,704	1	13,703	99.99
scaffold_1568	13,696	1	13,695	99.99
scaffold_1564	13,651	1	13,650	99.99
scaffold_1570	13,605	1	13,604	99.99
scaffold_1571	13,601	1	13,498	99.24
scaffold_1583	13,497	1	13,496	99.99
scaffold_1605	13,493	1	13,492	99.99
scaffold_1590	13,433	1	13,432	99.99
scaffold_1609	13,292	1	13,291	99.99
scaffold_1612	13,266	1	12,879	97.08
scaffold_1628	13,232	1	13,231	99.99
scaffold_1641	13,118	1	13,117	99.99
scaffold_1664	12,910	1	12,909	99.99
scaffold_1693	12,725	1	12,724	99.99
scaffold_1786	12,196	1	12,195	99.99
scaffold_1854	11,731	1	11,730	99.99

Supplementary Table I.7 | *C. milii* genes covered by segmental duplication.
Coverage_by_duplication: indicates whether the duplication covers fully or partially the sequence of the gene; Exons_in_duplication: number of exons fully or partially covered by the duplication; Exons_in_duplication: number of exons not covered by the duplication; Exons_details: all exons are listed indicating if they are not covered (e), partially covered (p) or, otherwise, totally covered by a duplication.

Scaffold	**Start**	**End**	**Gene name**	**Gene_type**	**Coverage_by_duplication**	**Exons_in_duplication**	**Exons_out_duplication**	**Exons_details**
#220	672127	676616	ADPRH	protein_coding	full	6	0	1, 2, 3, 4, 5, 6
#92	201179	203570	B0YN61_CALMI	protein_coding	partial	1	0	1 (p)
#92	204375	206781	B0YN62_CALMI	protein_coding	full	1	0	1
#92	207555	210039	B0YN63_CALMI	protein_coding	full	1	0	1
#92	210885	213291	B0YN64_CALMI	protein_coding	full	1	0	1
#92	214059	216543	B0YN65_CALMI	protein_coding	full	1	0	1
#92	217778	220262	B0YN66_CALMI	protein_coding	full	1	0	1
#92	220422	223917	B0YN67_CALMI	protein_coding	full	1	0	1
#92	224026	228083	B0YN68_CALMI	protein_coding	partial	1	0	1 (p)
#494	4924	5797	CCR1	protein_coding	full	2	0	1, 2
#1149	8326	9028	CXCR7	protein_coding	full	2	0	1, 2
#1559	9532	13182	DMBT1	protein_coding	full	4	0	1, 2, 3, 4
#1559	9625	13071	DMBT1	protein_coding	full	5	0	1, 2, 3, 4, 5
#1559	9559	13101	DMBT1	protein_coding	full	6	0	1, 2, 3, 4, 5, 6
#903	23623	26124	DMBT1	protein_coding	partial	1	3	1, 2 (e), 3 (e), 4 (e)

#1170	10396	16904	FAM55C	protein_coding	full	6	0	1, 2, 3, 4, 5, 6
#86	7649	12120	FASN	protein_coding	full	3	0	1, 2, 3
#371	41119	41992	GPR139	protein_coding	full	3	0	1, 2, 3
#1001	14311	17056	GPR139	protein_coding	full	2	0	1, 2
#635	44412	45312	GPR139	protein_coding	full	2	0	1, 2
#1158	13988	14876	GPR139	protein_coding	full	2	0	1, 2
#111	2431461	2432367	GPR139	protein_coding	full	1	0	1
#1389	211	12846	GPR139	protein_coding	full	4	0	1, 2, 3, 4
#1032	5975	8140	GPR139	protein_coding	full	2	0	1, 2
#958	7006	7870	GPR139	protein_coding	full	2	0	1, 2
#430	84068	84989	GPR139	protein_coding	full	2	0	1, 2
#858	12550	13650	GPR142	protein_coding	full	2	0	1, 2
#111	2408683	2409544	GPR142	protein_coding	full	2	0	1, 2
#325	203656	204794	GPR142	protein_coding	full	2	0	1, 2
#1141	6373	7243	GPR142	protein_coding	full	2	0	1, 2
#1207	7471	8305	GPR142	protein_coding	full	2	0	1, 2
#1401	9009	9822	GPR142	protein_coding	full	3	0	1, 2, 3
#1132	12016	13976	GPR142	protein_coding	full	2	0	1, 2
#1628	4838	5709	GPR142	protein_coding	full	2	0	1, 2
#1309	2534	3194	GPR142	protein_coding	full	2	0	1, 2
#1275	8056	8920	GPR142	protein_coding	full	2	0	1, 2
#694	21691	22432	GPR142	protein_coding	full	2	0	1, 2
#1354	7081	9821	GPR142	protein_coding	full	2	0	1, 2
#1195	16219	17079	GPR142	protein_coding	full	3	0	1, 2, 3
#220	222267	222990	GPR142	protein_coding	full	2	0	1, 2
#325	128435	129314	GPR142	protein_coding	full	3	0	1, 2, 3
#1234	9714	11416	GPR142	protein_coding	full	3	0	1, 2, 3
#1124	17952	19047	GPR142	protein_coding	full	2	0	1, 2

#111	2420309	2420978	GPR142	protein_coding	full	2	0	1, 2
#1092	14325	15541	GPR142	protein_coding	full	2	0	1, 2
#290	317836	318628	GPR142	protein_coding	full	2	0	1, 2
#951	10933	12756	GPR142	protein_coding	full	3	0	1, 2, 3
#285	3401	7529	HEBP2	protein_coding	full	4	0	1, 2, 3, 4
#180	548607	559469	ICAM2	protein_coding	full	4	0	1, 2, 3, 4
#1564	5935	12547	IGHM	protein_coding	full	10	0	1, 2, 3, 4, 5, 6, 7, 8, 9, 10
#1564	6043	12544	IGHM	protein_coding	full	4	0	1, 2, 3, 4
#1549	4394	9667	IGHM	protein_coding	full	3	0	1, 2, 3
#1570	3922	11892	IGHM	protein_coding	full	7	0	1, 2, 3, 4, 5, 6, 7
#1166	11038	17541	ighm	protein_coding	full	4	0	1, 2, 3, 4
#1564	2789	12544	IGHM	protein_coding	full	6	0	1, 2, 3, 4, 5, 6
#1570	3832	7976	IGHM	protein_coding	full	4	0	1, 2, 3, 4
#1549	2632	9667	IGHM	protein_coding	full	10	0	1, 2, 3, 4, 5, 6, 7, 8, 9, 10
#220	684076	693002	ighm	protein_coding	full	7	0	1, 2, 3, 4, 5, 6, 7
#1166	8118	11287	ighm	protein_coding	full	2	0	1, 2
#1253	10427	15902	IGHM	protein_coding	full	4	0	1, 2, 3, 4
#1344	930	1272	ighv13-2	protein_coding	full	1	0	1
#937	11847	12195	ighv13-2	protein_coding	full	1	0	1
#220	24649	31824	IGHV3-23	protein_coding	full	6	0	1, 2, 3, 4, 5, 6
#1436	6451	8539	IGHV3-30	protein_coding	full	2	0	1, 2
#793	20184	22836	IGHV3-53	protein_coding	full	2	0	1, 2
#1025	18449	20781	IGHV3-53	protein_coding	full	4	0	1, 2, 3, 4
#793	20258	21103	IGHV3-53	protein_coding	full	2	0	1, 2
#221	688465	689076	IGLC7	protein_coding	full	1	0	1
#1258	5853	6337	IGLV3-21	protein_coding	full	2	0	1, 2
#1258	10494	12420	IGLV3-21	protein_coding	full	2	0	1, 2
#1258	1421	1881	IGLV3-9	protein_coding	full	2	0	1, 2
#19	733875	7340338	IL8	protein_coding	full	4	0	1, 2, 3, 4

	9							
#19	7328782	7332034	IL8	protein_coding	partial	4	0	1, 2, 3, 4 (p)
#19	7355923	7358930	IL8	protein_coding	full	4	0	1, 2, 3, 4
#19	7328782	7340335	IL8	protein_coding	partial	4	0	1, 2, 3, 4 (p)
#19	7346619	7349535	IL8	protein_coding	full	4	0	1, 2, 3, 4
#942	265	5905	MOG	protein_coding	full	6	0	1, 2, 3, 4, 5, 6
#64	37635	71612	MUC16	protein_coding	full	15	0	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15
#220	38315	38990	NAT8L	protein_coding	full	1	0	1
#41	2437535	2452299	NPL	protein_coding	partial	1	11	1 (e), 2 (e), 3 (e), 4 (e), 5 (e), 6 (e), 7 (e), 8 (e), 9 (e), 10 (e), 11 (e), 12 (p)
#1693	1528	6511	olfcd3	protein_coding	full	9	0	1, 2, 3, 4, 5, 6, 7, 8, 9
#1169	9484	16375	olfcd3	protein_coding	full	7	0	1, 2, 3, 4, 5, 6, 7
#1612	207	4347	olfcd3	protein_coding	full	7	0	1, 2, 3, 4, 5, 6, 7
#1693	1322	6520	olfcd3	protein_coding	full	9	0	1, 2, 3, 4, 5, 6, 7, 8, 9
#86	2077	5233	PTCHD3	protein_coding	full	2	0	1, 2
#232	241596	247043	SIGLEC15	protein_coding	full	6	0	1, 2, 3, 4, 5, 6
#232	220280	224506	SIGLEC15	protein_coding	full	3	0	1, 2, 3
#232	220280	233979	SIGLEC15	protein_coding	full	4	0	1, 2, 3, 4
#232	165481	166798	SIGLEC15	protein_coding	full	2	0	1, 2
#232	241596	247055	SIGLEC15	protein_coding	full	4	0	1, 2, 3, 4
#232	250791	254708	SIGLEC15	protein_coding	full	2	0	1, 2
#232	243245	247057	SIGLEC15	protein_coding	full	4	0	1, 2, 3, 4
#232	220280	230895	SIGLEC15	protein_coding	full	4	0	1, 2, 3, 4
#814	3602	16893	SRCRB4D	protein_coding	full	12	0	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12

#1470	0	1618	SSC5D	protein_coding	full	3	0	1, 2, 3
#7	2595095	2599855	TCEB1	protein_coding	partial	2	1	1 (e), 2, 3
#1210	16500	17461	TRAV12-2	protein_coding	full	2	0	1, 2
#1210	15106	17460	TRAV12-2	protein_coding	full	2	0	1, 2
#220	591952	592327	TRAV12-3	protein_coding	full	1	0	1
#408	44622	50908	TRAV13-2	protein_coding	full	2	0	1, 2
#408	43240	44961	TRAV13-2	protein_coding	full	2	0	1, 2
#375	142411	143235	TRAV17	protein_coding	full	2	0	1, 2
#793	10036	10786	TRAV17	protein_coding	full	2	0	1, 2
#663	43790	44362	TRAV17	protein_coding	full	2	0	1, 2
#663	42136	44407	TRAV17	protein_coding	full	2	0	1, 2
#449	84188	84551	TRAV18	protein_coding	full	1	0	1
#375	119556	125234	TRAV18	protein_coding	full	4	0	1, 2, 3, 4
#1392	4953	14078	TRAV2	protein_coding	full	4	0	1, 2, 3, 4
#449	77919	82579	TRAV20	protein_coding	full	4	0	1, 2, 3, 4
#220	620191	623576	TRAV20	protein_coding	full	2	0	1, 2
#220	643480	648963	TRAV20	protein_coding	full	4	0	1, 2, 3, 4
#408	105456	108636	TRAV20	protein_coding	full	2	0	1, 2
#1126	12013	13125	TRAV20	protein_coding	full	2	0	1, 2
#220	645455	650475	TRAV20	protein_coding	full	3	0	1, 2, 3
#1126	14045	19303	TRAV20	protein_coding	full	4	0	1, 2, 3, 4
#1175	13295	13646	TRAV20	protein_coding	full	1	0	1
#449	97382	99310	TRAV20	protein_coding	full	3	0	1, 2, 3
#1126	5110	5452	TRAV21	protein_coding	full	1	0	1
#1382	5851	6349	TRAV23DV6	protein_coding	full	1	0	1
#375	129142	130985	TRAV3	protein_coding	full	2	0	1, 2
#408	85892	99466	TRAV36DV7	protein_coding	full	3	0	1, 2, 3
#408	96999	101327	TRAV36DV7	protein_coding	full	4	0	1, 2, 3, 4
#1155	9892	15306	TRAV38-1	protein_coding	full	4	0	1, 2, 3, 4

#1268	11774	16862	TRAV38-1	protein_coding	full	3	0	1, 2, 3
#793	3389	5490	TRAV38-1	protein_coding	full	2	0	1, 2
#1352	10750	14315	TRAV38-1	protein_coding	full	3	0	1, 2, 3
#1158	1173	1524	TRAV38-1	protein_coding	full	1	0	1
#1142	1556	9873	TRAV38-1	protein_coding	full	3	0	1, 2, 3
#1268	11789	12596	TRAV38-1	protein_coding	full	2	0	1, 2
#1155	11807	15327	TRAV38-1	protein_coding	full	4	0	1, 2, 3, 4
#1210	22	2088	TRAV38-1	protein_coding	full	3	0	1, 2, 3
#1346	8424	9325	TRAV38-2DV8	protein_coding	full	2	0	1, 2
#759	13577	13925	TRAV38-2DV8	protein_coding	full	1	0	1
#396	30851	34098	TRAV38-2DV8	protein_coding	full	3	0	1, 2, 3
#659	19105	21243	TRAV38-2DV8	protein_coding	full	3	0	1, 2, 3
#759	15385	15673	TRAV38-2DV8	protein_coding	full	1	0	1
#449	67671	68282	TRAV41	protein_coding	full	2	0	1, 2
#408	51937	52318	TRAV6	protein_coding	full	1	0	1
#396	66263	70992	TRAV8-4	protein_coding	full	3	0	1, 2, 3
#659	39636	39984	TRAV9-1	protein_coding	full	1	0	1
#793	31393	31750	TRAV9-1	protein_coding	full	1	0	1
#608	930	2249	TRBV10-1	protein_coding	full	2	0	1, 2
#1226	6285	7765	TRBV6-1	protein_coding	full	3	0	1, 2, 3
#608	9256	10133	TRBV6-1	protein_coding	full	1	0	1
#670	1947	3570	TRBV6-1	protein_coding	full	2	0	1, 2
#1226	10790	11576	TRBV6-1	protein_coding	full	2	0	1, 2
#1226	6692	8294	TRBV6-1	protein_coding	full	2	0	1, 2
#1226	1632	3926	TRBV6-1	protein_coding	full	3	0	1, 2, 3
#1226	11204	13957	TRBV6-1	protein_coding	full	3	0	1, 2, 3
#1226	5004	5361	TRBV6-4	protein_coding	full	1	0	1
#220	360282	360964	TRDV1	protein_coding	full	2	0	1, 2
#1382	13592	15075	TRDV1	protein_coding	full	3	0	1, 2, 3
#1268	779	5059	TRDV1	protein_coding	full	4	0	1, 2, 3, 4

#923	19979	23528	TRDV1	protein_coding	full	3	0	1, 2, 3
#1210	6346	9973	TRDV1	protein_coding	full	4	0	1, 2, 3, 4
#1155	5460	6159	TRDV1	protein_coding	full	2	0	1, 2
#220	258410	258774	TRDV3	protein_coding	full	2	0	1, 2
#1479	1090	5316	TRDV3	protein_coding	full	4	0	1, 2, 3, 4
#220	625615	634380	TRDV3	protein_coding	full	4	0	1, 2, 3, 4
#220	638168	638614	TRDV3	protein_coding	full	1	0	1
#859	11020	22072	TRIM60	protein_coding	full	8	0	1, 2, 3, 4, 5, 6, 7, 8
#325	273098	274897	TRIM69	protein_coding	full	4	0	1, 2, 3, 4
#7	2563387	2587471	UBE2W	protein_coding	partial	1	4	1, 2 (e), 3 (e), 4 (e), 5 (e)
#7	2557675	2587471	UBE2W	protein_coding	partial	1	3	1, 2 (e), 3 (e), 4 (e)
#180	554351	560657	VCAM1	protein_coding	full	6	0	1, 2, 3, 4, 5, 6
#1346	1264	1594	-	protein_coding	full	1	0	1
#396	88598	95985	-	protein_coding	partial	5	2	1, 2, 3, 4, 5 (p), 6 (e), 7 (e)
#220	198469	199300	-	protein_coding	full	2	0	1, 2
#1293	14321	14981	-	protein_coding	full	2	0	1, 2
#620	27315	35885	-	protein_coding	full	3	0	1, 2, 3
#122	1965004	1980576	-	protein_coding	partial	3	10	1 (e), 2 (e), 3 (e), 4 (e), 5 (e), 6 (e), 7 (e), 8 (e), 9 (e), 10 (e), 11, 12, 13
#122	1966443	1980576	-	protein_coding	partial	1	9	1 (e), 2 (e), 3 (e), 4 (e), 5 (e), 6 (e), 7 (e), 8 (e), 9 (e), 10
#1	13250226	13252702	-	protein_coding	full	4	0	1, 2, 3, 4
#375	161230	168071	-	protein_coding	full	6	0	1, 2, 3, 4, 5, 6
#993	1128	17517	-	protein_coding	full	6	0	1, 2, 3, 4, 5, 6
#670	43921	45025	-	protein_coding	full	2	0	1, 2

#109	251579	266923	-	protein_coding	full	13	0	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13
#221	674694	677638	-	protein_coding	full	3	0	1, 2, 3
#76	2449263	2450283	-	protein_coding	full	2	0	1, 2
#109	252613	266923	-	protein_coding	full	13	0	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13
#1854	51	8869	-	protein_coding	full	7	0	1, 2, 3, 4, 5, 6, 7
#396	131318	132153	-	protein_coding	full	2	0	1, 2
#1002	15640	18183	-	protein_coding	full	2	0	1, 2
#111	2455580	2455642	-	miRNA	full	1	0	1
#1054	14688	14743	-	miRNA	full	1	0	1
#76	2389721	2390829	-	protein_coding	full	2	0	1, 2
#915	21716	25822	-	protein_coding	full	4	0	1, 2, 3, 4
#122	1915365	1985819	-	protein_coding	partial	0	23	1 (e), 2 (e), 3 (e), 4 (e), 5 (e), 6 (e), 7 (e), 8 (e), 9 (e), 10 (e), 11 (e), 12 (e), 13 (e), 14 (e), 15 (e), 16 (e), 17 (e), 18 (e), 19 (e), 20 (e), 21 (e), 22 (e), 23 (e)
#1058	1963	19650	-	protein_coding	full	10	0	1, 2, 3, 4, 5, 6, 7, 8, 9, 10
#930	21359	22238	-	protein_coding	full	2	0	1, 2
#923	2291	12944	-	protein_coding	full	4	0	1, 2, 3, 4
#1	13237011	13239958	-	protein_coding	full	4	0	1, 2, 3, 4
#1251	4786	11122	-	protein_coding	full	4	0	1, 2, 3, 4
#608	6879	7562	-	protein_coding	full	1	0	1
#180	561935	569841	-	protein_coding	partial	4	2	1 (e), 2 (e), 3, 4, 5, 6

#76	2441286	2442206	-	protein_coding	full	2	0	1, 2
#408	15163	19479	-	protein_coding	full	6	0	1, 2, 3, 4, 5, 6
#937	16375	18581	-	protein_coding	full	2	0	1, 2
#220	407749	412341	-	protein_coding	full	5	0	1, 2, 3, 4, 5
#980	8976	23135	-	protein_coding	full	11	0	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11
#1511	6929	8705	-	protein_coding	full	5	0	1, 2, 3, 4, 5
#76	2395399	2396082	-	protein_coding	full	2	0	1, 2
#1141	18565	19150	-	protein_coding	full	2	0	1, 2
#76	2431399	2433044	-	protein_coding	full	2	0	1, 2
#1121	12363	17138	-	protein_coding	full	5	0	1, 2, 3, 4, 5
#180	564407	569841	-	protein_coding	partial	3	2	1 (e), 2 (e), 3, 4, 5
#659	48016	48714	-	protein_coding	full	2	0	1, 2
#1382	3257	3939	-	protein_coding	full	2	0	1, 2
#1251	4750	12310	-	protein_coding	full	4	0	1, 2, 3, 4
#1377	9812	12068	-	protein_coding	full	3	0	1, 2, 3
#1	13245312	13252334	-	protein_coding	full	4	0	1, 2, 3, 4
#76	2398865	2399757	-	protein_coding	full	2	0	1, 2
#220	522276	530193	-	protein_coding	partial	8	1	1 (e), 2, 3, 4, 5, 6, 7, 8, 9
#759	33101	33407	-	protein_coding	full	1	0	1
#220	502359	503235	-	protein_coding	full	2	0	1, 2
#746	7738	25875	-	protein_coding	full	6	0	1, 2, 3, 4, 5, 6
#1080	16837	18013	-	protein_coding	full	3	0	1, 2, 3
#41	2344597	2345563	-	protein_coding	full	1	0	1
#608	3039	5404	-	protein_coding	full	3	0	1, 2, 3

#759	20731	21761	-	protein_coding	full	2	0	1, 2
#1380	14176	15507	-	protein_coding	full	2	0	1, 2
#122	1915365	1983598	-	protein_coding	partial	0	21	1 (e), 2 (e), 3 (e), 4 (e), 5 (e), 6 (e), 7 (e), 8 (e), 9 (e), 10 (e), 11 (e), 12 (e), 13 (e), 14 (e), 15 (e), 16 (e), 17 (e), 18 (e), 19 (e), 20 (e), 21 (e)
#1531	2334	5388	-	protein_coding	full	2	0	1, 2
#1055	4022	5636	-	protein_coding	full	2	0	1, 2
#109	252326	266796	-	protein_coding	full	13	0	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13
#1481	8850	14414	-	protein_coding	full	8	0	1, 2, 3, 4, 5, 6, 7, 8
#1170	2358	4795	-	protein_coding	full	2	0	1, 2
#1121	4	16987	-	protein_coding	full	11	0	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11
#290	23908	25885	-	protein_coding	full	5	0	1, 2, 3, 4, 5
#1179	268	18453	-	protein_coding	full	22	0	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22
#449	48516	48975	-	pseudogene	full	3	0	1, 2, 3
#76	2417498	2418429	-	protein_coding	full	2	0	1, 2
#290	157092	157950	-	protein_coding	full	2	0	1, 2
#408	118509	119397	-	protein_coding	full	2	0	1, 2
#1055	3863	18619	-	protein_coding	full	5	0	1, 2, 3, 4, 5
#1346	13151	14098	-	protein_coding	full	2	0	1, 2
#1288	31	8455	-	protein_coding	full	6	0	1, 2, 3, 4, 5, 6
#937	13499	14208	-	protein_coding	full	2	0	1, 2
#1142	11290	12306	-	protein_coding	full	2	0	1, 2
#76	238464	2385606	-	protein_coding	full	3	0	1, 2, 3

	7							
#325	218511	219294	-	protein_coding	full	2	0	1, 2
#1	13245312	13249571	-	protein_coding	full	4	0	1, 2, 3, 4
#76	2409496	2410153	-	protein_coding	full	2	0	1, 2
#221	674694	678102	-	protein_coding	full	3	0	1, 2, 3
#396	137642	138100	-	protein_coding	full	2	0	1, 2
#76	2413953	2414837	-	protein_coding	full	2	0	1, 2
#1531	2432	5396	-	protein_coding	full	3	0	1, 2, 3
#41	2362833	2363997	-	protein_coding	full	1	0	1
#76	2374704	2375714	-	protein_coding	full	2	0	1, 2
#1088	18618	19536	-	protein_coding	full	2	0	1, 2
#1058	1968	9056	-	protein_coding	full	6	0	1, 2, 3, 4, 5, 6
#1664	8180	9140	-	protein_coding	full	2	0	1, 2
#1254	3926	4868	-	protein_coding	full	1	0	1
#670	42753	45025	-	protein_coding	full	3	0	1, 2, 3
#221	653935	678102	-	protein_coding	partial	1	2	1, 2 (e), 3 (e)
#109	246299	266796	-	protein_coding	partial	12	1	1 (e), 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13
#923	6343	7045	-	protein_coding	full	2	0	1, 2
#1073	11438	14611	-	protein_coding	full	3	0	1, 2, 3
#109	325479	516273	-	protein_coding	partial	1	44	1 (e), 2 (e), 3 (e), 4 (e), 5 (e), 6 (e), 7 (e), 8 (e), 9 (e), 10 (e), 11 (e), 12 (e), 13 (e), 14 (e), 15 (e), 16 (e), 17 (e), 18 (e), 19 (e), 20 (e), 21 (e), 22 (e), 23

								(e), 24 (e), 25 (e), 26 (e), 27 (e), 28 (e), 29 (e), 30 (e), 31 (e), 32 (e), 33 (e), 34 (e), 35 (e), 36 (e), 37 (e), 38 (e), 39 (e), 40 (e), 41 (e), 42 (e), 43 (e), 44 (e), 45 (p)
#746	11603	25323	-	protein_coding	full	4	0	1, 2, 3, 4
#1058	8	19288	-	protein_coding	full	6	0	1, 2, 3, 4, 5, 6
#670	5166	7862	-	protein_coding	full	3	0	1, 2, 3
#1121	1	17002	-	protein_coding	full	7	0	1, 2, 3, 4, 5, 6, 7
#1641	3560	12221	-	protein_coding	full	9	0	1, 2, 3, 4, 5, 6, 7, 8, 9
#375	134584	135525	-	protein_coding	full	2	0	1, 2

Supplementary Table II.1 | The average GC content and standard deviation of vertebrate genomes

These were computed over non-overlapping 3-kb windows.

Genome	Average GC content (%)	Standard deviation (%)
C. milii	42.34	4.47
Human	40.91	6.44
Chicken	41.29	5.64
Lizard	39.90	3.53
X. tropicalis	40.06	4.12
Zebrafish	36.60	4.54
Medaka	40.14	3.60
Stickleback	44.53	3.84
Fugu	45.46	4.23

Supplementary Table II.2 | Tests of GC compositional homogeneity in *C. milii* genome
Analysis was carried out for scaffolds > 4 Mb. Kruskal-Wallis tests were carried out on groups of 20-kb windows in 300-kb regions.

Scaffold	Length	Average GC (%)	Standard deviation	Kruskal-Wallis Test P-value	Heterogeneous? (P < 0.05)
1	18,507,834	40.23	3.26	~0	Yes
2	17,031,706	41.04	4.57	~0	Yes
3	16,461,339	39.08	3.65	~0	Yes
4	16,433,419	40.09	3.69	2.33E-15	Yes
5	15,003,573	41.25	3.94	~0	Yes
6	13,911,802	41.06	3.60	~0	Yes
7	13,485,325	38.95	3.39	~0	Yes
8	12,728,579	41.05	3.80	~0	Yes
9	12,614,711	39.78	3.65	~0	Yes
10	11,909,161	42.89	3.95	~0	Yes
11	11,573,033	40.27	3.57	6.02E-05	Yes
12	11,530,728	40.22	3.80	~0	Yes
13	10,529,718	41.21	3.51	7.59E-11	Yes
14	9,996,151	39.25	3.06	7.24E-09	Yes
15	9,908,685	43.68	3.54	~0	Yes
16	9,849,519	40.76	3.98	~0	Yes
17	9,843,783	41.14	3.84	~0	Yes
18	9,630,976	42.56	3.70	~0	Yes
19	9,386,079	41.23	3.55	1.21E-08	Yes
20	9,228,010	40.00	3.58	0.012267	Yes
21	9,046,933	41.66	3.81	0.00172	Yes
22	8,947,343	40.79	3.70	0.005888	Yes
23	8,803,159	41.23	3.82	3.64E-13	Yes
24	8,332,076	41.23	3.43	0.063744	No
25	8,056,818	40.18	3.66	0.083315	No
26	7,764,900	40.51	3.75	0.150473	No
27	7,641,004	40.18	3.60	1.08E-06	Yes
28	7,535,631	40.83	3.92	~0	Yes
29	7,318,916	40.09	4.21	~0	Yes
30	6,964,137	41.83	3.68	~0	Yes
31	6,938,060	41.03	3.54	0.004917	Yes
32	6,832,876	42.05	3.19	1.24E-14	Yes
33	6,506,902	42.38	3.80	4.89E-13	Yes
34	6,463,314	41.27	3.67	0.000351	Yes
35	6,429,418	41.01	3.73	2.64E-06	Yes
36	6,301,995	44.15	3.80	2.23E-10	Yes
37	5,873,764	44.65	4.13	0.004666	Yes
38	5,469,490	40.79	3.83	4.36E-10	Yes
39	5,419,759	38.80	2.97	0.00365	Yes
40	5,379,800	40.41	3.54	4.71E-05	Yes

41	5,334,932	41.25	3.41	~0	Yes
42	5,298,526	40.96	3.33	9.31E-13	Yes
43	5,220,276	40.47	3.43	8.12E-07	Yes
44	5,213,970	44.60	3.76	6.21E-08	Yes
45	5,104,675	41.30	4.49	~0	Yes
46	5,087,354	42.25	4.09	~0	Yes
47	5,048,387	45.00	4.65	0.049969	Yes
48	4,929,752	41.51	3.55	2.48E-11	Yes
49	4,897,056	43.16	3.55	4.63E-09	Yes
50	4,875,586	40.33	3.41	0.03009	Yes
51	4,861,936	40.60	3.71	0.391851	No
52	4,822,577	42.21	3.34	4E-07	Yes
53	4,771,929	42.10	3.59	3.16E-10	Yes
54	4,751,830	40.48	3.97	5.93E-07	Yes
55	4,608,585	40.69	3.55	0.048503	Yes
56	4,596,003	40.46	3.65	0.004326	Yes
57	4,591,479	41.38	3.41	6.02E-08	Yes
58	4,569,021	45.19	3.80	0.012924	Yes
59	4,564,766	43.31	3.46	7.67E-08	Yes
60	4,521,921	41.47	3.70	0.337046	No
61	4,508,736	42.64	4.38	~0	Yes
62	4,467,497	41.23	3.52	0.079865	No
63	4,339,253	43.21	3.91	3.36E-10	Yes
64	4,326,226	45.52	3.96	0.011714	Yes
65	4,283,084	42.37	3.01	0.002259	Yes
66	4,241,200	41.42	4.07	4.01E-09	Yes
67	4,186,129	40.40	3.40	0.36032	No
68	4,157,550	40.17	3.58	0.068569	No
69	4,147,603	40.56	3.48	0.03411	Yes
70	4,034,929	45.59	4.28	0.074826	No
Total					**61 ‘Yes’**

Supplementary Table II.3 | Summary of isochores in *C. milii* genome
(the largest 70 scaffolds; total 532.0 Mb)

Scaffold	No. of isochores	No. of isochores >300kb	Isochores >300 kb and homogeneous		
			Number	Avg. length (kb)	Percentage of the scaffold
1	59	14	7	1,698	64%
2	31	14	12	1,150	81%
3	62	17	15	928	85%
4	123	15	5	779	24%
5	43	15	10	1,083	72%
6	6	5	2	4,044	58%
7	39	11	7	1,282	67%
8	7	6	4	2,931	92%
9	66	14	6	745	35%
10	13	9	4	2,194	74%
11	79	13	5	739	32%
12	41	15	10	696	60%
13	53	17	3	505	14%
14	61	13	6	634	38%
15	17	9	3	907	27%
16	38	11	5	1,205	61%
17	38	13	8	705	57%
18	42	5	3	2,212	69%
19	10	6	2	1743	37%
20	81	8	2	512	11%
21	97	9	3	484	16%
22	105	10	4	399	18%
23	60	12	6	494	34%
24	75	9	4	600	29%
25	72	5	2	374	9%
26	74	8	1	414	5%
27	53	8	4	776	41%
28	22	9	9	790	94%
29	130	3	3	985	40%
30	20	7	4	962	55%
31	87	2	0	0	0%
32	34	5	2	1,517	44%
33	14	5	3	1,507	69%
34	5	4	2	1,646	51%
35	23	9	4	822	51%
36	74	5	4	393	25%

37	12	5	3	964	49%
38	23	4	3	994	55%
39	29	5	2	792	29%
40	67	4	1	552	10%
41	8	3	2	2,054	77%
42	29	6	3	786	45%
43	29	5	2	686	26%
44	10	4	1	3,798	73%
45	49	4	2	1,386	54%
46	22	5	4	848	67%
47	27	6	6	499	59%
48	46	4	2	750	30%
49	12	5	4	712	58%
50	74	4	3	503	31%
51	87	2	2	401	16%
52	38	6	3	468	29%
53	45	7	3	489	31%
54	43	5	2	809	34%
55	49	6	2	404	18%
56	49	6	2	491	21%
57	29	5	2	740	32%
58	16	2	1	3,357	73%
59	27	7	3	795	52%
60	40	4	0	0	0%
61	2	2	2	2,253	100%
62	2	1	0	0	0%
63	4	4	2	1,106	51%
64	76	1	1	2,955	68%
65	16	4	1	612	14%
66	49	4	3	677	48%
67	32	4	2	515	25%
68	216	1	1	390	9%
69	61	3	1	708	17%
70	41	1	1	792	20%
Total	**3,213**	**479**	**246**	**993.24**	**45.9%**

Supplementary Table II.4 | Descriptive statistics of isochores in lizard, *C. milii* and stickleback.

		Isochore families		
		L2	**H1**	**H2**
No. of isochores	*C. milii*	195	49	2
	Lizard	467	3	-
	Stickleback	-	60	11
Average GC content (%)	*C. milii*	39.9	43.0	44.4
	Lizard	39.4	42.5	-
	Stickleback	-	43.4	45.8
Average size of isochore (kb)	*C. milii*	917.5	1,243.9	2,233.5
	Lizard	496.3	1,137.0	-
	Stickleback	-	3,588.1	1,554.5
Percentage of isochoric sequence	*C. milii*	73.2	25.0	1.8
	Lizard	98.5	1.5	-
	Stickleback	-	92.6	7.4

Supplementary Table II.5 | Gene density of isochores in *C. milii*, lizard and stickleback, (measured as percentage of isochoric sequence that falls within protein-coding genes).

	Isochoric sequence (kb)			Genic sequence (kb)			Gene density (%)		
Species \ isochore family	**L2**	**H1**	**H2**	**L2**	**H1**	**H2**	**L2**	**H1**	**H2**
C. milii	178,920	60,951	4,467	85,703	31,006	1,527	47.9	50.9	34.2
Lizard	231,762	3,411	-	53,190	982	-	23.0	28.8	-
Stickleback	-	215,286	17,100	-	92,786	7,032	-	43.1	41.1

Supplementary Table III.7 | miRNA families present in elephant shark, mouse and human but lost in teleost fishes.

Sl.number	microRNA family	Function	Zebrafish	Stickleback	Fugu
1	mir-28	unknown	Absent	Absent	Absent
2	mir-32	anti-viral	Absent	Absent	Absent
3	mir-33	cholesterol metabolism	Absent	Present in Genome	Present in Genome
4	mir-147	inflammation response	Absent	Absent	Absent
5	mir-149	unknown	Absent	Absent	Absent
6	mir-150	B-cell development	Absent	Absent	Absent
7	mir-154	unknown	Absent	Absent	Absent
8	mir-191	cancer associated	Absent	Absent	Absent
9	mir-290	autophagy related	Absent	Absent	Absent
10	mir-378	cell survival	Absent	Absent	Absent
11	mir-425	unknown	Absent	Absent	Absent
12	mir-449	testis development	Absent	Absent	Absent
13	mir-467	unknown	Absent	Absent	Absent
14	mir-506	unknown	Absent	Absent	Absent
15	mir-551	unknown	Absent	Absent	Absent
16	mir-653	unknown	Absent	Absent	Absent
17	mir-676	unknown	Absent	Absent	Absent
18	mir-744	TGFBeta-1 regulation	Absent	Absent	Absent
19	mir-764	osteoblast differentiation	Absent	Absent	Absent
20	mir-873	unknown	Absent	Absent	Absent
21	mir-875	unknown	Absent	Absent	Absent
22	mir-1247	unknown	Absent	Absent	Absent

Supplementary Table IV.2 | Overlap of human orthologous gCNEs with known functional elements in the human genome

Type of functional elements	Percentage of observed overlaps (gCNEs)	Percentage of expected overlaps (random noncoding regions)	Fold enrichment	p-value
p300-binding sites (Visel et al. 2009)	3.94%	0.33%	12x	< 1e-200
Transcriptional enhancers (Visel et al. 2007)	3.01%	0.14%	22x	< 1e-200

Supplementary Table IV.3 | GeneOntology enrichment (molecular function section) of human genes associated with pan-gnathostome CNEs.
The top 20 GO terms with p-value < 0.05 are shown.

No.	GO term	Genes in genome	Genes associated with pan-gnathostome CNEs		p-value
			Observed	Expected	
1	sequence-specific DNA binding	721	99	22.49	1.70E-27
2	sequence-specific DNA binding transcription factor activity	1024	117	31.94	8.60E-25
3	RNA polymerase II distal enhancer sequence-specific DNA binding transcription factor activity	99	17	3.09	1.00E-08
4	chromatin binding	260	25	8.11	5.70E-07
5	enhancer sequence-specific DNA binding	17	7	0.53	3.20E-06
6	RNA polymerase II core promoter proximal region sequence-specific DNA binding transcription factor activity involved in positive regulation of transcription	42	9	1.31	4.60E-06
7	protein heterodimerization activity	302	25	9.42	9.60E-06
8	double-stranded DNA binding	155	15	4.83	9.00E-05
9	RNA polymerase II core promoter proximal region sequence-specific DNA binding	11	4	0.34	2.60E-04
10	RNA polymerase II transcription coactivator activity	11	4	0.34	2.60E-04
11	DNA binding, bending	56	8	1.75	3.20E-04
12	protein homodimerization activity	512	31	15.97	3.50E-04
13	steroid hormone receptor activity	57	8	1.78	3.70E-04
14	RNA polymerase II core promoter sequence-specific DNA binding transcription factor activity	6	3	0.19	5.60E-04
15	DNA binding	2390	167	74.55	6.00E-04
16	nucleic acid binding transcription factor activity	1026	119	32	6.70E-04
17	RNA polymerase II core promoter sequence-specific DNA binding	14	4	0.44	7.30E-04
18	transcription corepressor activity	169	16	5.27	1.39E-03
19	core promoter proximal region	20	7	0.62	2.16E-03

	sequence-specific DNA binding				
20	glutamate binding	9	3	0.28	2.20E-03

Supplementary Table IV.4 | GeneOntology enrichment (biological process section) of human genes associated with pan-gnathostome CNEs.
The top 20 GO terms with p-value < 0.05 are shown.

No.	GO term	Genes in genome	Genes associated with pan-gnathostome CNEs		p-value
			Observed	Expected	
1	positive regulation of transcription from RNA polymerase II promoter	618	66	19.8	3.40E-17
2	negative regulation of transcription from RNA polymerase II promoter	432	52	13.84	1.60E-15
3	negative regulation of neuron differentiation	54	15	1.73	4.30E-10
4	positive regulation of neuron differentiation	63	15	2.02	9.30E-10
5	dorsal spinal cord development	21	10	0.67	6.30E-07
6	forebrain development	267	44	8.56	1.70E-06
7	neural tube closure	66	12	2.11	3.30E-06
8	embryonic hindlimb morphogenesis	31	8	0.99	4.30E-06
9	metanephros development	76	15	2.44	6.60E-06
10	positive regulation of osteoblast differentiation	43	9	1.38	7.10E-06
11	camera-type eye development	232	31	7.43	7.10E-06
12	negative regulation of smoothened signaling pathway	17	6	0.54	9.60E-06
13	regulation of transcription, DNA-dependent	2996	191	96	1.30E-05
14	trigeminal nerve development	6	4	0.19	1.50E-05
15	lens induction in camera-type eye	6	4	0.19	1.50E-05
16	pituitary gland development	44	11	1.41	1.60E-05
17	embryonic digestive tract morphogenesis	19	6	0.61	2.00E-05
18	positive regulation of cartilage development	12	5	0.38	2.20E-05
19	anterior/posterior axis specification	41	7	1.31	3.40E-05
20	neuron fate commitment	56	16	1.79	4.60E-05

Supplementary Table IV.5 | Top 20 human genes with the highest number of pan-gnathostome CNEs

No.	Gene	No. of pan-gnathostome CNEs	Name	Description
1	ENSG00000108001	42	EBF3	early B-cell factor 3
2	ENSG00000143032	28	BARHL2	BarH-like homeobox 2
3	ENSG00000121297	21	TSHZ3	teashirt zinc finger homeobox 3
4	ENSG00000148655	20	C10orf11	chromosome 10 open reading frame 11
5	ENSG00000114861	19	FOXP1	forkhead box P1
6	ENSG00000175745	17	NR2F1	nuclear receptor subfamily 2, group F, member 1
7	ENSG00000205148	16	AC016251.1	Uncharacterized protein
8	ENSG00000165659	16	DACH1	dachshund homolog 1 (Drosophila)
9	ENSG00000185594	16	SPATA8	spermatogenesis associated 8
10	ENSG00000170549	15	IRX1	iroquois homeobox 1
11	ENSG00000153234	15	NR4A2	nuclear receptor subfamily 4, group A, member 2
12	ENSG00000091656	15	ZFHX4	zinc finger homeobox 4
13	ENSG00000169946	15	ZFPM2	zinc finger protein, multitype 2
14	ENSG00000176842	14	IRX5	iroquois homeobox 5
15	ENSG00000181355	14	OFCC1	orofacial cleft 1 candidate 1
16	ENSG00000167081	14	PBX3	pre-B-cell leukemia homeobox 3
17	ENSG00000256463	14	SALL3	sal-like 3 (Drosophila)
18	ENSG00000164651	14	SP8	Sp8 transcription factor
19	ENSG00000170430	13	MGMT	O-6-methylguanine-DNA methyltransferase
20	ENSG00000075891	13	PAX2	paired box 2

Supplementary Table IV.6 | Top 20 *C. milii* genes with the highest number of gCNEs lost in teleosts

No.	Gene	No. of associated gCNEs	No. of gCNEs lost in teleosts	Name	Description	Is gene detected in teleosts?
1	SINCAMG00000005569	94	91	ZNF608	zinc finger protein 608 [Homo sapiens]	detected
2	SINCAMG00000008354	118	81	LPHN2	latrophilin 2 [Homo sapiens]	detected
3	SINCAMG00000007237	179	76	nr2f2	nuclear receptor subfamily 2, group F, member 2 [Danio rerio]	detected
4	SINCAMG00000016163	104	71	RBFOX1	RNA binding protein, fox-1 homolog (C. elegans) 1 [Homo sapiens]	detected
5	SINCAMG00000012383	131	65	MGMT	O-6-methylguanine-DNA methyltransferase [Xenopus tropicalis]	detected
6	SINCAMG00000013188	121	62	ZEB2	zinc finger E-box binding homeobox 2 [Homo sapiens]	detected
7	SINCAMG00000010227	93	57	EBF1	early B-cell factor 1 [Homo sapiens]	not_detected
8	SINCAMG00000012676	126	55	NR2F1	nuclear receptor subfamily 2, group F, member 1 [Homo sapiens]	detected
9	SINCAMG00000015450	127	53	IRX1	iroquois homeobox 1 [Xenopus tropicalis]	detected
10	SINCAMG00000000087	61	52	TLE1	transducin-like enhancer of split 1 (E(sp1) homolog, Drosophila) [Homo sapiens]	detected
11	SINCAMG00000015913	72	51	MECOM	MDS1 and EVI1 complex locus [Homo sapiens]	detected
12	SINCAMG00000004616	117	49	irx3a	iroquois homeobox protein 3a [Danio rerio]	detected
13	SINCAMG00000006831	110	49	ZFHX4	zinc finger homeobox 4 [Xenopus tropicalis]	detected

14	SINCAMG00000004420	90	47	DACH1	dachshund homolog 1 (Drosophila) [Homo sapiens]	detected
15	SINCAMG00000003683	67	46	C16orf7	chromosome 16 open reading frame 7 [Homo sapiens]	detected
16	SINCAMG00000013602	62	46	EFNA5	ephrin-A5 [Xenopus tropicalis]	detected
17	SINCAMG00000003023	72	45	WWOX	WW domain-containing oxidoreductase [Gallus gallus]	detected
18	SINCAMG00000003179	82	44	ESRRG	estrogen-related receptor gamma [Xenopus tropicalis]	detected
19	SINCAMG00000004484	100	42	TSHZ3	teashirt zinc finger homeobox 3 [Homo sapiens]	detected
20	SINCAMG00000004614	59	42	FTO	FTO isoform 1 Fragment [Meleagris gallopavo]	detected

Supplementary Table IV.7 | Top 20 *C. milii* genes with the highest number of gCNEs lost in tetrapods

No.	Gene	No. of associated gCNEs	No. of gCNEs lost in tetrapods	Name	Description	Is gene detected in tetrapods?
1	SINCAMG00000007832	43	19	CASZ1	castor zinc finger 1 [Homo sapiens]	detected
2	SINCAMG00000017450	38	11	-	uncharacterized protein	not_detected
3	SINCAMG00000008909	18	9	FNDC7	fibronectin type III domain containing 7 [Homo sapiens]	not_detected
4	SINCAMG00000015450	127	8	IRX1	iroquois homeobox 1 [Xenopus tropicalis]	detected
5	SINCAMG00000016474	16	8	CHST3-like	carbohydrate sulfotransferase 3-like [Xenopus tropicalis]	detected
6	SINCAMG00000011290	22	7	HERC1	hect domain and RCC1-like domain 1 [Bos taurus]	not_detected
7	SINCAMG00000006768	42	6	FBRSL1	fibrosin-like 1 [Homo sapiens]	detected
8	SINCAMG00000017287	11	6	-	uncharacterized protein	not_detected
9	SINCAMG00000007237	179	5	nr2f2	nuclear receptor subfamily 2, group F, member 2 [Danio rerio]	detected
10	SINCAMG00000011467	30	5	ATP8B2	ATPase, class I, type 8B, member 2 [Homo sapiens]	detected
11	SINCAMG00000006880	27	5	B0YN98_CALMI	Protocadherin nu1 [Source:UniProtKB/TrEMBL;Acc:B0YN85]	detected
12	SINCAMG00000016478	22	5	CUEDC2	CUE domain containing 2 [Homo sapiens]	detected
13	SINCAMG00000009005	16	5	PAX1	paired box 1 [Xenopus tropicalis]	detected
14	SINCAMG00000004468	16	5	-	si:dkey-22o22.2 [Danio rerio]	detected
15	SINCAMG00000005907	11	5	IGFALS	insulin-like growth factor-binding protein	detected

					complex acid labile subunit [Bos taurus]	
16	SINCAMG00000004617	128	4	IRX5	iroquois homeobox 5 [Xenopus tropicalis]	not_detected
17	SINCAMG00000006156	71	4	C15orf41	chromosome 15 open reading frame 41 [Homo sapiens]	detected
18	SINCAMG00000001847	41	4	LRBA	LPS-responsive vesicle trafficking, beach and anchor containing [Homo sapiens]	detected
19	SINCAMG00000006817	37	4	HNF4G	hepatocyte nuclear factor 4, gamma [Homo sapiens]	detected
20	SINCAMG00000000509	29	4	dnah9l	dynein, axonemal, heavy polypeptide 9 like [Danio rerio]	detected

Supplementary Table V.1 | Evaluation of alternate topologies for the 13-chordate dataset using CONSEL.

Three selected topologies were evaluated using several tests as implemented in the program CONSEL. A concatenated peptide alignment of 699 one-to-one core orthologs was used for topology testing. The tests used are: approximately unbiased test (AU), bootstrap probability (NP, BP), Bayesian posterior probability (PP), Kishino-Hasegawa test (KH), Shimodaira-Hasegawa test (SH), weighted KH test (wKH) and weighted SH test (wSH). *p*-values derived from these tests for the various topologies are shown (higher is better). All the tests ranked "Chondrichthyes sister to bony vertebrates" (C,(Tel,(Lc,Tet))) as the most likely topology. This congruence in the top-ranked topology between different tests is a strong support for this topology. This topology was also inferred by phylogenetic analysis using Maximum likelihood and Bayesian inference methods.

Rank	Topology	AU	NP	BP	PP	KH	SH	wKH	wSH
1	(C,(A,(Lc,Tet)))	0.972	0.970	0.967	1.000	0.971	0.994	0.971	0.997
2	(A,(C,(Lc,Tet)))	0.028	0.030	0.033	1e-73	0.029	0.122	0.029	0.073
3	(((C,Tel),Lc),Tet)	4e-37	6e-15	0	0	0	0	0	0
4	(((C,Lc),Tel),Tet)	9e-43	8e-16	0	0	0	0	0	0
5	((C,(Lc,Tel)),Tet)	1e-74	2e-20	0	0	0	0	0	0

C, Chondrichthyes (*C. milii*); Tet, tetrapods; Tel, teleosts; Lc, coelacanth; A, Actinopterygii (stickleback and zebrafish)

Supplementary Table VI.1 | Relative rate tests of *C. milii* versus the other vertebrates using the 13-chordate protein dataset and sea lamprey as the outgroup.

The ‘slow’ column shows the ‘significantly’ slower evolving ingroup species based on P-value. The ‘identical’ and ‘divergent’ columns refer to sites where the amino acid residue is the same or different in all 3 sequences, respectively. ‘Ingroup1-specific’column refers to sites where ingroup 2 and outgroup share the same amino acid but not ingroup 1. The same applies for ‘ingroup2-specific’ and ‘outgroup-specific’.

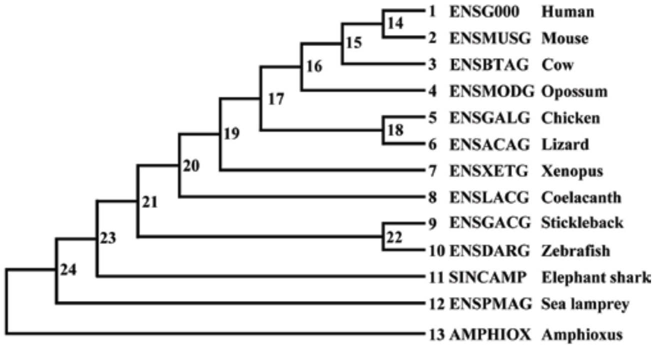
Ingroup1	Ingroup2	Outgroup	Genes	Identical	Divergent	Ingroup1 specific	Ingroup2 specific	Outgroup specific	Slow	CHI^2_test_statistic	P-value
Human	*C. milii*	Lamprey	699	141977	24814	15046	14154	41912	*C. milii*	27.25	1.79E-07
Mouse	*C. milii*	Lamprey	699	141468	25497	15555	14312	41071	*C. milii*	51.73	6.37E-13
Cow	*C. milii*	Lamprey	699	141608	24992	15413	14275	41612	*C. milii*	43.62	3.98E-11
Opossum	*C. milii*	Lamprey	699	141503	24928	15520	13726	42226	*C. milii*	110.05	9.57E-26
Chicken	*C. milii*	Lamprey	699	142197	24221	14822	13901	42753	*C. milii*	29.53	5.50E-08
Lizard	*C. milii*	Lamprey	699	141508	25149	15515	13999	41732	*C. milii*	77.87	1.10E-18
Xenopus	*C. milii*	Lamprey	699	139525	26285	17485	13982	40604	*C. milii*	389.96	8.43E-87
Coelacanth	*C. milii*	Lamprey	699	141714	22849	15306	13069	44958	*C. milii*	176.36	3.02E-40
Stickleback	*C. milii*	Lamprey	699	138142	28292	18880	15179	37409	*C. milii*	402.17	1.86E-89
Zebrafish	*C. milii*	Lamprey	699	139910	27181	17111	14934	38764	*C. milii*	147.90	5.00E-34

Supplementary Table VI.2 | Two-Cluster tests using Lintre for the 13-chordate dataset

The various pairwise comparisons analyzed and the respective topologies are shown to indicate the node and tip numbering. The fast or slow-evolving cluster at each node is denoted by '>' or '<'. Nodes involving *C. milii* comparisons are highlighted. Z, Z-statistics; CP, confidence probablilty.

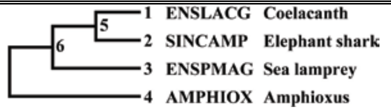
node	L		R	delta	s.e.	Z	CP	height	s.e.	bA	bB	bC
22	10	<	9	0.030014	0.001245	24.105183	99.96%	0.121251	0.000691	0.106244	0.136257	0.33579
18	6	>	5	0.014978	0.000938	15.968775	99.96%	0.092957	0.000567	0.100446	0.085468	0.275167
14	1	<	2	0.012579	0.000572	21.999766	99.96%	0.038799	0.00032	0.03251	0.045088	0.312768
15	14	<	3	0.000128	0.000534	0.239538	18.20%	0.038358	0.000279	0.038294	0.038422	0.340758
16	15	>	4	0.001146	0.000778	1.472955	85.84%	0.071154	0.00043	0.071727	0.070581	0.333678
17	16	>	18	0.004283	0.000851	5.035696	99.96%	0.108321	0.00052	0.110462	0.106179	0.348506
19	17	<	7	0.028137	0.001208	23.288071	99.96%	0.146	0.000703	0.131931	0.160069	0.3532
20	19	>	8	0.007565	0.001257	6.018724	99.96%	0.1496	0.000693	0.153382	0.145818	0.373758
21	20	<	22	0.033394	0.00139	24.032621	99.96%	0.191358	0.000807	0.174661	0.208055	0.447187
23	21	>	11	0.025452	0.001699	14.98295	99.96%	0.169774	0.000734	0.1825	0.157048	0.590516
24	23	<	12	0.088112	0.003553	24.79951	99.96%	0.317119	0.001398	0.273063	0.361175	0.634104

Q=3313.091139



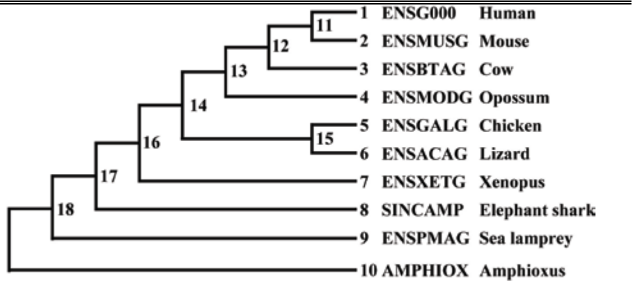
node	L		R	delta	s.e.	Z	CP	height	s.e.	bA	bB	bC
5	1	>	2	0.030782	0.00217	14.185666	99.96%	0.150801	0.00082	0.166192	0.13541	0.612204
6	5	<	3	0.095271	0.003762	25.321526	99.96%	0.312935	0.001459	0.2653	0.36057	0.634841

Q=877.093257



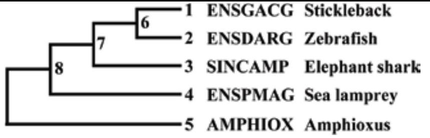
node	L		R	delta	s.e.	Z	CP	height	s.e.	bA	bB	bC
15	6	>	5	0.015188	0.000969	15.669913	99.96%	0.092961	0.000567	0.100555	0.085367	0.282403
11	1	<	2	0.013615	0.00058	23.483639	99.96%	0.0388	0.00032	0.031992	0.045607	0.31179
12	11	>	3	0.000035	0.000548	0.063594	4.78%	0.03836	0.000279	0.038377	0.038342	0.351337
13	12	>	4	0.001159	0.000818	1.416478	84.14%	0.071155	0.00043	0.071734	0.070575	0.359201
14	13	>	15	0.003564	0.000971	3.670968	99.96%	0.108324	0.00052	0.110106	0.106542	0.427537
16	14	<	7	0.029903	0.001494	20.008691	99.96%	0.146006	0.000703	0.131055	0.160957	0.486882
17	16	>	8	0.019875	0.001817	10.935982	99.96%	0.165862	0.000769	0.1758	0.155925	0.591651
18	17	<	9	0.094665	0.003641	25.996654	99.96%	0.314637	0.001426	0.267304	0.361969	0.633355

Q=2117.531370



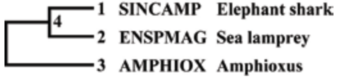
node	L		R	delta	s.e.	Z	CP	height	s.e.	bA	bB	bC
6	1	>	2	0.026826	0.001575	17.030176	99.96%	0.121249	0.000691	0.134662	0.107836	0.53405
7	6	>	3	0.042324	0.002091	20.243862	99.96%	0.192963	0.000931	0.214125	0.171801	0.575861
8	7	<	4	0.081202	0.003657	22.205508	99.96%	0.318735	0.001443	0.278134	0.359336	0.636151

Q=1243.697200



node	L		R	delta	s.e.	Z	CP	height	s.e.	bA	bB	bC
4	1	<	2	0.11097	0.003974	27.926665	99.96%	0.305393	0.001529	0.249908	0.360879	0.634594

Q=779.898602



Supplementary Table VI.3 | Takezaki Two-Cluster test using the 13-chordate dataset

Mean pairwise distances of different vertebrate ingroup clusters to the outgroup (amphioxus) are shown. The distances were obtained from the ML tree shown in Figure 1. Z-statistics were used to infer if the differences between the distances to the outgroup for the two ingroup clusters are significantly different from 0 or not. Standard error (S.E.) for calculating the Z-statistics were obtained using distances from 100 bootstrap replicates (MEGA5) and 100 random Bayesian trees. Mean pairwise distance to the outgroup for ingroup a (Lac) and ingroup b (Lbc) are shown; |δ| = |Lac-Lbc| is the absolute value of the difference between the distances of the two ingroup clusters to the outgroup; Z-statistics is |δ|/S.E.; CP (confidence probability) value is 1 – P-value.

Ingroup-a	**Ingroup-b**	L_{ac}	L_{bc}	**|δ|**	**Bayes100**				**MEGA**			
					S.E.	**Z-stat**	**P-value**	**CP value**	**S.E.**	**Z-stat**	**P-value**	**CP value**
C. milii	Coelacanth	0.93	0.96	0.03	0.0002	181.35	0	100.00%	0.0008	37.88	0	100.00%
C. milii	Tetrapods	0.93	1.016	0.09	0.0002	519.96	0	100.00%	0.0008	106.54	0	100.00%
C. milii	Teleosts	0.93	1.05	0.12	0.0002	646.79	0	100.00%	0.0010	127.54	0	100.00%
C. milii	Sea lamprey	0.93	1.04	0.12	0.0003	424.14	0	100.00%	0.0016	75.25	0	100.00%

Supplementary Table VI.4 | Pairwise distance to the outgroup (amphioxus) for the 13-chordate dataset.
Pairwise distances were calculated from the 13-chordate neutral tree using the R-package 'ape' (see Methods).

Species	Pairwise distance to amphioxus (substitutions per 4D site)
Human	2.3549287
Mouse	2.5619558
Cow	2.4145335
Opossum	2.3145022
Chicken	2.2832321
Lizard	2.3958730
Xenopus	2.551982
Coelacanth	2.121122
Stickleback	2.601429 (highest)
Zebrafish	2.413519
C. milii	**2.058743 (lowest)**
Sea lamprey	2.125137

Supplementary Table VII.1 | Intron gain and loss in deep gnathostome lineages.

Lgi, *Lottia gigantea*; Nve, *Nematostella vectensis*; Tad, *Trichoplax adhaerens*.

			Intron Presence/Absence			Non-chordate invertebrates	
Sl. No.	**Gene**	**Elephant shark gene ID**	**Elephant shark**	**Osteichthyes**	**Amphioxus**	**Present**	**Absent**
	Osteichthyes ancestor gains						
1	FAM46A	SINCAMT00000026881	-	+	-	-	Lgi, Nve, Tad
2	FBXO33	SINCAMT00000025248	-	+	-	-	Lgi
3	FBXO45	SINCAMT00000025801	-	+	-	-	Lgi, Nve
4	FEM1B	SINCAMT00000015243	-	+	-	-	Lgi, Nve, Tad
5	KLHDC4	SINCAMT00000006825	-	+	-	-	-
6	ALG2	SINCAMT00000026708	-	+	-	-	Lgi, Nve
7	ESPL1	SINCAMT00000016975	-	+	-	-	-
8	BRPF1	SINCAMT00000003885	-	+	-	-	Lgi, Nve, Tad
9	CBLB	SINCAMT00000023852	-	+	-	-	Lgi, Nve, Tad
10	NDUFA8	SINCAMT00000002244	-	+	-	-	Lgi, Nve, Tad
11	N4BP3	SINCAMT00000018579	-	+	-	-	Lgi
12	ITIH5	SINCAMT00000014592	-	+	-	-	Lgi
13	PIGS	SINCAMT00000021003	-	+	-	-	Lgi, Tad
	Osteichthyes ancestor losses						
1	ATP13A1	SINCAMT00000022114	+	-	+	NA	NA
2	MMP24	SINCAMT00000021968	+	-	+	-	Lgi, Nve
3	CPT2	SINCAMT00000007581	+	-	+	NA	NA
4	HSPA5	SINCAMT00000018894	+	-	+	NA	NA
5	ZYG11B	SINCAMT00000012990	+	-	+	NA	NA
6	EIF2C4	SINCAMT00000017848	+	-	+	NA	NA
7	SGK1	SINCAMT00000004343	+	-	+	NA	NA
8	DDX42	SINCAMT00000017140	+	-	+	NA	NA
9	MCM2	SINCAMT00000005748	+	-	+	NA	NA
10	HECW1	SINCAMT00000007670	+	-	+	NA	NA
	Elephant shark gains						
1	ATP13A1	SINCAMT00000022114	+	-	-	-	Lgi, Nve, Tad
2	FEM1B	SINCAMT00000015243	+	-	-	-	Lgi, Nve, Tad
3	NSA2	SINCAMT00000005981	+	-	-	-	Lgi, Nve, Tad
4	EXTL3	SINCAMT00000015064	+	-	-	-	Lgi, Nve
5	EXTL3	SINCAMT00000015064	+	-	-	-	Lgi, Nve
6	PMPCA	SINCAMT00000003405	+	-	-	-	Lgi, Nve, Tad
7	BRD4	SINCAMT00000021269	+	-	-	-	Lgi, Nve, Tad
8	SLITRK1	SINCAMT00000008887	+	-	-	-	Lgi, Nve, Tad
9	PTPRO	SINCAMT00000000077	+	-	-	NA	NA

	Elephant shark losses						
1	COG8	SINCAMT00000006291	-	+	+	NA	NA
2	ODZ1	SINCAMT00000025114	-	+	+	NA	NA
3	DGKE	SINCAMT00000017717	-	+	+	NA	NA
4	TIMM10	SINCAMT00000026952	-	+	+	NA	NA
5	NUDT19	SINCAMT00000026551	-	+	+	NA	NA
6	NUDT19	SINCAMT00000026551	-	+	+	NA	NA

Supplementary Table VIII.5 | *C. milii* scaffolds showing one-to-one gene synteny with either chicken or human chromosomes but not corresponding to the 40 ancestral GLGs identified by Nakatani et al. (2007).

Elephant shark scaffold	synteny with chicken		synteny with human	
	Syntenic chicken chr.	No. of orthologs	Syntenic human chr.	No. of orthologs
scaffold_110	chr_1	9	chr_21	9
scaffold_152	chr_1	10	chr_21	12
scaffold_170	chr_1	14	chr_21	12
scaffold_222	chr_1	4	chr_21	4
scaffold_255	chr_1	9	chr_21	8
scaffold_217	chr_3	7	chr_8	5
scaffold_218	chr_4	3	chr_2	6
scaffold_239	chr_7	5	chr_21	6
scaffold_246	chr_7	9	chr_3	5
scaffold_102	chr_7	6	chr_3,chr_X	8
scaffold_426	chr_14	4	chr_17	4
scaffold_214	chr_19	3	chr_7	10

Supplementary Table VIII.6 | *C. milii* scaffolds showing synteny with the macro and microchromosomes of chicken. Chicken chromosomes 11 to 28 (in bold) are microchromosomes. Scaffolds marked with an asterisk are syntenic to more than 1 chicken chromosome. Each of the 4 elephant shark scaffolds highlighted in red is syntenic to one microchromosome and one macrochromosome.

Chicken chr.	Elephant shark scaffolds
1	7*, 6*, 4*, 18, 20, 17*, 42, 39, 16*, 35, 53*, 63, 81, 113, 119, 133, 145, 118*, 235, 123, 170, 261, 237, 191, 29*, 181, 157, 152, 255, 110, 136, 206, 299, 267, 344, 327, 271, 222, 975, 438, 305, 291, 194
2	7*, 9*, 4*, 17*, 16*, 13, 53*, 48, 49*, 60, 79, 37, 56, 84, 83, 55, 40, 80, 72, 90, 86, 105, 104, 118*, 98, 129, 99, 29*, 187, 142, 207, 160, 150, 210, 178, 139, 89*, 108, 158, 245, 233, 307, 272
3	1*, 23, 34, 26, 31, 49*, 32, 68, 62, 61, 71, 134, 51, 135, 106, 109*, 164, 143, 128, 159, 120, 78*, 284, 217, 213, 124
4	2, 11, 25, 53*, 19*, 21*, 59, 50, 67, 43, 57, 138, 65*, 117, 96, 193*, 166, 269, 230, 78*, 208, 175, 335, 198, 268, 218
5	9*, 8, 28, 114, 153, 130, 249, 319, 216, 313, 602, 334, 179, 174
6	3, 115, 111, 146, 126, 241, 453, 352
7	14, 16*, 45*, 46, 76, 246, 288, 149, 102, 239, 301
8	7*, 41, 74, 234, 101, 112, 252, 283, 745
9	1*, 95
10	5, 211, 282, 203, 399
11	12, 197, 127, 189, 148
12	15, 107, 38, 231, 333, 347, 539
13	87, 19*, 204, 92, 156, 144, 168, 200
14	30, 147, 154, 131, 109*, 163, 212, 426, 312, 205
15	94, 91, 196, 66, 226, 356
17	77, 45*, 100, 258, 236

Supplementary Table VIII.6 | Cont'd.

Chicken chr.	Elephant shark scaffolds
18	10, 514, 242
19	47, 82, 155, 398, 397, 214
20	6*, 137, 247, 315, 223
21	93, 58, 400, 338, 183
22	368, 380, 532
23	36, 121, 227, 171, 162
24	44, 184, 228, 511
26	70, 33
27	88, 122, 185, 293, 256, 357, 251
28	64, 85, 209, 316, 292
Z	22, 16*, 21*, 24, 27, 52, 73, 54, 140, 75, 69, 103, 65*, 193*, 165, 173, 167, 89*, 232, 202, 172, 132, 591

Supplementary Table VIII.7 | Interchromosomal rearrangements in the medaka lineage.
Syntenic relationships based on comparison of *C. milii* scaffolds and medaka chromosomes are compared to the 13 teleost proto-chromosomes reconstructed by Kasahara et al. (2007). Novel interchromosomal rearrangements predicted in this study are highlighted in red.

Syntenic relationship identified by Kasahara et al. (2007)[1]		Syntenic relationship identified by comparison with elephant shark.	
Teleost proto-chromosome	**Medaka chromosomes**	**Elephant shark scaffold**	**Medaka chromosomes**
b	Ola11, Ola16	scaffold_13	Ola4, Ola11, Ola16, Ola20
c	Ola2, Ola21	scaffold_14	Ola2, Ola3, Ola17, Ola21
d	Ola1, Ola15, Ola19	scaffold_31	Ola1, Ola15, Ola24
e	Ola1, Ola8, Ola19	scaffold_10	Ola1, Ola8, Ola19
		scaffold_30	
f	Ola1, Ola10, Ola18	scaffold_2	Ola10, Ola14, Ola18
		scaffold_11	Ola1, Ola4, Ola10, Ola12, Ola18
g	Ola10, Ola14	scaffold_2	Ola10, Ola14, Ola18
h	Ola13, Ola14	scaffold_18	Ola4, Ola13, Ola14, Ola21
i	Ola9, Ola12	scaffold_22	Ola9, Ola12
		scaffold_73	
		scaffold_94	
		scaffold_140	
		scaffold_196	
j	Ola3, Ola6	scaffold_5	Ola3, Ola6
		scaffold_8	
		scaffold_12	
		scaffold_197	
		scaffold_211	
		scaffold_282	
k	Ola6, Ola23	scaffold_35	Ola6, Ola23
		scaffold_39	
		scaffold_119	
		scaffold_237	
l	Ola5, Ola7	scaffold_15	Ola5, Ola7
		scaffold_38	
		scaffold_33	
		scaffold_58	
		scaffold_70	
		scaffold_107	
		scaffold_137	
		scaffold_223	

Supplementary Table VIII.7 | Cont'd.

Syntenic relationship identified by Kasahara et al. (2007)[1]		Syntenic relationship identified by comparison with elephant shark.	
Teleost proto-chromosome	Medaka chromosomes	Elephant shark scaffold	Medaka chromosomes
m	Ola4, Ola13, Ola17, Ola20	scaffold_48	Ola11, Ola17, Ola20
		scaffold_20	Ola3, Ola4, Ola13, Ola21

[1] Kasahara M, *et al*. 2007. The medaka draft genome and insights into vertebrate genome evolution. Nature 447: 714-719.

Supplementary Table VIII.8 | Interchromosomal rearrangements in the zebrafish lineage.
Syntenic relationships based on comparison of *C. milii* scaffolds and zebrafish chromosomes are compared to the 13 teleost proto-chromosomes reconstructed by Kasahara et al. (2007). Novel interchromosomal rearrangements predicted in this study are highlighted in red.

Syntenic relationship identified by Kasahara et al. (2007)[1]		Syntenic relationship identified by comparison with elephant shark.	
Teleost proto-chromosome	Zebrafish chromosomes	Elephant shark scaffold	Zebrafish chromosomes
a	Dre13, Dre17, Dre20	scaffold_26	Dre4, Dre16, Dre17, Dre20, Dre23
		scaffold_3	Dre12, Dre13, Dre17
		scaffold_114	Dre13, Dre17, Dre20
		scaffold_34	Dre13, Dre17, Dre20, Dre22
b	Dre16, Dre19	scaffold_13	Dre2, Dre16, Dre19, Dre24
d	Dre1, Dre12, Dre13	scaffold_23	Dre1, Dre12, Dre13, Dre17
		scaffold_3	Dre12, Dre13, Dre17
e	Dre1, Dre3, Dre12	scaffold_23	Dre1, Dre12, Dre13, Dre17
		scaffold_30	Dre1, Dre3, Dre12, Dre24
f	Dre1, Dre7, Dre14	scaffold_25	Dre1, Dre14, Dre20, Dre23
		scaffold_11	Dre1, Dre7, Dre10, Dre14, Dre23
		scaffold_2	Dre5, Dre7, Dre14
g	Dre5, Dre10, Dre14, Dre21	scaffold_11	Dre1, Dre7, Dre10, Dre14, Dre23
		scaffold_50	Dre5, Dre10, Dre14, Dre21
		scaffold_44	Dre5, Dre10, Dre15, Dre18
		scaffold_47	Dre5, Dre10, Dre15, Dre21
		scaffold_27	Dre5, Dre10, Dre21
		scaffold_2	Dre5, Dre7, Dre14
		scaffold_22	Dre5, Dre8, Dre10
		scaffold_94	Dre5, Dre8, Dre10, Dre21
		scaffold_100	
h	Dre5, Dre10, Dre15, Dre21	scaffold_50	Dre5, Dre10, Dre14, Dre21
		scaffold_44	Dre5, Dre10, Dre15, Dre18
		scaffold_47	Dre5, Dre10, Dre15, Dre21
		scaffold_27	Dre5, Dre10, Dre21
		scaffold_22	Dre5, Dre8, Dre10
		scaffold_94	Dre5, Dre8, Dre10, Dre21
		scaffold_100	
		scaffold_18	Dre6, Dre9, Dre10, Dre15
i	Dre5, Dre8, Dre10, Dre21	scaffold_50	Dre5, Dre10, Dre14, Dre21
		scaffold_44	Dre5, Dre10, Dre15, Dre18
		scaffold_47	Dre5, Dre10, Dre15, Dre21
		scaffold_27	Dre5, Dre10, Dre21
		scaffold_22	Dre5, Dre8, Dre10

Supplementary Table VIII.8 | Cont'd.

Syntenic relationship identified by Kasahara et al. (2007)[1]		Syntenic relationship identified by comparison with elephant shark.	
Teleost proto-chromosome	**Zebrafish chromosomes**	**Elephant shark scaffold**	**Zebrafish chromosomes**
i	Dre5, Dre8, Dre10, Dre21	scaffold_94	Dre5, Dre8, Dre10, Dre21
		scaffold_100	
j	Dre7, Dre18, Dre25	scaffold_39	Dre4, Dre18, Dre25
		scaffold_12	Dre7, Dre17, Dre18, Dre25
		scaffold_8	Dre7, Dre18, Dre25
		scaffold_5	
		scaffold_211	
k	Dre4, Dre18, Dre25,	scaffold_39	Dre4, Dre18, Dre25
		scaffold_12	Dre7, Dre17, Dre18, Dre25
		scaffold_8	Dre7, Dre18, Dre25
		scaffold_5	
		scaffold_211	
l	Dre6, Dre11, Dre23	scaffold_14	Dre1, Dre6, Dre9, Dre11, Dre22
		scaffold_20	Dre1, Dre6, Dre9, Dre11, Dre24
		scaffold_10	Dre3, Dre6, Dre11, Dre12, Dre22
m	Dre2, Dre6, Dre8, Dre11, Dre15, Dre22, Dre24	scaffold_14	Dre1, Dre6, Dre9, Dre11, Dre22
		scaffold_20	Dre1, Dre6, Dre9, Dre11, Dre24
		scaffold_13	Dre2, Dre16, Dre19, Dre24
		scaffold_48	Dre2, Dre19, Dre24
		scaffold_74	Dre2, Dre6, Dre8, Dre20
		scaffold_41	Dre2, Dre6, Dre8, Dre20, Dre22
		scaffold_10	Dre3, Dre6, Dre11, Dre12, Dre22
		scaffold_18	Dre6, Dre9, Dre10, Dre15
-	-	scaffold_42	Dre9, Dre10, Dre22

1. Kasahara M, *et al*. 2007. The medaka draft genome and insights into vertebrate genome evolution. Nature 447: 714-719.

Supplementary Table IX.1 | Top 100 protein domains in *C. milii*

S/N	Protein domain	Pfam IDs	Protein count	Percentage of total proteins
1	Immunoglobulin V-set domain	PF07686	596	3.4155
2	7 transmembrane receptor (rhodopsin family)	PF00001	589	3.3754
3	Immunoglobulin domain	PF13895, PF00047, PF13927	527	3.0201
4	Protein kinase domain	PF00069	472	2.7049
5	Protein tyrosine kinase	PF07714	463	2.6533
6	Immunoglobulin I-set domain	PF07679	295	1.6905
7	Zinc finger, C3HC4 type (RING finger)	PF13920, PF13923, PF00097	287	1.6447
8	WD domain, G-beta repeat	PF00400	233	1.3352
9	C2H2-type zinc finger	PF13894, PF13912	225	1.2894
10	Zinc finger, C2H2 type	PF00096	222	1.2722
11	Ring finger domain	PF13639	214	1.2264
12	Zinc-finger double domain	PF13465	211	1.2092
13	Ankyrin repeats (3 copies)	PF12796	200	1.1461
14	Ankyrin repeat	PF00023, PF13606	198	1.1347
15	Homeobox domain	PF00046	198	1.1347
16	Leucine rich repeat	PF13504, PF13855	192	1.1003
17	Leucine Rich Repeat	PF00560	190	1.0888
18	Ankyrin repeats (many copies)	PF13637, PF13857	187	1.0716
19	RNA recognition motif. (a.k.a. RRM, RBD, or RNP domain)	PF00076, PF13893	181	1.0372
20	Ras family	PF00071	172	0.9857
21	B-box zinc finger	PF00643	166	0.9513
22	Miro-like protein	PF08477	164	0.9398
23	ADP-ribosylation factor family	PF00025	161	0.9226
24	RNA recognition motif (a.k.a. RRM, RBD, or RNP domain)	PF14259	153	0.8768
25	SPRY domain	PF00622	150	0.8596
26	RING-type zinc-finger, LisH dimerisation motif	PF13445	149	0.8539
27	PH domain	PF00169	148	0.8481
28	Variant SH3 domain	PF07653	148	0.8481
29	SH3 domain	PF00018	145	0.8309
30	Fibronectin type III domain	PF00041	139	0.7966

31	Serpentine type 7TM GPCR chemoreceptor Srsx	PF10320	130	0.7450
32	Leucine Rich repeats (2 copies)	PF12799	126	0.7221
33	BTB/POZ domain	PF00651	124	0.7106
34	PDZ domain (Also known as DHR or GLGF)	PF00595	122	0.6991
35	SPRY-associated domain	PF13765	122	0.6991
36	Leucine Rich repeat	PF13516	121	0.6934
37	EF-hand domain pair	PF13499, PF13833	119	0.6819
38	EF-hand domain	PF13405	111	0.6361
39	C2 domain	PF00168	110	0.6304
40	Tetratricopeptide repeat	PF07719, PF00515, PF13371, PF13181, PF13176, PF13174, PF13432, PF13429, PF13424, PF13374, PF13431, PF07721, PF07720, PF13428	105	0.6017
41	SH2 domain	PF00017	102	0.5845
42	Major Facilitator Superfamily	PF07690	101	0.5788
43	Ion transport protein	PF00520	99	0.5673
44	Immunoglobulin C1-set domain	PF07654	98	0.5616
45	EF hand	PF13202, PF00036, PF09068	96	0.5501
46	50S ribosome-binding GTPase	PF01926	95	0.5444
47	Helix-loop-helix DNA-binding domain	PF00010	90	0.5158
48	Leucine rich repeats (6 copies)	PF13306	90	0.5158
49	Calcium-binding EGF domain	PF07645	89	0.5100
50	Helicase conserved C-terminal domain	PF00271, PF13625	89	0.5100
51	Zinc-finger of C2H2 type	PF12874	88	0.5043
52	AAA domain	PF13476, PF13086, PF13604, PF13087, PF13173, PF13481, PF13304, PF13401, PF13207, PF13238, PF13671, PF13614	87	0.4986
53	Cadherin domain	PF00028	85	0.4871
54	SAM domain (Sterile alpha motif)	PF00536, PF07647	85	0.4871
55	Collagen triple helix repeat (20 copies)	PF01391	84	0.4814
56	Scavenger receptor cysteine-rich domain	PF00530	84	0.4814

57	CD80-like C2-set immunoglobulin domain	PF08205	81	0.4642
58	TPR repeat	PF13414	78	0.4470
59	Elongation factor Tu GTP binding domain	PF00009	75	0.4298
60	DEAD/DEAH box helicase	PF00270	73	0.4183
61	EGF-like domain	PF00008, PF07974	72	0.4126
62	Gtr1/RagA G protein conserved region	PF04670	72	0.4126
63	Receptor family ligand binding region	PF01094	70	0.4011
64	Trypsin	PF00089	70	0.4011
65	Kelch motif	PF01344, PF07646, PF13964, PF13854	67	0.3840
66	NACHT domain	PF05729	67	0.3840
67	RhoGAP domain	PF00620	65	0.3725
68	RhoGEF domain	PF00621	64	0.3668
69	Serpentine type 7TM GPCR chemoreceptor Srx	PF10328	64	0.3668
70	von Willebrand factor type A domain	PF13768, PF00092, PF13519	64	0.3668
71	ATPase family associated with various cellular activities (AAA)	PF00004, PF07726	61	0.3496
72	Protein-tyrosine phosphatase	PF00102	60	0.3438
73	7 transmembrane sweet-taste receptor of 3 GCPR	PF00003	59	0.3381
74	BTB And C-terminal Kelch	PF07707	59	0.3381
75	Thrombospondin type 1 domain	PF00090	59	0.3381
76	Calponin homology (CH) domain	PF00307	58	0.3324
77	short chain dehydrogenase	PF00106	58	0.3324
78	LIM domain	PF00412	57	0.3266
79	Galactose oxidase, central domain	PF13418, PF13415	56	0.3209
80	Methyltransferase domain	PF12847, PF13847, PF13489, PF13659, PF13649, PF08242, PF08241, PF13679, PF13578, PF13383	56	0.3209
81	PHD-finger	PF00628, PF13831	55	0.3152
82	Nine Cysteines Domain of family 3 GPCR	PF07562	54	0.3095
83	Alpha/beta hydrolase family	PF12695, PF12697	52	0.2980
84	Ion channel	PF07885	51	0.2923
85	Signal recognition particle receptor beta subunit	PF09439	50	0.2865

86	Laminin G domain	PF02210, PF00054	49	0.2808
87	Lectin C-type domain	PF00059	49	0.2808
88	7 transmembrane receptor (Secretin family)	PF00002	47	0.2693
89	CUB domain	PF00431	47	0.2693
90	Ligand-binding domain of nuclear hormone receptor	PF00104	47	0.2693
91	Periplasmic binding protein	PF13458	47	0.2693
92	Ubiquitin carboxyl-terminal hydrolase	PF00443, PF13423	47	0.2693
93	Kinesin motor domain	PF00225	46	0.2636
94	Mitochondrial carrier protein	PF00153	46	0.2636
95	Sugar (and other) transporter	PF00083	46	0.2636
96	Cytochrome P450	PF00067	45	0.2579
97	Dual specificity phosphatase, catalytic domain	PF00782	45	0.2579
98	K+ channel tetramerisation domain	PF02214	45	0.2579
99	FERM central domain	PF00373	44	0.2521
100	PMP-22/EMP/MP20/Claudin tight junction	PF13903	44	0.2521

Supplementary Table IX.2 | Protein domains uniquely present in *C. milii*

S/N	Protein domain	Percentage of total proteins in elephant shark
1	3-Oxoacyl-[acyl-carrier-protein (ACP)] synthase III	0.0057
2	Bacterial transcriptional activator domain	0.0057
3	Cupin	0.0115
4	DinB superfamily	0.0057
5	Herpesvirus alkaline exonuclease	0.0057
6	H-type lectin domain	0.0057
7	Leucine-zipper of ternary complex factor MIP1	0.0057
8	Marek's disease glycoprotein A	0.0057
9	PEP-utilising enzyme, mobile domain	0.0057
10	PHAT	0.0057
11	Poxvirus D5 protein-like	0.0115
12	Putative sugar-binding domain	0.0057
13	Pyruvate phosphate dikinase, PEP/pyruvate binding domain	0.0057
14	RbcX protein	0.0057
15	RnfC Barrel sandwich hybrid domain	0.0057
16	Transcriptional regulatory protein, C terminal	0.0057
17	Tyrosine phosphatase family C-terminal region	0.0057

Supplementary Table IX.3 | Protein domains present in bony vertebrates but absent in *C. milii*.

Teleosts include zebrafish and stickleback; reptiles include *Anolis* lizard and chicken; mammals include human, mouse, cow and opossum.

S/N	Protein domain	Percentage of total proteins			
		Teleosts	*Xenopus*	Reptiles	Mammals
1	Acetyl xylan esterase (AXE1)	0.0043	0.0054	0.0058	0.0071
2	Acetylcholinesterase tetramerisation domain	0.0043	0.0109	0.0058	0.0071
3	Adenosine deaminase z-alpha domain	0.0064	0.0054	0.0029	0.0083
4	Alanine-glyoxylate amino-transferase	0.0021	0.0054	0.0029	0.0036
5	Aluminium induced protein	0.0021	0.0054	0.0058	0.0036
6	Binding domain of DNA repair protein Ercc1 (rad10/Swi10)	0.0043	0.0054	0.0058	0.0048
7	BRCA2, helical	0.0043	0.0054	0.0058	0.0048
8	BRCA2, oligonucleotide/oligosaccharide-binding, domain 1	0.0043	0.0054	0.0058	0.0048
9	BRCA2, oligonucleotide/oligosaccharide-binding, domain 3	0.0043	0.0054	0.0058	0.0048
10	Brf1-like TBP-binding domain	0.0085	0.0054	0.0058	0.0048
11	Brinker DNA-binding domain	0.0064	0.0380	0.0116	0.0048
12	CAAX protease self-immunity	0.0085	0.0054	0.0029	0.0048
13	CHDNT (NUC034) domain	0.0149	0.0163	0.0145	0.0143
14	Chromosome passenger complex (CPC) protein INCENP N terminal	0.0043	0.0054	0.0116	0.0036
15	Ciliary neurotrophic factor	0.0043	0.0109	0.0145	0.0107
16	CLN3 protein	0.0043	0.0109	0.0029	0.0048
17	COQ9	0.0043	0.0054	0.0058	0.0048
18	Dolichol-phosphate mannosyltransferase subunit 3 (DPM3)	0.0043	0.0054	0.0029	0.0048
19	DSBA-like thioredoxin domain	0.0107	0.0109	0.0058	0.0059
20	EMG1/NEP1 methyltransferase	0.0043	0.0054	0.0058	0.0036
21	Evolutionarily conserved signalling intermediate in Toll pathway	0.0043	0.0054	0.0029	0.0048
22	Fanconi anemia group F protein (FANCF)	0.0043	0.0054	0.0058	0.0048
23	Fibrinogen alpha C domain	0.0043	0.0054	0.0058	0.0048
24	Flavodoxin-like fold	0.0234	0.0054	0.0116	0.0107
25	GIY-YIG catalytic domain	0.0043	0.0054	0.0029	0.0059

26	Growth arrest and DNA-damage-inducible proteins-interacting protein 1	0.0043	0.0054	0.0029	0.0048
27	Histone chaperone domain CHZ	0.0021	0.0054	0.0029	0.0048
28	Hormone-sensitive lipase (HSL) N-terminus	0.0085	0.0054	0.0029	0.0048
29	Hydantoinase/oxoprolinase	0.0043	0.0054	0.0058	0.0036
30	Hydantoinase/oxoprolinase N-terminal region	0.0043	0.0054	0.0029	0.0036
31	Interleukin 11	0.0085	0.0054	0.0029	0.0036
32	Iron/zinc purple acid phosphatase-like protein C	0.0043	0.0054	0.0029	0.0048
33	Jacalin-like lectin domain	0.0021	0.0109	0.0087	0.0143
34	K167R (NUC007) repeat	0.0021	0.0054	0.0029	0.0048
35	Mandelate racemase / muconate lactonizing enzyme, C-terminal domain	0.0043	0.0054	0.0029	0.0036
36	Mandelate racemase / muconate lactonizing enzyme, N-terminal domain	0.0043	0.0054	0.0029	0.0036
37	MazG nucleotide pyrophosphohydrolase domain	0.0043	0.0054	0.0029	0.0048
38	Mediator complex subunit 25 synapsin 1	0.0021	0.0054	0.0058	0.0048
39	Menin	0.0064	0.0054	0.0029	0.0036
40	Methylpurine-DNA glycosylase (MPG)	0.0043	0.0054	0.0058	0.0048
41	Mitochondrial protein from FMP27	0.0043	0.0054	0.0058	0.0048
42	NADH:flavin oxidoreductase / NADH oxidase family	0.0021	0.0054	0.0029	0.0012
43	NADPH-dependent FMN reductase	0.0234	0.0054	0.0116	0.0107
44	NIF3 (NGG1p interacting factor 3)	0.0043	0.0054	0.0058	0.0048
45	non-SMC mitotic condensation complex subunit 1, N-term	0.0043	0.0054	0.0058	0.0048
46	Olfactory receptor	0.3749	2.1379	0.9525	4.0990
47	Pancreatic ribonuclease	0.0064	0.0054	0.0318	0.0737
48	Peptidase M60-like family	0.0085	0.0217	0.0087	0.0095
49	Pex19 protein family	0.0043	0.0054	0.0029	0.0071
50	Polynucleotide kinase 3 phosphatase	0.0043	0.0054	0.0029	0.0048
51	PP1-regulatory protein, Phostensin N-terminal	0.0021	0.0054	0.0058	0.0143
52	Pregnancy-associated plasma protein-A	0.0128	0.0109	0.0116	0.0095

53	Prothymosin/parathymosin family	0.0085	0.0109	0.0087	0.0083
54	PTN/MK heparin-binding protein family, C-terminal domain	0.0107	0.0054	0.0116	0.0083
55	Quinolinate phosphoribosyl transferase, C-terminal domain	0.0021	0.0054	0.0029	0.0048
56	Quinolinate phosphoribosyl transferase, N-terminal domain	0.0021	0.0054	0.0029	0.0048
57	Rab geranylgeranyl transferase alpha-subunit, insert domain	0.0043	0.0054	0.0029	0.0048
58	Rap1 Myb domain	0.0043	0.0054	0.0029	0.0048
59	Ribosomal protein L32	0.0043	0.0109	0.0029	0.0083
60	Ribosomal protein S13/S18	0.0043	0.0054	0.0029	0.0131
61	RNA polymerase Rpb1 C-terminal repeat	0.0021	0.0054	0.0029	0.0071
62	RNA polymerase Rpb1, domain 6	0.0043	0.0054	0.0029	0.0048
63	RNA polymerase Rpb1, domain 7	0.0043	0.0054	0.0029	0.0048
64	RNA polymerases N / 8 kDa subunit	0.0021	0.0109	0.0058	0.0048
65	rRNA small subunit methyltransferase G	0.0043	0.0054	0.0029	0.0048
66	SCAN domain	0.0170	0.0868	0.3300	0.2247
67	Semialdehyde dehydrogenase, NAD binding domain	0.0021	0.0109	0.0174	0.0071
68	Seminal vesicle autoantigen (SVA)	0.0192	0.0326	0.0145	0.0190
69	Serum albumin family	0.0021	0.0109	0.0203	0.0214
70	Siah interacting protein, N terminal	0.0043	0.0054	0.0058	0.0048
71	SLBB domain	0.0043	0.0054	0.0029	0.0059
72	Sox developmental protein N terminal	0.0234	0.0163	0.0174	0.0143
73	Srg family chemoreceptor	0.0021	0.0054	0.0029	0.0024
74	Succinate dehydrogenase/Fumarate reductase transmembrane subunit	0.0043	0.0054	0.0029	0.0048
75	Tellurite resistance protein TehB	0.0107	0.0054	0.0029	0.0012
76	Telomere regulation protein Stn1	0.0021	0.0054	0.0058	0.0048
77	TFIIH C1-like domain	0.0043	0.0054	0.0058	0.0083
78	Thrombomodulin like fifth domain, EGF-like	0.0064	0.0054	0.0029	0.0048
79	Timeless protein	0.0043	0.0054	0.0029	0.0036
80	Timeless protein C terminal region	0.0043	0.0054	0.0029	0.0036
81	Topoisomerase II-associated protein PAT1	0.0064	0.0109	0.0087	0.0095
82	Touch receptor neuron protein Mec-17	0.0043	0.0054	0.0029	0.0119
83	Tower	0.0043	0.0054	0.0058	0.0048
84	Transcription factor Tfb2	0.0043	0.0109	0.0029	0.0107

85	Transcriptional regulatory protein LGE1	0.0064	0.0054	0.0058	0.0059
86	Transducer of regulated CREB activity, N terminus	0.0192	0.0109	0.0029	0.0107
87	Translation machinery associated TMA7	0.0043	0.0054	0.0058	0.0071
88	Trehalase	0.0043	0.0054	0.0029	0.0048
89	UvrD/REP helicase N-terminal domain	0.0064	0.0054	0.0145	0.0083
90	Vitelline membrane outer layer protein I (VOMI)	0.0064	0.0109	0.0116	0.0048
91	VWA domain containing CoxE-like protein	0.0128	0.0054	0.0029	0.0071
92	Zeta toxin	0.0064	0.0217	0.0261	0.0250

Supplementary Table IX.4 | Protein domains present in *C. milii* and tetrapods but lost in teleosts

Teleosts include zebrafish, stickleback, medaka and fugu; reptiles include *Anolis* lizard and chicken; mammals include human, mouse, cow and opossum.

S/N	Protein domain	Percentage of total proteins			
		Elephant shark	***Xenopus***	**Reptiles**	**Mammals**
1	Acyl-CoA reductase (LuxC)	0.0057	0.0109	0.0058	0.0155
2	Alpha 1,4-glycosyltransferase conserved region	0.0115	0.0651	0.0116	0.0083
3	Apolipoprotein B100 C terminal	0.0057	0.0054	0.0058	0.0036
4	Endoplasmic reticulum protein ERp29, C-terminal domain	0.0057	0.0054	0.0087	0.0048
5	ERp29, N-terminal domain	0.0057	0.0054	0.0087	0.0048
6	Fanconi anaemia group A protein	0.0057	0.0054	0.0058	0.0048
7	Formin Homology Region 1	0.0057	0.0054	0.0087	0.0083
8	Glycosyl transferase family 11	0.0287	0.0217	0.0029	0.0155
9	Glycosyltransferase sugar-binding region containing DXD motif	0.0115	0.0651	0.0116	0.0083
10	Mis12-Mtw1 protein family	0.0057	0.0054	0.0058	0.0048
11	piRNA pathway germ-plasm component	0.0057	0.0054	0.0058	0.0048
12	Progesterone receptor	0.0057	0.0054	0.0029	0.0036
13	Serine-rich domain associated with BRCT	0.0057	0.0054	0.0058	0.0048

Supplementary Table IX.5 | Protein domains present in *C. milii* and teleosts but lost in tetrapods

Teleosts include zebrafish and stickleback; tetrapods include *Xenopus*, lizard, chicken and mammals.

S/N	Protein domain	Percentage of total proteins	
		Elephant shark	Teleosts
1	Helitron helicase-like domain at N-terminus	0.0057	0.0107
2	Malate/L-lactate dehydrogenase	0.0057	0.0021
3	non-haem dioxygenase in morphine synthesis N-terminal	0.0115	0.0043
4	Peptide-N-glycosidase F, C terminal	0.0057	0.0043
5	Peptide-N-glycosidase F, N terminal	0.0057	0.0043
6	Sea anemone cytotoxic protein	0.0057	0.0085

Supplementary Table X.1 | List of major gene families involved in bone formation in vertebrates

Human gene name	Human protein ID	Elephant shark gene ID	Elephant shark protein ID	Remarks
Hedgehog signaling				
Dispatched	NP_116279.2	SINCAMG00000008471	SINCAMP00000012967	
Scube1	NP_766638.2	-	-	
Scube2	NP_001164161.1	SINCAMG00000006934	SINCAMP00000010589	
Scube3	NP_689966.2	SINCAMG00000010708	SINCAMG00000010708	Conserved synteny with TCP11L1 and ZNF76
HHAT	NP_060664.2	SINCAMG00000014826	SINCAMP00000022770	Linked to KCNH1
Gas1	NP_002039.2	SINCAMG00000016790	SINCAMP00000025879	
Cdo	NP_001230526.1	SINCAMG00000012747	SINCAMP00000019513	
Ptc1	NP_000255.2	SINCAMG00000008503	SINCAMP00000013041	
Ptc2	NP_001159764.1	SINCAMG00000008875	SINCAMP00000013634	
Gpc3	NP_001158089.1	SINCAMG00000014624	SINCAMP00000022513	Conserved synteny with HS6ST2, MBNL3 in the 3’ end and PHF6, HPRT1 in the 5’ end.
Smo	NP_005622.1	SINCAMG00000000727	SINCAMP00000001123	
Kif7	NP_940927.2	SINCAMG00000011775	SINCAMP00000018064	
Sufu	NP_001171604.1	SINCAMG00000013703	SINCAMP00000021033	
Gli1	NP_001153517.1	SINCAMG00000011910	SINCAMP00000018279	
Gli2	NP_005261.2	SINCAMG00000012504	SINCAMP00000019143	
Gli3	NP_000159.3	SINCAMG00000008696	SINCAMP00000013329	
HHIP	NP_071920.1	SINCAMG00000005083	SINCAMP00000007760	
Ihh	NP_002172.2	SINCAMG00000012200	SINCAMP00000018690	
Evc	NP_714928.1	SINCAMG00000015540	SINCAMP00000023908	
Evc2	NP_001159608.1	SINCAMG00000015491	SINCAMP00000023850	

BMP signaling				
Noggin	NP_005441.1	SINCAMG00000016918	SINCAMP00000026007	
Chordin	NP_003732.2	SINCAMG00000006552	SINCAMP00000009976	
Follistatin	NP_006341.1	SINCAMG00000011273	SINCAMP00000017319	
Gremlin	NP_037504.1	SINCAMG00000006221 SINCAMG00000017247	SINCAMP00000009456 SINCAMP00000026336	
Bmp2	NP_001191.1	SINCAMG00000010544	SINCAMP00000016184	
Bmp4	NP_001193.2	SINCAMG00000006702	SINCAMP00000010200	
Bmp5	NP_066551.1	SINCAMG00000013534	SINCAMP00000020750	Confirmed SINCAMG00000013534 by synteny with HMGCLL1 and COL21A1. Removed SINCAMG00000011678 which could be BMP8A based on synteny with MACF1 and NDUFS5.
Bmp6	NP_001709.1	SINCAMG00000000807	SINCAMP00000001237	Found by synteny with TXNDC5 and SNRNP48
Bmp7	NP_001710.1	SINCAMG00000009190	SINCAMP00000014149	Found by synteny with SPO11 and RAE1
Bmpr1a	NP_004320.2	SINCAMG00000016451	SINCAMP00000025412	
Bmpr1b	NP_001243722.1	SINCAMG00000005868	SINCAMP00000008965	
Bmpr2	NP_001195.2	SINCAMG00000010999	SINCAMP00000016909	
Smad1	NP_001003688.1	SINCAMG00000004379 SINCAMG00000005057	SINCAMP00000006697 SINCAMP00000007720	
Smad4	NP_005350.1	SINCAMG00000000232	SINCAMP00000000372	
Smad5	NP_001001419.1	SINCAMG00000005235	SINCAMP00000007989	
Smad6	NP_005576.3	SINCAMG00000010879	SINCAMP00000016713	
Smad7	NP_005895.1	SINCAMG00000000277	SINCAMP00000000443	
Smurf1	NP_065162.1	SINCAMG00000004299	SINCAMP00000006589	
Smurf2	NP_073576.1	SINCAMG00000003087	SINCAMP00000004786	
FGF signalling				
FGF1	NP_000791.1	SINCAMG00000007092	SINCAMP00000010808	

FGF23	NP_065689.1	SINCAMG00000000070	SINCAMP00000000105	Found by synteny with FGF6 and C12orf5
FGF2	NP_001997.5	SINCAMG00000011418	SINCAMP00000017546	
SPRY1	NP_001244967.1	SINCAMG00000016815	SINCAMP00000025904	
SPRY2	NP_005833.1	SINCAMG00000016813	SINCAMP00000025902	
SPRY3	NP_005831.1	SINCAMG00000012081	SINCAMP00000018519	Found by synteny with VAMP7 and TMLHE (scaffold_2)
SPRY4	NP_112226.2	SINCAMG00000017659	SINCAMP00000026750	
FGFR1	NP_075598.2	SINCAMG00000011704 SINCAMG00000013753	SINCAMP00000017959 SINCAMP00000021191	
FGFR2	NP_000132.3	SINCAMG00000001668	SINCAMP00000002567	Found by synteny with ATE1 and WDR11
FGFR3	NP_000133.1	SINCAMG00000005971	SINCAMP00000009100	
FGFR4	NP_002002.3	SINCAMG00000014055	SINCAMP00000021672	
MAPK1	NP_002736.3	SINCAMG00000013172	SINCAMP00000020187	
MAPK8	NP_002741.1	SINCAMG00000004248	SINCAMP00000006507	
RAF1	NP_002871.1	SINCAMG00000002947	SINCAMP00000004516	
RAC1	NP_008839.2	SINCAMG00000006050 SINCAMG00000006052 SINCAMG00000013416	SINCAMP00000009211 SINCAMP00000009224 SINCAMP00000020554	
KRAS	NP_004976.2	SINCAMG00000013095	SINCAMP00000020025	
HRAS	NP_005334.1	SINCAMG00000009949	SINCAMP00000015284	
Transcription factors				
Sox5	NP_008871.3	SINCAMG00000015324	SINCAMP00000023607	
Sox6	NP_059978.1	SINCAMG00000006623	SINCAMP00000010101	
Sox9	NP_000337.1	SINCAMG00000017284	SINCAMP00000026373	
Runx2	NP_001019801.3	SINCAMG00000004114	SINCAMP00000006257	
Osterix/sp7	NP_001166938.1	SINCAMG00000017568	SINCAMP00000026659	

Bapx1	NP_001180.1	SINCAMG00000001225 SINCAMG00000006458	SINCAMP00000001872 SINCAMP00000009804	SINCAMG00000006458 is a fragment; possibly both genes are the same gene. SINCAMG00000001225 is syntenic to CPEB2 and RAB28.
LIM mineralised protein 1 (LMP-1)/mec-3 homolog/four and a half LIM domain containing	NP_005442.2	SINCAMG00000014924	SINCAMP00000022911	
Sp3	NP_003102.1	SINCAMG00000000544	SINCAMP00000000861	
Atf4	NP_001666.2	SINCAMG00000010382 SINCAMG00000010483	SINCAMP00000015933 SINCAMP00000016078	
Twist1	NP_000465.1	SINCAMG00000007976	SINCAMP00000012183	
Twist2	NP_001258822.1	SINCAMG00000003851	SINCAMP00000005857	
sox8	NP_055402.2	SINCAMG00000002752	SINCAMP00000004204	
Pu.1/ Spi-1	NP_001074016.1	SINCAMG00000012985	SINCAMP00000019840	
MITF	NP_937802.1	SINCAMG00000014146	SINCAMP00000021797	
NFATc1	NP_006153.2	SINCAMG00000007254	SINCAMP00000011077	
c-FOS	NP_005243.1	SINCAMG00000009790	SINCAMP00000015032	
Msx1	NP_002439.2	SINCAMG00000015571	SINCAMP00000023935	
Msx2	NP_002440.2	SINCAMG00000000703	SINCAMP00000001085	
Differentiation genes				
Matrilin-1	NP_002370.1	SINCAMG00000009781	SINCAMP00000015025	
Matrilin-3	NP_002372.1	SINCAMG00000007214	SINCAMP00000011003	
Col2a1	NP_001835.3	SINCAMG00000011525 SINCAMG00000000850	SINCAMP00000017882 SINCAMP00000001308	
Col10a1	NP_000484.2	SINCAMG00000004805	SINCAMP00000007366	
Col11a2 (XI alpha	NP_542411.2	SINCAMG00000008728	SINCAMP00000013503	

1a)				
Col1a1	NP_000079.2	SINCAMG00000013351	SINCAMP00000021079	
Col1a2	NP_000080.2	SINCAMG00000012400	SINCAMP00000019519	
Cox-2 (cyclooxygenase-2)	NP_000954.1	SINCAMG00000015316	SINCAMP00000023527	
Leptin receptor	NP_002294.2	SINCAMG00000010469	SINCAMP00000016069	
MMPI	NP_002412.1	SINCAMG00000001350	SINCAMP00000002074	
Osteocalcin / Bone gamma-carboxyglutamate protein	NP_954642.1	SINCAMG00000013388	SINCAMP00000020498	
Alkaline phosphatase	NP_000469.3	SINCAMG00000007089	SINCAMP00000010809	
Fam20C	NP_064608.2	SINCAMG00000004863 SINCAMG00000012860	SINCAMP00000007441 SINCAMP00000019666	
Chondroitin sulfate proteoglycan 4	NP_001888.2	SINCAMG00000008980 SINCAMG00000011254	SINCAMP00000013791 SINCAMP00000017285	
Chondroitin sulfate proteoglycan 5	NP_006565.2	KA353634	KA353634	Missing from assembly
galactosamine (N-acetyl)-6-sulfate sulfatase	NP_000503.1	SINCAMG00000010142	SINCAMP00000015577	
Heparan sulfate 2-O-sulfotransferase 1, 2, 3	NP_036394.1	SINCAMG00000008013 SINCAMG00000016544 SINCAMG00000016995 SINCAMG00000017642	SINCAMP00000012248 SINCAMP00000025558 SINCAMP00000026084 SINCAMP00000026733	SINCAMG00000008013 is syntenic to PKN2 and LMO4.
Collagenase 3/ Matrix	NP_002418.1	SINCAMG00000013249 SINCAMG00000013297	SINCAMP00000020318 SINCAMP00000020361	

metallopeptidase 13				
Bmp1	NP_001190.1	SINCAMG00000006960	SINCAMP00000010637	
Osteocrin	NP_937827.1	SINCAMG00000003547 SINCAMG00000007587	SINCAMP00000005401 SINCAMP00000011580	
Arachidonate 12-lipoxygenase/ALOX12	NP_000688.2	SINCAMG00000005263	SINCAMP00000008046	ALOX12 is SINCAMG00000005263 (scaffold_584), while ALOX5 is SINCAMG00000003391 (scaffold_126). Relationship is confirmed by NJ tree with human and zebrafish ALOX5 and 12.
Calcium-sensing receptor/CaSR	NP_001171536.1	SINCAMG00000010557	SINCAMP00000016223	
Osteopotentia homolog	NP_055098.1	SINCAMG00000015944	SINCAMP00000024593	
Sost/Sclerostin	NP_079513.1	SINCAMG00000012988	SINCAMP00000019846	
Sostdc1	NP_056279.1	SINCAMG00000008090	SINCAMP00000012373	
CRTAP/cartilage associated protein	NP_006362.1	SINCAMG00000006872	SINCAMP00000010489	
Phospho1	NP_001137276.1	SINCAMG00000006482	SINCAMP00000009844	
Phospho2	NP_001008489.1	SINCAMG00000000453	SINCAMP00000000731	
ANKH protein/ankylosis protein	NP_473368.1	SINCAMG00000008556	SINCAMP00000013099	
Atp2b1a	NP_001001323.1	SINCAMG00000013705	SINCAMP00000021203	
Cant1	NP_001153244.1	SINCAMG00000002262	SINCAMP00000003453	
PHEX	NP_000435.3	SINCAMG00000014306	SINCAMP00000022041	
CD44	NP_000601.3	SINCAMG00000009938	SINCAMP00000015264	
Chondroadherin	NP_001258.2	SINCAMG00000004002 SINCAMG00000011743	SINCAMP00000006076 SINCAMP00000018016	
ENPP1 - Ectonucleotide	NP_006199.2	SINCAMG00000006314	SINCAMP00000009586	Labelled as ENPP3 in the genome browser, but confirmed to be ENPP1 by NJ tree.

pyrophosphatase/phosphodiesterase family member 1				
Entpd5	NP_001240.1	SINCAMG00000002450	SINCAMP00000003761	
Sptbn1	NP_003119.2	SINCAMG00000006476 SINCAMG00000013852	SINCAMP00000009994 SINCAMP00000021579	
Adamts18	NP_955387.1	SINCAMG00000012186	SINCAMP00000018694	
Rspo3	NP_116173.2	SINCAMG00000006603	SINCAMP00000010032	
Galnt3	NP_004473.2	SINCAMG00000000219	SINCAMP00000000360	
Fam3c	NP_001035109.1	SINCAMG00000009455	SINCAMP00000014556	
Xylt1	NP_071449.1	SINCAMG00000015951	SINCAMP00000024585	
Ext1	NP_000118.2	SINCAMG00000000630	SINCAMP00000000984	Conserved synteny with MED30 and SAMD12
Ext2	NP_000392.3	SINCAMG00000007585	SINCAMP00000011587	
Extl3	NP_001431.1	SINCAMG00000009719	SINCAMP00000014926	Conserved synteny with FZD3, INTS9, HMBOX1
Papst1	NP_835361.1	SINCAMG00000007847	SINCAMP00000011993	
Uxs1	NP_001240804.1	SINCAMG00000009298	SINCAMP00000014319	
Has2	NP_005319.1	SINCAMG00000006923	SINCAMP00000010549	
Mgp	NP_001177768.1	SINCAMG00000013395	SINCAMP00000020504	Conserved synteny with ERP27 and PDE6H. Adjacent to osteocalcin/BGLAP.
Osteoclast regulators				
EGR1	NP_001955.1	SINCAMG00000008709	SINCAMP00000013366	
CSF1R	NP_005202.2	SINCAMG00000010472	SINCAMP00000016088	
M-CSF/CSF/GM-CSF	NP_000748.3	-	-	
RANK(TNFRSF11A)	NP_003830.1	SINCAMG00000003288 SINCAMG00000009035 SINCAMG00000009040	SINCAMP00000005005 SINCAMP00000013887 SINCAMP00000013891	

RANKL (TNFSF11)	NP_003692.1	SINCAMG00000001809	SINCAMP00000002766	Labelled as TNFSF10-like but confirmed by synteny with AKAP11 and EPSTI1.
TNFRSF11B/OPG	NP_002537.3	SINCAMG00000001571	SINCAMP00000002408	
Mcp1	NP_002973.1	JW879915 SINCAMG00000014559	JW879915 SINCAMP00000022338	Two possible orthologs that resemble both Mcp1 (CCL2) and Mcp3 (CCL7).
CCR2	NP_001116513.2	SINCAMG00000008234 SINCAMG00000009465 SINCAMG00000009559	SINCAMP00000012599 SINCAMP00000014568 SINCAMP00000014697	
CCR4	NP_005499.1	SINCAMG00000006551 SINCAMG00000014673 SINCAMG00000016780	SINCAMP00000009950 SINCAMP00000022504 SINCAMP00000025869	
Osteoclast stimulating factor 1	NP_036515.4	SINCAMG00000005702	SINCAMP00000008714	
SPARC, SPARCL1 and related SCPP genes				
Sparc/Osteonectin	NP_003109.1	SINCAMG00000007163	SINCAMP00000010957	
Sparcl1	NP_001121782.1	SINCAMG00000010206	SINCAMP00000015668	
SIBLING				
DSPP	NP_055023.2	Not found		
DMP1	NP_004398.1	Not found		
IBSP	NP_004958.2	Not found		
MEPE	NP_001171623.1	Not found		
SPP1 (OPN)	NP_001238759.1	Not found		
P/Q rich SCPP (Enamel)				
AMEL	NP_872621.1	Not found		
ENAM	NP_114095.2	Not found		
AMBN	NP_057603.1	Not found		
AMTN	NP_997722.1	Not found		

ODAM (APIN)	NP_060325.3	Not found		
Proteoglycans				
Fibronectin	NP_002017.1	SINCAMG00000001164	SINCAMP00000001858	
leprecan	NP_001230175.1	SINCAMG00000006696	SINCAMP00000010226	
Tuftelin	NP_064512.1	SINCAMG00000008225	SINCAMP00000012588	
Podocan	NP_714914.2	SINCAMG00000002366 SINCAMG00000013633	SINCAMP00000003624 SINCAMP00000020895	
SLRP gene family				
Fibromodulin	NP_002014.2	SINCAMG00000009120 SINCAMG00000009122	SINCAMP00000014034 SINCAMP00000014037	Possible duplicates located 1 gene apart in opposite orientation.
Prolargin	NP_002716.1	SINCAMG00000009123	SINCAMP00000014040	
Opticin	NP_055174.1	SINCAMG00000009121	SINCAMP00000014035	Based on synteny with FMOD, PRELP and ATP2B4
Extracellular matrix protein 2	NP_001384.1	SINCAMG00000014021	SINCAMP00000021532	
Asporin	NP_060150.4	KC707914	KC707914	
Osteomodulin	NP_005005.1	KC707915	KC707915	
Osteoglycin/Mimecan	NP_148935.1	SINCAMG00000014048	SINCAMP00000021567	
Decorin	NP_001911.1	SINCAMG00000013640	SINCAMP00000020961	SINCAMG00000013640 is syntenic to BTG1 and LUM
Lumican	NP_002336.1	SINCAMG00000013688	SINCAMP00000020965	
Keratocan	NP_008966.1	SINCAMG00000013691	SINCAMP00000020972	
Epiphycan	NP_004941.2	SINCAMG00000013696	SINCAMP00000020978	
ECM2L	XP_003960142.1	-	-	
Biglycan	NP_001702.1	JW874743	JW874743	Missing from assembly
LecticanHAPLN gene family				
Hyaluronan and	NP_068589.1	JW872610	JW872610	

proteoglycan link protein 2				
Brevican	NP_068767.3	KC707909	KC707909	
Versican	NP_004376.2	SINCAMG00000004695	SINCAMP00000007216	
Hyaluronan and proteoglycan link protein 1	NP_001875.1	SINCAMG00000004694	SINCAMP00000007214	
Aggrecan	NP_037359.3	SINCAMG00000011028	SINCAMP00000016953	
Hyaluronan and proteoglycan link protein 3	NP_839946.1	SINCAMG00000011037	SINCAMP00000016955	
Neurocan	NP_004377.2	SINCAMG00000006847	SINCAMP00000010414	
Hyaluronan and proteoglycan link protein 4	NP_075378.1	SINCAMG00000005547	SINCAMP00000008481	

Supplementary Table XI.1 | Genes involved in antigen presentation

	H. sapiens gene	*H. sapiens* protein	*C. milii* protein	Remarks
	CTSS	ENSP00000357981		not detected in *C. milii* and *G. cirratum* databases
	CTSL1	ENSP00000345344	SINCAMP00000014412	Partial sequence is found in Scaffold_202: 800,531-813,636, but there are ~8 genes found in Scaffold_85 as well, one of them may be CTSL2. A total of ~13 genes are present in the *C. milii* genome
	CTSL2	ENSP00000259470		see CTSL1
	PSMB10	ENSP00000351314	SINCAMP00000003834	Scaffold_1084: 19,439-14,269
	ERAP1	ENSP00000296754		See ERAP2
	ERAP2	ENSP00000400376	SINCAMP00000019922	There are three genes in tandem on scaffold_27: 3,636,463-3,650,610; 3,655,296-3,672,395; 3,681,159-3,704,136. Human ERAP1 and ERAP2 genes are also in tandem.
	TAP1	ENSP00000346206	SINCAMP00000021912	Scaffold_5491:31-3,636
	TAP2	ENSP00000364032		Not present in *C. milii* databases, but present in *G. cirratum* transcriptome (KC814638)
	LGMN	ENSP00000334052	SINCAMP00000005108	Scaffold_114: 1,861,488-1,851,553
	IFI30	ENSP00000384886	SINCAMP00000009088	Scaffold_64: 3,752,082-3,755,990
	POMP	ENSP00000370222	SINCAMP00000014238	Scaffold_81: 626,392-610,628
	PSME3	ENSP00000466794	JW876569 SINCAMP00000019935	Scaffold_256: 4,028-929
	PSME1	ENSP00000372155	JX052270	Scaffold_18508: 459-27
	PSME2	ENSP00000216802	JW876766	split into three short scaffolds_ 15485: 437-1,033; 14587: 1110-237; 12568: 1,228-1,142
	PSMB8	ENSP00000406878	SINCAMP00000003812	scaffold_1084:966-8810
	PSMB5	ENSP00000355325	JW876303	AAVX01617944.1
	PSMB6	ENSP00000270586	SINCAMG00000011687	Possibly pseudogene on scaffold_4135:898-1,604; but see JW876745
	PSMB9	ENSP00000363993		Not detected in *C. milii* databnases, but present in *G. cirratum* transcriptome (KC814632)
	PSMB11	ENSP00000386212	SINCAMG00000014611 SINCAMG00000002681	possibly pseudogene; two scaffolds were identified (Scaffold_2497: 1,191-1,875; Scaffold_1219: 9,744-10,417); not present in transcriptomes
	ACE	ENSP00000290866	SINCAMP00000003105	Scaffold_185: 138,164-115,875
	CD74	ENSP00000430614	SINCAMP00000016296	Scaffold_204: 654,822-649,039

Supplementary Table XI.2 | Pathogen receptors and intracellular signalling components

H. sapiens gene	*H. sapiens* protein	*C. milii* protein	Remarks
TLR-like receptors			
TLR1	ENSP00000421259	SINCAMP00000002713	TLR1 or TLR6 or TLR10
		SINCAMP00000025836	TLR1 or TLR6 or TLR10
TLR2	ENSP00000260010	SINCAMP00000026272	TLR2 or TLR5
		SINCAMP00000026273	TLR2 or TLR5
TLR3	ENSP00000296795	SINCAMP00000019722	
TLR4	ENSP00000363089		not detected; pseudogene fragment related to TLR4 located in syntenic region on scaffold_45 between DBC1 and ASTN2 at nt ~ 3,886,000
TLR5	ENSP00000440643		see TLR2
TLR6	ENSP00000371376		see TLR1
TLR7	ENSP00000370034	SINCAMP00000026415	
TLR8	ENSP00000312082	SINCAMP00000026416	
TLR9	ENSP00000353874	SINCAMP00000004690	
TLR10	ENSP00000308925		see TLR1
		SINCAMP00000021648	Unclear orthology
		SINCAMP00000026056	Unclear orthology
Signalling molecules			

LBP	ENSP00000217407		Not detected; however, several BPI homologs are present:
			SINCAMG00000009082;
			SINCAMG00000009082;
			SINCAMG00000008957;

CD14	ENSP00000304236		not detected in region syntenic with human genome (scaffold_92: at nt ~ 40,000)
MD-2	ENSP00000284818		not detected in region syntenic with human genome (scaffold_7: at nt ~ 2,000,000)
MYD88	ENSP00000379625	SINCAMP00000017569	
TICAM1/TRIF	ENSP00000248244	SINCAMP00000010683	Scaffold_83: 2,123,051-2,120,931
TICAM2/TRAM	ENSP00000415139	KC707910	Scaffold_54: 3,008,897-3,009,622
TIRAP	ENSP00000279992	SINCAMP00000018345	Scaffold_44: 97,453-99,641
SARM	ENSP00000406738	SINCAMP00000021487	
TRAF6	ENSP00000337853	SINCAMP00000000213	
TMEM173/STING	ENSP00000331288	SINCAMP00000000648 SINCAMP00000000649	Scaffold_156: 1,377,664-1,370,681 (SINCAMP00000000648 and SINCAMP00000000649 are the same sequence)
RIPK2	ENSP00000220751	SINCAMP00000000865	Scaffold_105: 557,246-512,111

CARD8	ENSP00000428883		not detected
NAIP	ENSP00000429839	SINCAMP00000012328	Scaffold-91: 3,291,033-3,298,496
IRAK1	ENSP00000358997	SINCAMP00000006958	Scaffold_15: 8,101,045-8,089,531
IRAK4	ENSP00000390651	SINCAMP00000002110	Scaffold_133: 1,251,110-1,243,329
TBK1	ENSP00000329967	JW863716	Scaffold_39: 770,168-786,498
IRF3	ENSP00000470431	JW870582	split into two scaffolds; Scaffold_8024:1,865-408 & Scaffold_4730: 3,040-1,707
IRF7	ENSP00000329411	SINCAMP00000012194	Scaffold_8024: 1,874-423
MALT1	ENSP00000376445	SINCAMP00000001082	Scaffold_103: 1,483,317-1,507,222
TANK	ENSP00000376505	SINCAMP00000000233	Scaffold_14: 499,943-515,116

Supplementary Table XI.3 | Intracellular pathogen receptors and inflammasome components

H. sapiens gene	*H. sapiens* protein	*C. milii* protein	Remarks
Intracellular pathogen receptors			
DHX58	ENSP00000369213	SINCAMP00000020725	Scaffold_256: 384,966-392,934
IFIH1	ENSP00000263642	SINCAMP00000000285	
DDX58	ENSP00000251642	SINCAMP00000000024	
NOD1	ENSP00000222823	SINCAMP00000003714	
		SINCAMP00000003723	Similar to SINCAMP00000003714
NOD2	ENSP00000300589		Unclear orthology; possibly SINCAMP00000003723
Inflammasome components			
NLRP1	ENSP00000324366	SINCAMP00000025689	assignment not supported by synteny; sequence possibly incomplete
NLRP3	ENSP00000337383	SINCAMP00000014888 SINCAMP00000015962 SINCAMP00000000673 SINCAMP00000010530 JW863066 JW863420 JW863446 JW864590 JW864704 JW874152 JW879245	Altogether 57 NLRP3-like genes were identified. Only11 representatives are listed here.
IPAF/NLRC4	ENSP00000354159		Not detected in *C. milii*, but present in *G. cirratum* transcriptome: KC814625
NLRC5	ENSP00000262510	SINCAMP00000005561	Scaffold_189: 586,610-564,330
AIM2	ENSP00000357112		Not detected
ASC/PYCARD	ENSP00000247470	KA353642 JW877662	Scaffold_57: 1,447,688-1,447,419
CASP1	ENSP00000410076	SINCAMP00000018377	Scaffold_47: 296,823-301,237
IL1B	ENSP00000263341	SINCAMP00000003504	
IL18	ENSP00000280357	SINCAMP00000011586	

Supplementary Table XI.4 | Genes involved in the complement system

H. sapiens gene	*H. sapiens* protein	*C. milii* protein	Remarks
Substrate binding proteins			
C1QA	ENSP00000363773	JW875837	Gene assignment is arbitrary; There are three C1q genes in tandem in Scaffold_183 (SINCAMP00000001380- 317,277-316,155; SINCAMP00000001383-310,974-309,945; SINCAMP00000001381- 306,228-304,489)
C1QB	ENSP00000313967	SINCAMP00000001383 JW870590	See C1QA
C1QC	ENSP00000363770	SINCAMP00000001381 JX052985	See C1QA
C3	ENSP00000245907	SINCAMP00000012624	Scaffold_738: 9,385-36,070
C4A	ENSP00000396688	SINCAMP00000015348	There are two C4 genes and their assginment is arbitrary; Scaffold_276: 34,055-86,869
C4B	ENSP00000415941	SINCAMP00000014484	See C4A for gene assignment; Scaffold_296: 341,703-370,982
MBL2	ENSP00000363079		not detected
Ligands			
C1R	ENSP00000290575	SINCAMP00000015476 SINCAMP00000018513	There are two C1r genes that are ~28% similar to each other. SINCAMP00000015476 is found in Scaffold_2144: 4,543-10,376, while SINCAMP00000018513 is found in Scaffold_387: 3,561-14,620.
C1RL	ENSP00000266542		not detected
C1s	ENSP00000385035	SINCAMP00000016215 JW866600	There are two C1s genes that are ~31% similar to each other.; SINCAMP00000016215 is located on Scaffold_290: 382,925-373,480.
C2	ENSP00000299367	JW864721	Part of this sequence found in AAVX01271959.1; best BLAST hit on reverse blast with human C2
BF	ENSP00000410815	JW865377	There are two Bf-like (including JW864721) assigned for C2/Bf based on the similarity to *G. cirratum* sequences; a fragment of JW865377 is found on Scaffold_16329: 560-384.
CFD	ENSP00000332139		
MASP1	ENSP00000336792	SINCAMP00000018490	
MASP2	ENSP00000383690	SINCAMP00000010528	assignment supported by synteny
Membrane attack complex			
C5	ENSP00000223642	SINCAMP00000001718	Scaffold_100: 178,210-351,038

C6	ENSP00000263413	SINCAMP00000014669	Scaffold_2701: 8,328-155
C7	ENSP00000322061	JW870687	Possibly split into two scaffolds; Scaffold_15155: 75-308 & Scaffold_13170: 1,243-609.
C8A	ENSP00000354458	SINCAMP00000003517	Scaffold_112: 2,367,520-2,358,153
C8B	ENSP00000360281	SINCAMP00000003507	Scaffold_112: 2,346,130-2,356,566; C8A and B genes are found in tandem in tail-to-tail orientation in both *C. milii* and human.
C8G	ENSP00000224181	SINCAMP00000008133	Scaffold_77: 951,256-947,097
C9	ENSP00000263408	JW867730	Split into three scaffolds: Scaffold_666: 1,069-899 & Scaffold_9346: 57-1,362 & Scaffold_7982: 962-1,108
Regulatory proteins			
CFP	ENSP00000380189	JW870148	Not found on scaffolds
SERPING1	ENSP00000278407	JW873856	Scaffold_4727: 268-4,017
CFI	ENSP00000378130	SINCAMP00000023069	Scaffold_269: 99,328-79,965
CFH	ENSP00000356399	SINCAMP00000025042	In addition to CFH, the human genome encodes a further 5 CFH-related genes (CFHR1-CFHR5); orthology is uncertain
C4BPA	ENSP00000356037	SINCAMP00000014228	
C4BPB	ENSP00000243611		
CD46	ENSP00000350893	SINCAMP00000014229 SINCAMP00000026655	There seems to be multiple (>5) genes in Scaffold_70: some of them are located 3,545,903-3,550,226; 3,529,871-3,561943; and between 3,494,241 and 3,561,943. It is not clear all of them are CD46 orthologue.
CD55/DAF	ENSP00000356031	SINCAMP00000014220	
CD59	ENSP00000340210		not detected in *C. milii* databases; present in *G. cirratum* transcriptome (KC814621)
C1qBP	ENSP00000225698	SINCAMP00000021701	
CLU	ENSP00000315130	SINCAMP00000023969	
VTN	ENSP00000226218	SINCAMP00000021550 ENSP00000226218	possibly two VTN-like genes
Complement/anaphylatoxin receptors			
CR1/CD35	ENSP00000383744		
CR1L	ENSP00000421736		
CR2/CD21	ENSP00000356025		
ITGAM/CD11b	ENSP00000441691	JW862483	ITGAM and ITGAX are very similar; assignment of the *C. milii* transcript is arbitrary
ITGAX/CD11c	ENSP00000268296	JW862389	see ITGAM

ITGB2	ENSP00000347279	SINCAMP00000006134	
C3AR1	ENSP00000302079		see C5AR1
C5AR1/CD88	ENSP00000347197	SINCAMP00000026154	annotated as C3AR1 in *C. milii* assembly; orthology unclear

Supplementary Table XI.5 | Chemokines in *C. milii*

H. sapiens gene	*H. sapiens* protein	*C. milii* protein	Remarks
CC chemokines			
CCL19	ENSP00000368077	SINCAMP00000007281	CCL19-like?; scaffold_618: 28,517-33,903
		SINCAMP00000025569	CCL19-like ?; scaffold_1: 12,442,330-12,445,534
		SINCAMP00000025570	CCL19-like ?; scaffold_1: 12,451,296-12,454,760
		SINCAMP00000025571	CCL19-like ?; scaffold_1: 12,461,823-12,464,611
		SINCAMP00000007279	CCL19-like ?; scaffold_618: 18,302-21,011
CCL20	ENSP00000351671	SINCAMP00000025574	CCL20-like ?; scaffold_1: 12,502,559-12,504,326
		SINCAMP00000025573	CCL20-like ?; scaffold_1: 12,489,660-12,491,794
		SINCAMP00000025572	CCL20-like ?; scaffold_1: 12,475,065-12,476,676
CCL24	ENSP00000222902	SINCAMP00000022338	
CCL25	ENSP00000375086	SINCAMP00000008771	
		SINCAMP00000025566	Unclear homology; scaffold_1: 12,404,817-12,408,740
		SINCAMP00000025567	Unclear homology; scaffold_1: 12,412,827-12,415,963
		SINCAMP00000025568	Unclear homology; scaffold_1: 12,418,635-12,421,525
		SINCAMP00000014990	Unclear homology; scaffold_2132: 1,130-2,917
CXC chemokines			
CXCL8	ENSP00000306512	SINCAMP00000013346	
CXCL12	ENSP00000379140	SINCAMP00000025277	scaffold_3: 14,257,952-14,263,963
		SINCAMP00000025276	scaffold_3: 14,224,143-14,241,859; duplicated version of CXCL12 gene
CXCL14	ENSP00000337065	SINCAMP00000007926	
CXCL16	ENSP00000460145	SINCAMP00000017269	
		SINCAMP00000012935	scaffold_689: 5,738-8,378; unclear homology
		SINCAMP00000012947	scaffold_689: 15,580-17,483; unclear homology
		SINCAMP00000014691	scaffold_2246: 7,195-8,941; unclear homology
		INCAMP00000016228	scaffold_19: 7,328,783-7,340,338; unclear homology
		SINCAMP00000016232	scaffold_19: 7,346,620-7,349,535; unclear homology
		SINCAMP00000016234	scaffold_19: 7,355,924-7,358,930; unclear homology
		SINCAMP00000016225	scaffold_19: 7,281,632-7,284,526; unclear

			homology
		SINCAMP00000025268	scaffold_3: 14,170,157-14,173,501; related to CXCL9/CXCL10
		SINCAMP00000025270	scaffold_3: 14,178,344-14,181,991; related to CXCL9/CXCL10
		SINCAMP00000008162	scaffold_87: 3,288,334-3,291,514; related to CXCL11/CXCL14
		SINCAMP00000008169	scaffold_87: 3,296,328-3,298,495; related to CXCL11/CXCL14
		SINCAMP00000008173	scaffold_87: 3,3 01,114-3,303,079; related to CXCL11/CXCL14

Blue: homeostatic chemokines; **Red**: inflammatory chemokines **Green**: dual-function chemokines.

No unambiguous orthologues of human CCL1, CCL2 ,CCL3, CCL3L1, CCL3L3, CCL4, CCL4L1, CCL4L2, CCL5, CCL7, CCL8, CCL11,CCL13, CCL14, CCL15, CCL16, CCL17, CCL18, CCL21, CCL22, CCL23, CCL24, CCL26,CCL27, CCL28, CXCL1, CXCL2, CXCL3, CXCL4, CXCL4L1, CXCL5, CXCL6, CXCL7, CXCL9, CXCL10, CXCL11, CXCL13, CXCL17, XCL1, XCL2, CX3CL1 genes could be detected in *C. milii* genome.

Supplementary Table XI.6 | Chemokine receptors in *C. milii*

H. sapiens gene	*H. sapiens* protein	*C. milii* protein	Remarks
CCR4	ENSP00000332659	SINCAMP00000025869	Closely related to CCR8
CCR6	ENSP00000339393	SINCAMP00000023769	
CCR7	ENSP00000246657	SINCAMP00000017205	
CCR9	ENSP00000348260	SINCAMP00000009969	
CXCR1	ENSP00000295683	SINCAMP00000026575	Closely related to CXCR2; human genes are situated next to each other on chromosome 2; in *C. milii*, these two genes occur on different scaffolds (scaffold_407: 127,075-130,083 [SINCAMP00000026575] and scaffold_1109: 1-5,040 [SINCAMP00000002221].
CXCR2	ENSP00000319635		See CXCR1
CXCR4	ENSP00000241393	SINCAMP00000026434	
CXCR5	ENSP00000292174	SINCAMP00000018503	Two CXCR5-like genes exist in *C. milii*
		SINCAMP00000026400	
CXCR6	ENSP00000304414	SINCAMP00000025870	Two CXCR6-like genes exist in *C. milii* (scaffold_83: 374,347-375,432 [SINCAMP00000025870] and scaffold_88: 1,006,050-1,007,186 [SINCAMP00000025828]
CXCR7	ENSP00000272928	SINCAMP00000006060	
CCRL1	ENSP00000249887	SINCAMP00000025929	
CCBP2	ENSP00000273145	SINCAMP00000022504	
XCR1	ENSP00000438119	SINCAMP00000025868	Two XCR1-like genes exist in *C. milii* (scaffold_83: 201,892-202,767 [SINCAMP00000025868] and scaffold_83: 314,547-319,885 [SINCAMP00000009950]).
		SINCAMP00000026574	This gene on scaffold_407: 122,146-123,873 has no clear homologue in humans

Blue: receptors for homeostatic chemokines; **Red**: receptors for inflammatory chemokines;

Green: receptors for homeostatic, inflammatory and/or dual-function [homeostatic/inflammatory]

Chemokines. No clear orthologues could be detected for human CCR1, CCR2, CCR3, CCR5, CCR8

(closely related to CCR4), CXCR3, CX3CR1, CCRL2, DUFFY/DARC.

Supplementary Table XI.7a | Overview: Cytokines and interleukins in *C. milii*

Receptor Composition			Ligand(s)
Component 1	Component 2	Component 3	
IL2RG	IL2RB		IL2
IL2RG	IL2RB	IL2RA	IL2
IL2RG	IL2RB	IL15RA	IL15
IL2RG	IL4R		IL4
IL2RG	IL7R		IL7
IL2RG	IL9R		IL9
IL2RG	IL21R		IL21
	IL7R	CRLF2	TSLP
IL13RA1			IL13
IL13RA2			IL13
IL13RA1	IL4R		IL4, IL13
IL6ST	IL6R		IL6
IL6ST	IL11RA		IL11
IL6ST	IL27RA		IL27 (EBI3+IL27)
IL6ST	LIFR		LIF, CTF1
IL6ST	OSMR		OSM, IL31
IL6ST	LIFR	CNTFR	CNT (CLCF1+CLF1)
IL31R	OSMR		IL31
CSF1R			CSF1, IL34
CSF2RB	CSF2RA		CSF2
CSF2RB	IL3RA		IL3
CSF2RB	IL5RA		IL5
CSF3R			CSF3
IL10RB	IL10RA		IL10
IL10RB	IL28RA		IL28A, IL28B, IL29
IL10RB	IL22RA1		IL22
IL10RB	IL20RA		IL26
IL20RB	IL20RA		IL20, IL24, IL19
IL20RB	IL22RA1		IL24
IL12RB1	IL12RB2		IL12 (IL12B+IL12A)

IL12RB1	IL23R		IL23 (IL12B+IL23A)
FTL3			FLT3LG
KIT			KITLG

Colour code: green, orthologue confidently identified; orange, orthologue present in transcriptome database, but not in genome assembly; blue, orthologue may be present, but classification uncertain; grey, not detected and syntenic region also not identified; red, not detected and absent from syntenic region.

Supplementary Table XI.7b | Interleukin and cytokine receptors

H. sapiens gene	*H. sapiens* protein	*C. milii* protein	Remarks
Receptors			
IL1R1	ENSP00000386380	SINCAMP00000014355	
IL1RAP	ENSP00000072516	SINCAMP00000014578	
IL1R2	ENSP00000330959		
IL2RA	ENSP00000369293		likely absent, see IL15RA; not found in *G. cirratum* transcriptome
IL2RB	ENSP00000216223	SINCAMP00000015978	
IL2RG	ENSP00000363318	SINCAMP00000025339 SINCAMP00000025342	local duplication on scaffold_2: 16,442,958-16,450,451 (SINCAMP00000025339) and scaffold_2: 16,451,578-16,464,056 (SINCAMP00000025342); a similar situation occurs in *G. cirratum* (transcripts KC814626 and KC814631)
IL3RA	ENSP00000327890		may be present on scaffold_18 (see below)
IL4R	ENSP00000379111	SINCAMP00000006430	scaffold_147: 1,516,756-1,525,752; flanking genes, including IL21R are conserved; however, a third interleukin receptor is present
IL5RA	ENSP00000256452		may be present on scaffold_18 (see below)
IL6R	ENSP00000357470		
IL6ST	ENSP00000370698	SINCAMP00000017100 SINCAMP00000017111 SINCAMP00000017124	scaffold_22: 5,312,432-5,354,496 (SINCAMP00000017100); scaffold_22: 5,394,661-5,417,010 (SINCAMP00000017111); scaffold_22: 5,418,018-5,464,539 (SINCAMP00000017124); all annotated as IL6ST. This assignment is supported by synteny with the human IL6ST/IL31RA cluster, but three cytokine receptor-like genes are present on scaffold_22.
IL7R	ENSP00000306157	JW868881	So far, identified in *C. milii* transcript library only; also present in *G. cirratum* transcript library (KC814633)
IL9R	ENSP00000358431		Candidate on scaffold_147: 1,564,089-1,567,854 (SINCAMP00000006439); although annotated as IL9R, this is not supported by synteny. However, BLASTp against human sequences suggests sequence is related to IL21R, IL9R and to a lesser extent IL2RB.
IL10RA	ENSP00000227752	SINCAMP00000020946	
IL10RB	ENSP00000290200	SINCAMP00000004357	
IL11RA	ENSP00000326500	SINCAMP00000001165	
IL12RB1	ENSP00000314425	SINCAMP00000009045	
IL12RB2	ENSP00000262345	SINCAMP00000015780	scaffold_252: 389-13,922; assignment supported by sequence and sytenic genes on one side; gene is located at the end of the scaffold; in the human genome, the IL23R gene is linked to the IL12RB2 gene.
IL13RA1	ENSP00000360730	SINCAMP00000022206	scaffold_2: 9,098,254-9,108,206; annotated as

H. sapiens gene	*H. sapiens* protein	*C. milii* protein	Remarks
			IL13RA2, but based on BLAST results and linkage to DOCK11 and WDR44, this is most likely IL13RA1.
IL13RA2	ENSP00000361004	SINCAMP00000021159	scaffold_2: 8,615,795-8,627,435; although annotated as IL13RA1, BLAST results and linkage to LRCH2 and HTR2A/C suggest that this is most likely IL13RA2. The predicted protein is longer than human IL13RA2, and may have signalling capability.
IL15RA	ENSP00000369312	SINCAMP00000000976	scaffold_17: 388,812-420,074; likely represents IL15RA (based on sequence), but region exhibits synteny to the human locus which contains both IL15RA and IL2RA; a related sequence is found in the *G. cirratum* transcriptome (KC814623).
IL17RA	ENSP00000320936	SINCAMP00000015994	
IL17RB	ENSP00000288167	SINCAMP00000005730	
IL17RC	ENSP00000295981	SINCAMP00000018499	
IL17RD	ENSP00000296318	SINCAMP00000007112	
IL17RE	ENSP00000295980		
IL17REL	ENSP00000342520	SINCAMP00000001567	
IL20RA	ENSP00000314976	SINCAMP00000024622 SINCAMP00000019428	scaffold_26: 6,773,021-6,792,900 (SINCAMP00000024622); assignment is supported by sequence and synteny. A potential second IL20RA-like gene appears on scaffold_2: 4,260,224-4,278,160 (SINCAMP00000019428); annotated as IL20RA, but there is no synteny conservation with the human IL20RA locus; BLASTp suggest this to be additional copy of either IL20RA or IL22R. See also analysis of IFN gene family.
IL20RB	ENSP00000328133	SINCAMP00000025423	scaffold_1: 9,304,161-9,320,889; annotated as IL20RB, but there is no synteny conservation with the human IL20RA locus.
IL21R	ENSP00000338010	SINCAMP00000006436	
IL22RA1	ENSP00000270800	SINCAMP00000014733	
IL22RA2	ENSP00000296980	SINCAMP00000024615 SINCAMP00000024618	scaffold_26: 6,748,991-6,751,747 (SINCAMP00000024615); scaffold_26: 6,758,402-6,762,933 (SINCAMP00000024618); both annotated as IL22RA2, which is supported by synteny, but a second IL22RA2-like gene appears to be present. These genes are part of the human locus that contains the IFNGR1/IL22RA2/IL20RA genes.
IL23R	ENSP00000321345		
IL27RA	ENSP00000263379	SINCAMP00000016756	annotated as IL12RB2, but linked gene RLN3 suggests that this gene may be the equivalent of human IL27R.
IL31RA	ENSP00000415900		likely absent, but potentially present on scaffold_22 (see IL6ST)
CNTFR	ENSP00000368265	SINCAMP00000001166	
CRLF1	ENSP00000376188	SINCAMP00000008821	
CRLF2 (TSLPR)	ENSP00000383641		may be present on scaffold_18 (see below)

H. sapiens gene	*H. sapiens* protein	*C. milii* protein	Remarks
CSF1R	ENSP00000286301	SINCAMP00000016088	
CSF2RA	ENSP00000394227		may be present on scaffold_18 (see below)
CSF2RB	ENSP00000262825	SINCAMP00000016009	
CSF3R	ENSP00000362195		absent from syntenic region, potential hit on scaffold_64: 3,989,408-3,997,431 (SINCAMP00000009149)
FLT3	ENSP00000241453	SINCAMP00000014257	
KIT	ENSP00000288135	SINCAMP00000018568	
LIFR	ENSP00000263409	SINCAMP00000020153	
OSMR	ENSP00000274276		not found; this gene is linked to LIFR in the human genome; however, scaffold_329 is very short
		SINCAMP00000004531	scaffold_18: 373,875-396,948; annotated as IL5RA, located in a cluster of 5 cytokine receptor-like genes, in a region of strongly conserved synteny to human CRLF2/CSF2RA/IL3RA locus.
		SINCAMP00000004537	scaffold_18: 406,745-415,193; annotated as IL2RG, located in a cluster of 5 cytokine receptor-like genes, in a region of strongly conserved synteny to human CRLF2/CSF2RA/IL3RA locus.
		SINCAMP00000004543	scaffold_18: 428,106-443,452; annotated as IL2RG, located in a cluster of 5 cytokine receptor-like genes, in a region of strongly conserved synteny to human CRLF2/CSF2RA/IL3RA locus.
		SINCAMP00000004549	scaffold_18: 522,558-534,051; annotated as CSF2RA, located in a cluster of 5 cytokine receptor-like genes, in a region of strongly conserved synteny to human CRLF2/CSF2RA/IL3RA locus.
		SINCAMP00000004574	scaffold_18: 559,144-658,702; annotated as IL2RG, located in a cluster of 5 cytokine receptor-like genes, in a region of strongly conserved synteny to human CRLF2/CSF2RA/IL3RA locus.

Supplementary Table XI.7c | Interleukins, cytokines and ligands

H. sapiens gene	*H. sapiens* protein	*C. milii* protein	Remarks
Ligands			
CLCF1	ENSP00000434122		
CNTF	ENSP00000355370		
CRLF1	ENSP00000376188	SINCAMP00000008821	
CSF1	ENSP00000327513		
CSF2	ENSP00000296871		
CSF3	ENSP00000225474	SINCAMP00000017148	scaffold_251: 60,582-61,495; partial sequence only, but assignment supported by synteny.
CTF1	ENSP00000279804		
FLT3LG	ENSP00000204637		
KITLG	ENSP00000228280	SINCAMP00000021317	
IL1A	ENSP00000263339		No BLAST hit
IL1B	ENSP00000263341	SINCAMP00000003504	
IL2	ENSP00000226730		Absent from syntenic region; no BLAST hit; also absent from *G. cirratum* transcriptomes.
IL3	ENSP00000296870		Absent from syntenic region, but scaffold has gaps
IL4	ENSP00000231449		Absent from syntenic region; no BLAST hit
IL5	ENSP00000231454		Absent from syntenic region; no BLAST hit
IL6	ENSP00000385675	SINCAMP00000024395 SINCAMP00000024402	Duplicated in elephant shark
IL7	ENSP00000263851	SINCAMP00000024043	identified by synteny (IL7 gene overlaps the ZC2HC1A gene); identified in *G. cirratum* transcriptome (KC814622)
IL9	ENSP00000274520		Absent from syntenic region; no BLAST hit
IL10	ENSP00000412237	SINCAMP00000014174	scaffold_70: 3,215,896-3,221,395; note that this scaffold contains a total of 3 cytokine genes, while the equivalent human locus contains 4 cytokine genes (IL10, IL19, IL20 and IL24).
IL11	ENSP00000264563		
IL12A	ENSP00000303231		Absent from syntenic region; no BLAST hit
IL12B (p40)	ENSP00000231228	SINCAMP00000015768 SINCAMP00000018511	scaffold_19: 3,370,635-3,375,004 (SINCAMP00000015768); assignment supported by one closely linked gene and generally conserved synteny. scaffold_198: 312,760-316,682 appears to contain a second version of Il12B (SINCAMP00000018511); this copy is located on a short scaffold and linked to the DOCK2, FAM196B genes, which are also present on scaffold_19.
IL13	ENSP00000304915		Absent from syntenic region
IL15	ENSP00000296545	SINCAMP00000011712	Also identified in *G. cirratum* transcriptome (KC814629)

IL17A	ENSP00000344192		IL17A and IL17F are very similar; two related genes are found in the *C. milii* genome (SINCAMP00000016734 and SINCAMP00000017945)
IL17B	ENSP00000261796		IL17B and IL17D are very similar; two related genes are found in *C. milii* (SINCAMP00000019791 and SINCAMP00000004112)
IL17C	ENSP00000244241	SINCAMP00000015534	
IL17D	ENSP00000302924		IL17B and IL17D are very similar; two related genes are found in *C. milii* (SINCAMP00000019791 and SINCAMP00000004112)
IL17E/IL25	ENSP00000328111		not detected
IL17F	ENSP00000337432		IL17A and IL17F are very similar; two related genes are found in the *C. milii* genome (SINCAMP00000016734 and SINCAMP00000017945)
IL18	ENSP00000280357	SINCAMP00000011586	
IL19	ENSP00000343000		Apart from IL10, scaffold_70 contains two additional related genes (scaffold_70: 3,248,099-3,250,677 [SINCAMP00000014175] and scaffold_70: 3,257,735-3,259,786 [SINCAMP00000014177]; both are annotated as IL24. Note that this scaffold contains 3 cytokine genes, while the equivalent human locus contains 4 cytokine genes (IL10, IL19, IL20 and IL24).
IL20	ENSP00000356063		Apart from IL10, scaffold_70 contains two additional related genes (scaffold_70: 3,248,099-3,250,677 [SINCAMP00000014175] and scaffold_70: 3,257,735-3,259,786 [SINCAMP00000014177]; both are annotated as IL24. Note that this scaffold contains 3 cytokine genes, while the equivalent human locus contains 4 cytokine genes (IL10, IL19, IL20 and IL24).
IL21	ENSP00000264497		Absent from syntenic region; no BLAST hit; absent from *G. cirratum* transcriptomes
IL22	ENSP00000329384	SINCAMP00000002010	
IL23A (p19)	ENSP00000228534		
IL24	ENSP00000294984		Apart from IL10, scaffold_70 contains two additional related genes (scaffold_70: 3,248,099-3,250,677 [SINCAMP00000014175] and scaffold_70: 3,257,735-3,259,786 [SINCAMP00000014177]; both are annotated as IL24. Note that this scaffold contains 3 cytokine genes, while the equivalent human locus contains 4 cytokine genes (IL10, IL19, IL20 and IL24).
IL26	ENSP00000229134		Likely absent from IFNG/IL22 locus
IL27 (p28)	ENSP00000349365		
IL31	ENSP00000366234		Absent from syntenic region; no BLAST hit
IL34	ENSP00000288098	SINCAMP00000003055	
EBI3 (IL27B)	ENSP00000221847	SINCAMP00000010280	

LIF	ENSP00000249075	SINCAMP00000007825	scaffold_94: 1,383,634-1,386,116; no annotation, but BLASTp and syntenic genes indicate that this is LIF. In the human genome, LIF and OSM are neighbours, but only a single cytokine-like gene is present on this scaffold.
OSM	ENSP00000215781		Absent from syntenic region on scaffold_94; no BLAST hit
TSLP	ENSP00000339804		Not detected
TGFB1	ENSP00000221930		JW872116; fragment on scaffold_19954
TGFB2	ENSP00000092961	SINCAMP00000004915 SINCAMP00000016426	SINCAMP00000004915 assignment supported by synteny; scaffold_125: 1,335,119-1,390,302. SINCAMP00000016426 may be duplicated TGBF2-like gene.
TGFB3	ENSP00000238682		KC795563; fragment on scaffold_17347

Supplementary Table XI.8 | Tumour Necrosis Factor Ligand and Receptor Superfamilies

H. sapiens gene	*H. sapiens* protein	*C. milii* protein	Remarks
TNFRSF1A	ENSP00000162749	SINCAMP00000017647	may be present on scaffold_317: 91,489-96,623: annotated as TNFRSF14, however this is not supported by syntenic evidence, which suggests that this is a potential orthologue of TNFRSF1A or TNFRSF3; the gene is flanked by genes that map to H. sapiens chromosome 12q13, although they are not immediate neighbours to the H. sapiens TNFRSF1A or TNFRSF3 genes.
TNFRSF1B	ENSP00000365435	SINCAMP00000011509	identified by synteny and phylogeny
TNFRSF3 (LTBR)	ENSP00000228918		see TNFRSF1A
TNFRSF4 (OX-40, CD134)	ENSP00000368538		may be present on scaffold_176: 383,470-386,989; annotated TNFRSF11A, however this is not supported by synteny. This scaffold is syntenic with *H. sapiens* chromosome 1p36.33, which contains the genes TNFRSF4 and TNFRSF18
TNFRSF5 (CD40)	ENSP00000361359		three potential candidates on scaffold_6: 6,846,214-6,857,904 (SINCAMP00000013887; this is annotated as TNFRSF11A, however this is not supported by syntenic evidence); scaffold_6: 6,867,950-6,876,790 (SINCAMP00000013895); scaffold_6: 6,884,069-6,898,430 (SINCAMP00000013897): this scaffold shows clear synteny with *H. sapiens* chromosome 20q13.12, which contains the TNFRSF5 gene; however, the *C. milii* locus contains 3 TNFRSF genes, all are TNFRSF5 candidates.
TNFRSF6 (Fas.CD95)	ENSP00000347979	SINCAMP00000021843	identified by synteny and phylogeny
TNFRSF6B (DcR3)	ENSP00000359013	SINCAMP00000013141	identified by synteny and phylogeny
TNFRSF7 (CD27)	ENSP00000266557		syntenic region not identified, no BLAST hit
TNFRSF8 (CD30)	ENSP00000263932	SINCAMP00000011502	identified by synteny and phylogeny
TNFRSF9 (4-1BB)	ENSP00000366729	SINCAMP00000011609	likely present on scaffold_93: 2,393,004-2,400,059; however, although annotated as TNFRSF9, note that this scaffold is generally syntenic with *H. sapiens* chromosome 1p36.33-p36.22, which contains both TNFRSF9 and TNFRSF14 genes

H. sapiens gene	*H. sapiens* protein	*C. milii* protein	Remarks
TNFRSF10A (DR4)	ENSP00000221132		syntenic region not identified, flanking genes are spread over several short scaffolds, no BLAST hit
TNFRSF10B (DR5)	ENSP00000276431		syntenic region not identified, flanking genes are spread over several short scaffolds, no BLAST hit
TNFRSF10C (DcR1)	ENSP00000349324		syntenic region not identified, flanking genes are spread over several short scaffolds, no BLAST hit
TNFRSF10D (DcR2)	ENSP00000310263		syntenic region not identified, flanking genes are spread over several short scaffolds, no BLAST hit
TNFRSF11A (RANK)	ENSP00000269485	SINCAMP00000022509	identified by synteny and phylogeny
TNFRSF11B (OPG)	ENSP00000297350	SINCAMP00000002408	identified by synteny and phylogeny
TNFRSF12A	ENSP00000326737		syntenic region not identified, not BLAST hit
TNFRSF13B (TACI)	ENSP00000261652	SINCAMP00000019332	one linked gene, ambiguous phlyogeny
TNFRSF13C (BAFF-R)	ENSP00000291232		syntenic region not identified, no BLAST hit
TNFRSF14 (HVEM)	ENSP00000347948	SINCAMP00000011984	likely present on scaffold_93: 3,284,965-3,296,018; however, although annotated as TNFRSF14, note that this scaffold is generally syntenic with *H. sapiens* chromosome 1p36.33-p36.22, which contains both TNFRSF9 and TNFRSF14 genes
TNFRSF16 (NGFR)	ENSP00000172229	SINCAMP00000009871	identified by synteny and phylogeny
TNFRSF17 (BCMA)	ENSP00000053243	SINCAMP00000023620	identified by synteny and phylogeny
TNFRSF18 (GITR)	ENSP00000368570		see TNFRSF4 (OX-40, CD134)
TNFRSF19	ENSP00000371693	SINCAMP00000014331	identified by synteny and phylogeny
TNFRSF19L (RELT)	ENSP00000064780	SINCAMP00000006225	identified by synteny and phylogeny
TNFRSF21 (DR6)	ENSP00000296861	SINCAMP00000009531	identified by synteny (one linked gene) and phylogeny
TNFRSF25 (DR3)	ENSP00000349341	SINCAMP00000011512	identified by synteny (one linked gene)
TNFR gene family members with unclear orthology			
		SINCAMP00000025459	located on scaffold_2: 16,799,375-16,804,852; annotated as TNFRSF19, which is supported by phylogenetic but not syntenic evidence; because this scaffold generally exhibits syteny with the *H. sapiens* chromosome X, a gene on scaffold_81 is better supported as

H. sapiens gene	*H. sapiens* protein	*C. milii* protein	Remarks
			the true TNFRSF19 orthologue. Hence, this is possibly the result of a TNFRSF19 duplication.
		SINCAMP00000001420	scaffold_183: 613,185-620,272; annotated as TNFRSF14; this region is generally syntenic with *H. sapiens* chromosome 1p36.1-p33, which contains the human TNFRSF14 gene. The entire scaffold contains three TNFRSF genes that cannot be confidently identified by direct synteny.
		SINCAMP00000001426	scaffold_183: 624,938-635,267; annotated as CD40 (TNFRSF5); however this is not supported by syntenic evidence (the human TNFRSF5 gene is located on chromosome 20q12-q13.2). This region of scaffold_183 is generally syntenic with *H. sapiens* chromosome 1p36.1-p33, which contains the human TNFRSF14 gene. The entire scaffold contains three TNFRSF genes that cannot be confidently identified by direct synteny.
		SINCAMP00000001427	scaffold_183: 639,657-643,484; annotated as TNFRSF14; this region is generally syntenic with *H. sapiens* chromosome 1p36.1-p33, which contains the human TNFRSF14 gene. The entire scaffold contains three TNFRSF genes that cannot be confidently identified by direct synteny.
		SINCAMP00000003806	scaffold_186: 977,404-982,181
		SINCAMP00000002421	scaffold_187: 881,959-888,494; annotated as TNFRSF6B, which is supported by phylogenetic, but not by syntenic evidence. Based on synteny, this may be a duplicated TNFRSF11B gene.
		SINCAMP00000017647	scaffold_317: 91,489-96,623; annotated as TNFRSF14; however, this is not supported by syntenic evidence, which rather suggests that this is a potential orthologue of TNFRSF1A or TNFRSF3. This gene is flanked by genes that map to *H. sapiens* chromosome 12q13, although they are not immediate neighbours to the *H. sapiens* TNFRSF1A or TNFRSF3 genes.
		SINCAMP00000017342	scaffold_2235: 2,243-8,778; annotated as TNFRSF14. Its location on single gene scaffold precludes more precise assignment, but could be paralogue of TNFRSF14.
		SINCAMP00000015481	scaffold_4123: 930-3,873; annotated as NGFR (TNFRSF16); located on single gene scaffold. An additional TNFRSF16 gene was identified on scaffold_88, indicating that there may be at least two copies of this gene in the *C. milii* genome

H. sapiens gene	*H. sapiens* protein	*C. milii* protein	Remarks
		SINCAMP00000022562	scaffold_4448: 115-4,153; annotated as TNFRSF14. Its location on single gene scaffold precludes more precise assignment, but could be paralogue of TNFRSF14.
		SINCAMP00000010661	scaffold_5741: 89-3,527; annotated as TNFRSF14. Its location on single gene scaffold precludes more precise assignment, but could be paralogue of TNFRSF14.
Ligands (selection)			
TNFSF1	ENSP00000407133		not identified in MHC region of *C. milii*; absent from transcriptomes
TNFSF2	ENSP00000365290		Not detected in *C. milii* assembly and transcriptome; but candidate identified in nurse shark transcriptome (KC814628)
TNFSF3	ENSP00000410481		not identified in MHC region of *C. milii*; absent from transcriptomes
TNFSF5 (CD40LG)	ENSP00000359663	SINCAMP00000023076	identified by synteny and phylogeny
TNFSF6 (FASLG)	ENSP00000356694	SINCAMP00000024563	identified by synteny and phylogeny
TNFSF7 (CD70)	ENSP00000245903		syntenic region not identified; no BLAST hit
TNFSF9	ENSP00000245817	SINCAMP00000008977	potential fragment located on scaffold_7902: 825-1,901; potential fragment of TNFSF9 or TNFSF10
TNFSF10 (TRAIL)	ENSP00000241261	SINCAMP00000021674	identified by synteny (one linked gene) and phylogeny
TNFSF11 (RANKL)	ENSP00000239849	SINCAMP00000002766	identified by synteny and phylogeny
TNFSF14 (LIGHT)	ENSP00000245912		syntenic region not identified; no BLAST hit

Supplementary Table XI.9 | Lymphoid lineage determinants and regulators

H. sapiens gene	*H. sapiens* protein	*C. milii* protein	Remarks
Lymphocyte lineage regulators			
ZBTB7A/LRF	ENSP00000323670	SINCAMP00000010356	
ZBTB7B/ThPOK	ENSP00000292176	KC763333	See Supplementary Fig. XI.6 and transcript in *G. cirratum* transcriptome KC763332
ZBTB7C/KrPOK	ENSP00000328732	SINCAMP00000000448	
RUNX1	ENSP00000300305	SINCAMP00000004083	
RUNX2	ENSP00000319087	SINCAMP00000006257	
RUNX3	ENSP00000308051	SINCAMP00000004283	
SOX13	ENSP00000356172	SINCAMP00000013945	
FOXP3	ENSP00000365369		Related gene present; see Supplementary Fig. XI.5
PAX5	ENSP00000350844	SINCAMP00000013433	
PU.1/SPI1	ENSP00000367799	SINCAMP00000019840	
SPIB	ENSP00000471921	SINCAMP00000018290	
TCF3	ENSP00000262965	SINCAMP00000014972	
TCF12	ENSP00000267811	SINCAMP00000010590	
EBF1	ENSP00000322898	SINCAMP00000015692	
AIOLOS/IKZF3	ENSP00000344544	SINCAMP00000009354	
IKAROS/IKZF1	ENSP00000331614	SINCAMP00000007591	
NOTCH1	ENSP00000277541	SINCAMP00000003506	
DLL4	ENSP00000249749	SINCAMP00000016519	
PRDM1/BLIMP1	ENSP00000358092	SINCAMP00000008650	
BCL6	ENSP00000384371	SINCAMP00000019087	
XBP1	ENSP00000216037	SINCAMP00000008423	
IRF4	ENSP00000370343	SINCAMP00000002901	

H. sapiens gene	*H. sapiens* protein	*C. milii* protein	Remarks
IRF8	ENSP00000268638	SINCAMP00000004830	
FOXOA3	ENSP00000300134	SINCAMP00000012650	
FOXO1	ENSP00000368880	SINCAMP00000006546	
ZBTB17/MIZ1	ENSP00000364895	SINCAMP00000005658	
ZBTB16/PLZF	ENSP00000338157	SINCAMP00000020857	
PATZ1/MAZR	ENSP00000266269	KC707912	AAVX01308284.1; not annotated on scaffold_10363; see also KC707912
RORC	ENSP00000327025		Not identified; syntenic region absent; absent from *G. cirratum* transcriptomes
RORA	ENSP00000261523	SINCAMP00000011961	
TOX	ENSP00000354842	SINCAMP00000023003	
BATF	ENSP00000286639	SINCAMP00000015024	
CMYB	ENSP00000356788	SINCAMP00000024877	
GATA3	ENSP00000368632	SINCAMP00000014487	
BCL11B	ENSP00000349723	SINCAMP00000000128	
EOMES	ENSP00000295743	SINCAMP00000023351	
ETS1	ENSP00000376436	SINCAMP00000018427	
TBX21/TBET	ENSP00000177694	SINCAMP00000017134	
RBPJ	ENSP00000305815	SINCAMP00000002525	
CMAF	ENSP00000327048	KC707913	Only in transcriptomes, not in genomic assembly; see KC707913
MAFA	ENSP00000328364	SINCAMP00000011056	
MAFB	ENSP00000362410	SINCAMP00000005813	
MAFF	ENSP00000345393	SINCAMP00000007860	Candidate; supported by sequence only, no synteny information available
MAFG	ENSP00000350369	SINCAMP00000003223	
MAFK	ENSP00000344903	SINCAMP00000005482	
SATB1	ENSP00000341024	SINCAMP00000007530	

***H. sapiens* gene**	***H. sapiens* protein**	***C. milii* protein**	**Remarks**
FOS	ENSP00000306245	SINCAMP00000015032	Scaffold_319: 287,401-285,004
JUN (AP1)	ENSP00000360266	SINCAMP00000003449 KA353648	Two genes annotated as JUN (SINCAMP00000003449- Scaffold_112: 2,106,049-2,107,096; KA353648- Scaffold_180: 993,391-993,032)
NFATC2	ENSP00000379330	SINCAMP00000013500	Scaffold_6: 3,359,487-3,323,329
NFATC1 (NFAT2)	ENSP00000316553	SINCAMP00000011077	Scaffold_80: 3,266,390-3,329,765
NFATC4 (NFAT3)	ENSP00000250373		Not found in *C. milli*, but found in *G. cirratum* RNAseq (KC814620)
NFATC3 (NFAT4)	ENSP00000300659	SINCAMP00000007125	Scaffold_12: 10,047,001-10,088,986
NFAT5	ENSP00000396538	SINCAMP00000005998	Scaffold_12: 3,686,605-3,712,316
NFKB1	ENSP00000226574	SINCAMP00000015490	Scaffold_21:649,500-718,426
NFKB2	ENSP00000358983	SINCAMP00000025350	Scaffold_3: 14,563,467-14,584,725
RELA (NFKB3)	ENSP00000311508	JW870713	no *C. milii* genomic scaffold
Signalling			
STAT1	ENSP00000354394	SINCAMP00000001520	
STAT3	ENSP00000264657	SINCAMP00000020398	
STAT4	ENSP00000351255	SINCAMP00000001542	
STAT6	ENSP00000300134	SINCAMP00000012650	
LCK	ENSP00000337825	SINCAMP00000004899	
JAK1	ENSP00000343204	SINCAMP00000016030	
JAK3	ENSP00000432511	SINCAMP00000008456	
SYK	ENSP00000364898	SINCAMP00000013604	Scaffold_75: 3,316,637-3,362,460
ZAP70	ENSP00000264972	SINCAMP00000009420	Scaffold_85: 203,759-179,395
SLP76	ENSP00000046794	SINCAMP00000015990	Scaffold_19: 4,885,242-4,945,932
FYN	ENSP00000346671	SINCAMP00000024162	Scaffold_84: 1,944,409-1,931,723
LYN	ENSP00000428924	SINCAMP00000001754	scaffold_17:9609151-9661378

H. sapiens gene	*H. sapiens* protein	*C. milii* protein	Remarks
GRB2	ENSP00000339007	SINCAMP00000002020	Scaffold_10: 2,135,018-2,155,074
AKT1	ENSP00000270202	SINCAMP00000010992	Scaffold_9: 10,358,725-10,376,375
MAP2K1	ENSP00000302486	SINCAMP00000008737	Scaffold_5: 591,857-578,620
MAP2K2	ENSP00000262948	SINCAMP00000010354	Scaffold_85: 1,649,633-1,671,781
MAPK1	ENSP00000215832	SINCAMP00000020187	Scaffold_221: 97-56,125
MAPK8 (JNK1)	ENSP00000353483	SINCAMP00000006509	Scaffold_126: 1,334,017-1,317,695
MAPK9	ENSP00000345524	SINCAMP00000008135	Scaffold_87: 3,213,495-3,203,552
MAP10	ENSP00000352157	SINCAMP00000022605	Scaffold_50: 742,187-786,297
DAG1	ENSP00000312435	SINCAMP00000006741	Scaffold_15: 6,710,560-6,724,418
PRKCB	ENSP00000305355	SINCAMP00000005760	scaffold_147: 1,003,841-1,022,084
PRKCE	ENSP00000306124	SINCAMP00000023222	Scaffold_49: 2,429,044-2,511,423
PRKCA	ENSP00000408695	SINCAMP00000005295	Scaffold_10: 9,147,406-9,121,308
PRKCD	ENSP00000378217	SINCAMP00000006692	Scaffold_15: 6,563,594-6,546,572
PRKCI	ENSP00000295797	SINCAMP00000025114	Scaffold_1: 8,317,543-8,294,238
PRKCQ	ENSP00000263125	SINCAMP00000014448	Scaffold_191: 413,423-392,166
PRKCZ	ENSP00000367830	SINCAMP00000011330	Scaffold_93: 1,615,209-1,724,106
PLCG1	ENSP00000244007	SINCAMP00000006093	scaffold_137:1438981-1479687
PLCG2	ENSP00000352336	SINCAMP00000005324	Scaffold_12: 2,361,357-2,304,355
PPP3CA	ENSP00000378323	SINCAMP00000015438	Scaffold-21: 234,725-111,793
KCNH8	ENSP00000328813	SINCAMP00000007524	Scaffold_56: 2,287,582-2,161,766
mTOR	ENSP00000354558	SINCAMP00000010413	Scaffold_58: 1,411,767-1,582,650
HAVCR2 (TIM3)	ENSP00000312002		not detected
MS4A1 (CD20)	ENSP00000314620		not detected
FKBP1A	ENSP00000383003	SINCAMP00000021614	Scaffold_51: 1,621,800-1,632,540
FKBP11	ENSP00000449751	SINCAMP00000007723	Scaffold_672: 15,631-19,912
FKBP2	ENSP00000310935	JW877518	no *C. milii* genomic scaffold

H. sapiens gene	*H. sapiens* protein	*C. milii* protein	Remarks
FKBP1B	ENSP00000370373	SINCAMP00000021561	Scaffold_51: 1,536,858-1,572,480
FKBP7	ENSP00000413152	SINCAMP00000001047	Scaffold_14: 5,708,007-5,703,851
FKBP6	ENSP00000252037	SINCAMP00000018642	Scaffold_47: 747,639-743,584
FKBP8	ENSP00000388891	SINCAMP00000008870	Scaffold_64: 3,207,376-3,210,462
FKBP15	ENSP00000416158	SINCAMP00000008031	Scaffold_77: 319,821-300,043
FKBP3	ENSP00000216330	SINCAMP00000004049	Scaffold_114: 519,108-522,862
FKBP9	ENSP00000242209	SINCAMP00000010517	Scaffold_83: 2,033,873-2,048,529
FKBP14	ENSP00000222803	SINCAMP00000005315	Scaffold_178: 940,186-949,670
PTPN6	ENSP00000391592	JW866484	Scaffold_8122: 1,059-32
SHP2/PTPN11	ENSP00000340944	SINCAMP00000011807	Scaffold_91: 2,622,571-2,645,717
FASLGa	ENSP00000356694	SINCAMP00000024564SINCAMP00000024563 SINCAMP00000024571	There are three genes in tandem in scaffold_41 (SINCAMP00000024564- 3,537,743-3,535,459; SINCAMP00000024563- 3,519,386-3,517,964; SINCAMP00000024571- 3,544,543-3,542,925)
LRRC32	ENSP00000260061	KA353666	Scaffold_3992: 2,638-4,596
PIAS1	ENSP00000438574	SINCAMP00000015122	Scaffold_282: 148,838-110,015
PIAS2	ENSP00000381648	SINCAMP00000009685	Scaffold_591: 27,010-11,084
PIAS3	ENSP00000376765	KA353655	Scaffold_10165: 983-409
PIAS4	ENSP00000262971		not detected
TYK2	ENSP00000264818	SINCAMP00000003024	Scaffold_180: 50,328-67,947
HSCT (DAP10)	ENSP00000246551		not detected
TYROBP (DAP12)	ENSP00000262629		Not found from *C. milli*, but found in *G. cirratum* transcriptome (KC814624)
CD3zeta	ENSP00000354782	SINCAMP00000024473	Scaffold_42: 3,590,223-3,580,480
SH2D1A (SAP)	ENSP00000360181	SINCAMP00000024872	Scaffold_2: 14,759,749-14,746,474
SH2D1B (EAT2)	ENSP00000356906	SINCAMP00000024259	Two genes in tandem in Scaffold_42 (2,845,258-2,841,947 & 2,864,486-

H. sapiens gene	*H. sapiens* protein	*C. milii* protein	Remarks
			2,859,893)
Costimulation/ adhesion			
CTLA4	ENSP00000303939	SINCAMP00000005234	Scaffold_16:388,005-395,357
		SINCAMP00000005238	Scaffold_16:421,101-428,838; duplicated gene?
CD28	ENSP00000393648	SINCAMP00000024276	
CD276	ENSP00000454940	SINCAMP00000014095 SINCAMP00000011201	Assignment of SINCAMP00000014095 is supported by sequence and synteny; SINCAMP00000011201 is CD276-like gene
ITGAL (LFA-1)	ENSP00000349252	SINCAMP00000000567	Scaffold_1523: 269-11,354
ICAM-1	ENSP00000264832		not detected
VCAM1/ICAM2	ENSP00000388666	SINCAMP00000003744	There may be three genes in tandem in Scaffold_180: 552,659-548,609; 560,649-565,744; 548,914-548,609
SELP	ENSP00000356764	SINCAMP00000022922	Scaffold_41:1,200,630-1,214,578; There seems to be 3 genes in tandem in this region (SELP, SELL, SELE?)
CD44		KA353646	
ITGAE	ENSP00000263087	SINCAMP00000019447	Scaffold_47: 1,626,683-1,601,694
CD83	ENSP00000368450		not detected
SELL	ENSP00000236147	SINCAMP00000022922	see SELP
SELE	ENSP00000331736		not detected in *C. milli*, but present in *G. cirratum* RNA seq (KC814630)
SELPLG	ENSP00000228463	KA353638; KA353638	Scaffold_91: 2,601,623-2,602,414
Sirp beta	ENSP00000279477	KA353665	Scaffold_328: 180,483-185,178
Coreceptors			
CD4	ENSP00000011653		Related gene present; see Supplementary Figs. XI.8;9;10.

H. sapiens gene	*H. sapiens* protein	*C. milii* protein	Remarks
CD8A	ENSP00000283635	SINCAMP00000009412	see Supplementary Fig. XI.7;8.
CD8B	ENSP00000331172	SINCAMP00000009419	see Supplementary Fig. XI.7;8.
Other lymphocyte genes			
RAG1	ENSP00000299440	SINCAMP00000025908	
RAG2	ENSP00000308620	SINCAMP00000025909	
DNTT	ENSP00000360216	SINCAMP00000019696	
AICDA	ENSP00000229335	KC707911	partial sequence in AAVX01329030.1
ADA	ENSP00000361965	SINCAMP00000013030	
DCLRE1C/ ARTEMIS	ENSP00000350349	SINCAMP00000020621	
LIG4	ENSP00000349393	SINCAMP00000025957	
NHEJ1	ENSP00000349313	SINCAMP00000019288	
ORAI1	ENSP00000328216	SINCAMP00000026452	
PNP	ENSP00000354532	SINCAMP00000012687	
PRF1	ENSP00000316746	SINCAMP00000026251	
STIM1	ENSP00000300737	SINCAMP00000022129	
STX11	ENSP00000356540	SINCAMP00000022396	
UNC13D	ENSP00000207549	SINCAMP00000005853	
CIITA	ENSP00000316328	SINCAMP00000005130	
IgJ (J chain)	ENSP00000440066	SINCAMP00000017282	Scaffold_208: 110,393-116,382
Lymphoid organ regulators			
TLX1/HOX11	ENSP00000359215	SINCAMP00000006174	
FOXN1	ENSP00000226247	SINCAMP00000020745	
AIRE	ENSP00000291582	SINCAMP00000025355	

Supplementary Table XI.10 | Genes involved in the apoptosis pathway

H. sapiens gene	*H. sapiens* protein	*C. milii* protein	Remarks
BCL2	ENSP00000381185	SINCAMP00000012079	Scaffold_79: 2,457,321-2,389,102
BCLX	ENSP00000365230	SINCAMP00000004879	Scaffold_137: 40,678-42,413
BCL2L2	ENSP00000250405		
MCL1	ENSP00000358022	SINCAMP00000014943 JW867639	Two candidates were detected: (1) SINCAMP00000014943 in scaffold_2238: 4,046-5,878 (2) a gene distributed across two scaffolds (scaffold-2286: 636-388; scaffold_2177: 2,464-2,577)
APAF1	ENSP00000448165	SINCAMP00000017275	Scaffold_194: 891,703-743,621
BID	ENSP00000318822	SINCAMP00000002239	Scaffold_133: 1,997,934-1,994,695
BAK1	ENSP00000363591	SINCAMP00000016573	Scaffold_33: 1,350,963-1,363,138
BAX	ENSP00000293288	SINCAMP00000014512 SINCAMP00000014463	two genes ~46% similar to each other are found in tandem in Scaffold_349 (SINCAMP00000014512:189,755-191,483; SINCAMP00000014463: 175,142-173,881)
BLK	ENSP00000259089		not found in the *C. milii* genomic scaffold, but found in *G. cirratum* transcriptome (KC814634)
HRK	ENSP00000257572		not detected
BCL2L11	ENSP00000376943	SINCAMP00000023919	Scaffold_31: 4,997,189-4,977,379
BAD	ENSP00000309103	KA353645	not found in the genomic scaffold
NOXA	ENSP00000269518		not detected
BBC3	ENSP00000404503		not found in the *C. milii* databases, but found in the *S. canicula* (Scanicula_Contig2889; ORF:47..574 Frame -2)
BMF	ENSP00000346697	SINCAMP00000023259	Scaffold_28: 491,310-513,131
FADD	ENSP00000301838	SINCAMP00000012311	Scaffold_8: 10,116,178-10,117,903

CFLAR	ENSP00000312455		not found in *C. milii* databases, but found in *L. erinacea* and *S. canicula* (Lerinacea_Contig16392; Scanicula_Contig94283 ORF:137..1300 Frame +2)
DIABLO	ENSP00000398495	SINCAMP00000007630	Scaffold_94: 581,047-582,770
XIAP	ENSP00000360242	SINCAMP00000024937	Scaffold_2: 14,881,599-14,876,090
TP53	ENSP00000269305	AEW46988.1	not found in the genomic scaffold

Supplementary Table XI.11 | Vertebrate CD8 sequences used for phylogenetic analyses

CD8A sequences				
Species	**Abbreviation**	**Gene name**	**Protein Sequence ID**	**Length (aa)**
Ailuropoda melanoleuca	Am	CD8A	ENSAMEP00000010525	268
Bos taurus	Bt	CD8A	ENSBTAP00000028175	242
Callithrix jacchus	Cj	CD8A	Q3LRP5_CALJA	235
Canis lupus familiaris	Clf	CD8A	ENSCAFP00000011083	239
Cavia porcellus	Cp	CD8A	ENSCPOP00000019634	237
Dasypus novemcintus	Dn	CD8A	ENSDNOP00000001766	238
Echinops telfairi	Et	CD8A	ENSETEP00000012809	236
Equus caballus	Eca	CD8A	ENSECAP00000011123	244
Erinaceus europaeus	Ee	CD8A	ENSEEUP00000013542	203
Felis catus	Fc	CD8A	ENSFCAP00000013754	239
Gorilla gorilla	Ggo	CD8A	ENSGGOP00000003948	219
Homo sapiens	Hs	CD8A	ENSP00000283635	235
Ictidomys tridecemlineatus	It	CD8A	ENSSTOP00000010260	238
Loxodonta africana	La	CD8A	ENSLAFP00000003269	235
Macaca mulatta	Mmu	CD8A	ENSMMUP00000004695	235
Macropus eugenii	Me	CD8A	ABX79404.1	241
Monodelphis domesticata	Md	CD8A	ENSMODP00000011883	235
Mus musculus	Mm	CD8A	ENSMUSP00000068123	247
Mustela putorius furo	Mpf	CD8A	ENSMPUP00000009228	242
Nomascus leucogenys	Nl	CD8A	ENSNLEP00000003542	275
Oryctolagus cuniculus	Oc	CD8A	ENSOCUP00000008088	234
Otolemur garnettii	Og	CD8A	ENSOGAP00000006471	276
Pan troglodytes	Pt	CD8A	ENSPTRP00000020853	276
Pongo abelii	Pa	CD8A	ENSPPYP00000013612	272
Procavia capensis	Pc	CD8A	ENSPCAP00000013977	235
Pteropus vampyrus	Pv	CD8A	ENSPVAP00000008380	236
Rattus noveticus	Rn	CD8A	ENSRNOP00000009516	236
Sus scrofa	Sus	CD8A	ENSSSCP00000008772	236
Anas playrhynchos	Ap	CD8A	ACL52151.1	237
Anas poecilorhyncha	Apo	CD8A	AFP48734.1	237
Anser anser	Aa	CD8A	AFG26509.1	236
Cairina moschata	Cmo	CD8A	AAW63064.1	237
Gallus gallus	Gg	CD8A	ENSGALP00000025510	235
Meleagris gallopavo	Mg	CD8A	ENSMGAP00000013908	238
Pelodiscus sinensis	Ps	CD8A	ENSPSIP00000017952	266
Taeniopygia guttata	Tg	CD8A	ENSTGUP00000011178	216
Xenopus laevis	Xl	CD8A	ENSXETP00000004085	219
Xenopus tropicalis	Xt	CD8A	XP_002937320.1	234
Callorhinchus milii	Cm	CD8A	SINCAMP00000009412	211
Ctenopharyngodon idella	Ci	CD8A	ACU30711.1	215
Danio rerio	Dr	CD8A	ENSDARP00000065841	216
Dicentrarchus labrax	Dl	CD8A	AAZ66439.1	222
Epinephelus coioides	Ec	CD8A	ACS68183.1	227
Ginglymostoma cirratum	Gc	CD8A	KC707917	220
Gasterosteus aculeatus	Ga	CD8A	ENSGACP00000011825	224

Hippoglossus hippoglossus	Hh	CD8A	ACF04751.1	216
Ictalurus punctatus	Ip	CD8A	NP_001187260.1	223
Oncorhynchus mykiss	Om	CD8A	NP_001117735.1	226
Oreochromis niloticus	On	CD8A	ENSONIP00000025678	222
Oryzias latipes	Ol	CD8A	ENSORLP00000010686	228
Paralichthys olivaceus	Po	CD8A	BAC66490.1	225
Rhinobatos productus	Rp	CD8A	ABQ85060.1	219
Salmo salar	Ss	CD8A	NP_001117055.1	226
Siniperca chuatsi	Sc	CD8A	ADK56159.1	221
Takifugu rubripes	Tr	CD8A	ENSTRUP00000038110	219
Tetraodon nigroviridis	Tn	CD8A	ENSTNIP00000000873	231
Xiphophorus maculatus	Xm	CD8A	ENSXMAP00000002139	224
Outgroup:				
Homo sapiens	Hs	CD7	ENSP00000312027	240
CD8B sequences				
Species	**Abbreviation**	**Gene name**	**Protein Sequence ID**	**Length (aa)**
Ailuropoda melanoleuca	Am	CD8B	ENSAMEP00000010534	205
Bos taurus	Bt	CD8B	ENSBTAP00000032971	210
Callithrix jacchus	Cj	CD8B	ENSCJAP00000035959	245
Canis lupus familiaris	Clf	CD8B	ENSCAFP00000011081	210
Cavia porcellus	Cp	CD8B	ENSCPOP00000004922	210
Dipodomys ordii	Do	CD8B	ENSDORP00000003622	231
Equus caballus	Eca	CD8B	ENSECAP00000000773	210
Gorilla gorilla	Ggo	CD8B	ENSGGOP00000007291	209
Homo sapiens	Hs	CD8B	ENSP00000331172	243
Ictidomys tridecemlineatus	It	CD8B	ENSSTOP00000003724	210
Macaca mulatta	Mmu	CD8B	ENSMMUP00000004703	245
Macropus eugenii	Me	CD8B	ABX79406.1	207
Microcebus murinus	Mmur	CD8B	ENSMICP00000010382	246
Monodelphis domesticata	Md	CD8B	ENSMODP00000011916	206
Mus musculus	Mm	CD8B	ENSMUSP00000070131	213
Mustela putorius furo	Mpf	CD8B	ENSMPUP00000009237	209
Myotis lucifugus	Ml	CD8B	ENSMLUP00000003241	200
Nomascus leucogenys	Nl	CD8B	ENSNLEP00000000411	196
Oryctolagus cuniculus	Oc	CD8B	ENSOCUP00000008094	206
Otolemur garnettii	Og	CD8B	ENSOGAP00000006477	210
Pan troglodytes	Pt	CD8B	ENSPTRP00000059042	243
Pongo abelii	Pa	CD8B	ENSPPYP00000013609	243
Procavia capensis	Pc	CD8B	ENSPCAP00000010341	241
Rattus noveticus	Rn	CD8B	ENSRNOP00000009392	208
Sarcophilus harrisii	Sh	CD8B	ENSSHAP00000013243	214
Sus scrofa	Sus	CD8B	ENSSSCP00000008768	209
Tursiops truncatus	Tt	CD8B	ENSTTRP00000013776	206
Ambystoma mexicanum	Ame	CD8B	AAF61253.1	226
Ficedula albicollis	Fa	CD8B	ENSTGUP00000011182_1	206
Gallus gallus	Gg	CD8B	ENSGALP00000031383	207
Meleagris gallopavo	Mg	CD8B	XP_003206080.1	207
Melopsittacus undulatus	Mu	CD8B	TGUHOMP00000011182_1	207

Pelodiscus sinensis	Ps	CD8B	ENSPSIP00000017761	214
Taeniopygia guttata	Tg	CD8B	ENSTGUP00000011182	205
Xenopus laevis	Xl	CD8B	ADV71261.1	220
Callorhinchus milii	Cm	CD8B	SINCAMP00000009419	209
Ctenopharyngodon idella	Ci	CD8B	ACU30712.1	210
Danio rerio	Dr	CD8B	ENSDARP00000076041	208
Dicentrarchus labrax	Dl	CD8B	CBN81109.1	197
Epinephelus coioides	Ec	CD8B	ACS68186.1	212
Ginglymostoma cirratum	Gc	CD8B	KC814635	214
Gasterosteus aculeatus	Ga	CD8B	ENSGACP00000011833	205
Hippoglossus hippoglossus	Hh	CD8B	ACF04750.1	212
Ictalurus punctatus	Ip	CD8B	NP_001187190.1	210
Oncorhynchus mykiss	Om	CD8B	NP_001117480.1	213
Oreochromis niloticus	On	CD8B	ENSONIP00000025675	213
Oryzias latipes	Ol	CD8B	ENSORLP00000010674	212
Paralichthys olivaceus	Po	CD8B	BAM65617.1	212
Salmo salar	Ss	CD8B	NP_001117056.1	214
Siniperca chuatsi	Sc	CD8B	ADV78595.1	212
Takifugu rubripes	Tr	CD8B	ENSTRUP00000038072	210
Tetraodon nigroviridis	Tn	CD8B	ENSTNIP00000000172	219
Xiphophorus maculatus	Xm	CD8B	ENSXMAP00000002132	217

Supplementary Table XI.12 | Vertebrate CD4 and related sequences used for phylogenetic analyses

CD4 sequences				
Species	**Abbreviation**	**Gene name**	**Protein Sequence ID**	**Length (aa)**
Ailuropoda melanoleuca	Am	CD4	ENSAMEP00000014814	412
Bos taurus	Bt	CD4	ENSBTAP00000037566	455
Canis lupus familiaris	Clf	CD4	NP_001003252.1	463
Cavia porcellus	Cp	CD4	ENSCPOP00000011087	460
Equus caballus	Eca	CD4	ENSECAP00000007248	461
Felis catus	Fc	CD4	NP_001009250.1	474
Gorilla gorilla	Ggo	CD4	ENSGGOP00000003173	458
Homo sapiens	Hs	CD4	ENSP00000011653	458
Ictidomys tridecemlineatus	It	CD4	ENSSTOP00000015536	456
Loxodonta africana	La	CD4	ENSLAFP00000023728	450
Macropus eugenii	Me	CD4	ABR22561.1	464
Myotis lucifugus	Ml	CD4	ENSMLUP00000020857	447
Mus musculus	Mm	CD4	ENSMUSP00000024044	457
Macaca mulatta	Mmu	CD4	ENSMMUP00000017342	458
Mustela putorius furo	Mpf	CD4	ENSMPUP00000016704	462
Nomascus leucogenys	Nl	CD4	ENSNLEP00000005682	458
Ornithorhynchus anatinus	Oa	CD4	ENSOANP00000010475	499
Oryctolagus cuniculus	Oc	CD4	ENSOCUP00000016703	463
Pongo abelii	Pa	CD4	ENSPPYP00000004788	414
Pan troglodytes	Pt	CD4	ENSPTRP00000054833	458
Rattus noveticus	Rn	CD4	ENSRNOP00000021915	457
Sarcophilus harrisii	Sh	CD4	ENSSHAP00000018659	485
Sus scrofa	Sus	CD4	ENSSSCP00000031048	457
Tursiops truncatus	Tt	CD4	ENSTTRP00000004366	454
Anser anser	Aa	CD4	AFG26508.1	480
Anas playrhynchos	Ap	CD4	ENSAPLP00000002157	482
Cairina moschata	Cmo	CD4	AAW63065.1	482
Gallus gallus	Gg	CD4	ENSGALP00000036316	487
Meleagris gallopavo	Mg	CD4	ENSMGAP00000014929	488
Pelodiscus sinensis	Ps	CD4	ENSPSIP00000016561	458
Xenopus laevis	Xl	CD4	NP_001233240.1	473
Danio rerio	Dr	CD4-2	ENSDARP00000112097	383
Hippoglossus hippoglossus	Hh	CD4-2	ADP55206.1	308
Ictalurus punctatus	Ip	CD4-like protein 2	NP_001187156.1	412
Oryzias latipes	Ol	CD4-like	ENSORLP00000015990	302
Oncorhynchus mykiss	Om	CD4-related (1)	AAY42071.1	334
Oncorhynchus mykiss	Om	CD4-related (2)	NP_001118012.1	323
Oreochromis niloticus	On	CD4-2	ENSONIP00000016374	307
Paralichthys olivaceus	Po	CD4-2	BAM65616.1	302
Salmo salar	Ss	T-cell	NP_001139880.1	314

		surface glycoprotein CD4 (CD4-T)		
Tetraodon nigroviridis	Tn	CD4-2	ABU95652.1	309
Tetraodon nigroviridis	Tn	CD4-like	CAF97820.1	312
Takifugu rubripes	Tr	CD4-like	XP_003966373.1	258
Xiphophorus maculatus	Xm	CD4	ENSXMAP00000011077	315
Ctenopharyngodon idella	Ci	CD4-like	ACU30713.1	469
Dicentrarchus labrax	Dl	CD4	CAO98731.1	480
Danio rerio	Dr	CD4	ENSDARP00000121210	474
Epinephelus coioides	Ec	CD4	ADM47441.1	469
Gasterosteus aculeatus	Ga	CD4	ENSGACT00000013008	467
Hippoglossus hippoglossus	Hh	CD4	ACM50925.1	462
Ictalurus punctatus	Ip	CD4-like protein 1	NP_001187155.1	471
Lateolabrax japonicus	Lj	CD4	AFK73394.1	470
Oryzias latipes	Ol	CD4	ENSORLP00000015999	469
Oncorhynchus mykiss	Om	CD4	NP_001118011.1	489
Oreochromis niloticus	On	CD4-1	ENSONIP00000016380	538
Paralichthys olivaceus	Po	CD4-1	BAM65615.1	464
Siniperca chuatsi	Sc	CD4	ADV78594.1	549
Salmo salar	Ss	CD4-like	NP_001117083.1	490
Tetraodon nigroviridis	Tn	CD4-4a	ABU95653.1	466
Tetraodon nigroviridis	Tn	CD4-4b	ABU95654.1	454
Takifugu rubripes	Tr	CD4	ENSTRUP00000027426	464
LAG3 sequences				
Species	**Abbreviation**	**Gene name**	**Protein Sequence ID**	**Length (aa)**
Ailuropoda melanoleuca	Am	LAG3	ENSAMEP00000014826	523
Bos taurus	Bt	LAG3	ENSBTAP00000045595	516
Callithrix jacchus	Cj	LAG3	ENSCJAP00000014207	531
Canis lupus familiaris	Clf	LAG3	ENSCAFP00000021640	476
Cavia porcellus	Cp	LAG3	ENSCPOP00000011084	528
Dipodomys ordii	Do	LAG3	ENSDORP00000001447	500
Equus caballus	Eca	LAG3	ENSECAP00000002333	521
Gorilla gorilla	Ggo	LAG3	ENSGGOP00000023756	533
Homo sapiens	Hs	LAG3	ENSP00000203629	525
Ictidomys tridecemlineatus	It	LAG3	ENSSTOP00000013947	514
Loxodonta africana	La	LAG3	ENSLAFP00000020155	518
Monodelphis domesticata	Md	LAG3	ENSMODP00000022611	484
Macropus eugenii	Me	LAG3	ENSMEUP00000006085	518
Myotis lucifugus	Ml	LAG3	ENSMLUP00000012548	514
Mus musculus	Mm	LAG3	ENSMUSP00000032217	521
Macaca mulatta	Mmu	LAG3	ENSMMUP00000017334	533
Mustela putorius furo	Mpf	LAG3	ENSMPUP00000016720	521
Nomascus leucogenys	Nl	LAG3	ENSNLEP00000005651	496

Oryctolagus cuniculus	Oc	LAG3	ENSOCUP00000009632	526
Otolemur garnettii	Og	LAG3	ENSOGAP00000016914	520
Ochontona princeps	Op	LAG3	ENSOPRP00000003416	467
Pongo abelii	Pa	LAG3	ENSPPYP00000004787	525
Pan troglodytes	Pt	LAG3	ENSPTRP00000042500	527
Pteropus vampyrus	Pv	LAG3	ENSPVAP00000012921	433
Rattus noveticus	Rn	LAG3	ENSRNOP00000036771	525
Sarcophilus harrisii	Sh	LAG3	ENSSHAP00000018480	488
Sus scrofa	Sus	LAG3	ENSSSCP00000000734	507
Gallus gallus	Gg	LAG3	XP_416510.2	504
Oncorhynchus mykiss	Om	LAG3	NP_001182204.1	481
Oreochromis niloticus	On	LAG3	ENSONIP00000016263	476
Takifugu rubripes	Tr	LAG3	XP_003966355.1	460
Xiphophorus maculatus	Xm	LAG3	ENSXMAP00000011431	459
CD2 sequences				
Species	**Abbreviation**	**Gene name**	**Protein Sequence ID**	**Length (aa)**
Ailuropoda melanoleuca	Am	CD2	ENSAMEP00000005474	345
Bos taurus	Bt	CD2	ENSBTAP00000022936	338
Callithrix jacchus	Cj	CD2	ENSCJAP00000011022	352
Canis lupus familiaris	Clf	CD2	ENSCAFP00000014445	339
Cavia porcellus	Cp	CD2	ENSCPOP00000000585	348
Echinops telfairi	Et	CD2	ENSETEP00000000399	346
Equus caballus	Eca	CD2	ENSECAP00000017053	347
Felis catus	Fc	CD2	ENSFCAP00000001827	336
Gorilla gorilla	Ggo	CD2	ENSGGOP00000005147	351
Homo sapiens	Hs	CD2	ENSP00000358490	351
Ictidomys tridecemlineatus	It	CD2	ENSSTOP00000003746	343
Loxodonta africana	La	CD2	ENSLAFP00000012048	343
Macaca mulatta	Mmu	CD2	ENSMMUP00000016101	351
Microcebus murinus	Mmur	CD2	ENSMICP00000010840	351
Monodelphis domesticata	Md	CD2	ENSMODP00000028992	350
Mus musculus	Mm	CD2	ENSMUSP00000029456	344
Mustela putorius furo	Mpf	CD2	ENSMPUP00000001610	343
Myotis lucifugus	Ml	CD2	ENSMLUP00000002902	350
Nomascus leucogenys	Nl	CD2	ENSNLEP00000006260	351
Oryctolagus cuniculus	Oc	CD2	ENSOCUP00000007321	350
Otolemur garnettii	Og	CD2	ENSOGAP00000021497	334
Pan troglodytes	Pt	CD2	ENSPTRP00000051908	351
Pongo abelii	Pa	CD2	ENSPPYP00000001142	351
Procavia capensis	Pc	CD2	ENSPCAP00000001074	345
Pteropus vampyrus	Pv	CD2	ENSPVAP00000005946	344
Rattus noveticus	Rn	CD2	ENSRNOP00000021268	344
Sarcophilus harrisii	Sh	CD2	ENSSHAP00000020941	361
Sus scrofa	Sus	CD2	ENSSSCP00000007184	340
Tupaia belangeri	Tb	CD2	ENSTBEP00000002274	344
Tursiops truncatus	Tt	CD2	ENSTTRP00000002236	344
Pelodiscus sinensis	Ps	CD2	ENSPSIP00000011387	335

Xenopus tropicalis	Xt	CD2	XP_002943736.1	369
Callorhinchus milii	Cm	CD2	SINCAMP00000024576	333
Danio rerio	Dr	CD2	ENSDARP00000104683	311
Gadus morhua	Gm	CD2	ENSGMOP00000021281	353
Gasterosteus aculeatus	Ga	CD2	ENSGACP00000019026	347
Ginglymostoma cirratum	Gc	CD2	KC814637	401
Oreochromis niloticus	On	CD2 (3of3)	ENSONIP00000026394	333
Oryzias latipes	Ol	CD2	XP_004069561.1	333
Xiphophorus maculatus	Xm	CD2 (2of2)	ENSXMAP00000004170	358
JAM3 sequences				
Species	**Abbreviation**	**Gene name**	**Protein Sequence ID**	**Length (aa)**
Ailuropoda melanoleuca	Am	JAM3	ENSAMEP00000007551	329
Bos taurus	Bt	JAM3	ENSBTAP00000004124	302
Canis lupus familiaris	Clf	JAM3	ENSCAFP00000014414	310
Cavia porcellus	Cp	JAM3	ENSCPOP00000005883	332
Dipodomys ordii	Do	JAM3	ENSDORP00000012016	351
Felis catus	Fc	JAM3	ENSFCAP00000025232	310
Gorilla gorilla	Ggo	JAM3	ENSGGOP00000006618	355
Homo sapiens	Hs	JAM3	ENSP00000299106	310
Ictidomys tridecemlineatus	It	JAM3	ENSSTOP00000004915	310
Loxodonta africana	La	JAM3	ENSLAFP00000016317	311
Macaca mulatta	Mmu	JAM3	ENSMMUP00000009302	308
Macropus eugenii	Me	JAM3	ENSMEUP00000003790	360
Monodelphis domesticata	Md	JAM3	ENSMODP00000006642	310
Mus musculus	Mm	JAM3	ENSMUSP00000034472	310
Mustela putorius furo	Mpf	JAM3	ENSMPUP00000009024	305
Myotis lucifugus	Ml	JAM3	ENSMLUP00000008832	310
Ochontona princeps	Op	JAM3	ENSOPRP00000009042	351
Oryctolagus cuniculus	Oc	JAM3	ENSOCUP00000004882	341
Otolemur garnettii	Og	JAM3	ENSOGAP00000019903	308
Pan troglodytes	Pt	JAM3	ENSPTRP00000007683	355
Pongo abelii	Pa	JAM3	ENSPPYP00000004683	315
Pteropus vampyrus	Pv	JAM3	ENSPVAP00000003050	354
Rattus noveticus	Rn	JAM3	ENSRNOP00000012247	310
Sarcophilus harrisii	Sh	JAM3	ENSSHAP00000014916	308
Sus scrofa	Sus	JAM3	ENSSSCP00000016179	312
Tursiops truncatus	Tt	JAM3	ENSTTRP00000002203	332
Anolis carolinensis	Ac	JAM3	ENSACAP00000000996	313
Gallus gallus	Gg	JAM3	ENSGALP00000002228	293
Meleagris gallopavo	Mg	JAM3	ENSMGAP00000000979	312
Pelodiscus sinensis	Ps	JAM3	ENSPSIP00000018013	310
Callorhinchus milii	Cm	JAM3	SINCAMP00000019910	312
Danio rerio	Dr	jam3b	ENSDARP00000082979	333
Gasterosteus aculeatus	Ga	JAM3 (1of2)	ENSGACP00000026659	309
Oreochromis niloticus	On	JAM3 (2of3)	ENSONIP00000014007	307
Oryzias latipes	Ol	JAM3 (1of2)	ENSORLP00000000363	312

Takifugu rubripes	Tr	JAM3 (1of2)	ENSTRUP00000013636	305
Xiphophorus maculatus	Xm	JAM3 (1of2)	ENSXMAP00000006065	312

Supplementary Table XI.13 | CD4/LAG3-related sequences in cartilaginous fishes.

Protein domains were identified using InterproScan.

Protein Sequence ID	**Transcript**	**Gene location**	**Protein Length (aa)**	**Predicted IgSF domains**	**C-terminal TM domain**	**CxC motif**	**CxH motif**
Human CD4 (ENSP00000011653)		Chr 12p13.31	458	4	yes	yes	yes
Mouse CD4 (ENSMUSP00000024044)		Chr 6	457	4	yes	yes	yes
Chicken CD4 (ENSGALP00000036316)		Chr 1	487	4	yes	yes	no
Turkey CD4 (ENSMGAP00000014929)		Chr 1	488	3	yes	yes	no
Zebrafish CD4 (ENSDARP00000121210)		Chr 16	474	2	yes	yes	no
Zebrafish CD4-2 (ENSDARP00000112097)		Chr 16	383	3	yes	yes	no
Lamprey CD4-like (AAU09669.1)		unknown	378	2	yes	no	no
C. milii (SINCAMP00000010635)	Partial, no start codon (or complete short variant)	scaffold_92: 1,246,629-1,251,771	≥467 (or 424)	≥3 (or 3)	yes	no	no
C. milii (ES_spleen.CUFF.63498.1_prot/KC795564)	Partial, no start codon	scaffold_92: 1,238,976-1,243,099	≥292	≥2	?	?	?

G. cirratum (NS_thymus.comp347834_c8_prot/KC814636)	Complete	unknown	825	5	yes	no	no
G. cirratum (NS_thymus.comp347425_c9_prot/KC707916)	Complete	unknown	628	4	yes	no	yes